Research and Innovation Action (RIA)
HORIZON-CL4-2022-DATA-01-05: Extreme data mining, aggregation and analytics technologies and solutions

CLOUDMINER – Cloud-Native and Intelligent Polyglot Data Mining and Analytics Pipelines

Coordinating person

Name: Scott Hansen
E-mail: s.hansen@opengroup.org
Fax: +32-2-6757721

Participant No.	Participant Organisation Name	Part. Short Name	Country
1 (coordinator)	TOG	TOG	United Kingdom
2	University of L'Aquila	UDA	Italy
3	University of York	YORK	United Kingdom
4	CEA List	CEA	France
5	Maastricht University	MUN	Netherlands
6	Edge Hill University	EHU	United Kingdom
7	Institut für angewandte Systemtechnik Bremen GmbH	ATB	Germany
8	Contemporary Learning Management Systems	CLMS	Greece
9	Infotripla OY	INFT	Finland
10	Kachelmann GMBH (Meteologix)	MLGX	Switzerland
11	CNET Centre For New Energy Technologies SA	EDP	Portugal
12	Maieutica Cooperativa de Ensino Superior CRL	IPM	Portugal
13	Municipio Da Maia	CMM	Portugal
14	Continental Automotive Technologies GmbH	CONTI	Germany

1 Excellence

1.1 Objectives and Ambition

The increasing volume, heterogeneity and importance of data in an ever-more-digital world, has created a strong need for intelligent and scalable data mining and analytics technologies. At the same time, the need for agile data-based decision making and the shortage of skilled software developers with appropriate domain expertise, has boosted the demand for data mining solutions that can be effectively used by domain experts with little/no computer programming background to extract actionable insights [1].

Data mining is typically carried out through multi-step pipelines consisting of activities such as data collection and scrubbing, statistical analysis, machine model training and testing and result visualisation. Figure 1 shows a concrete example of a data analytics pipeline specified in the widely-used Orange[1] low-code tool. The pipeline aims to produce a machine learning model that can classify instances of three sub-species of the Iris flower (Setosa, Verticolor, and Virginica), based on the width and length of their petals and sepals. The pipeline starts with loading a relevant dataset from a CSV file, it then trains a Random Forest model on a subset of it, it evaluates the produced model on the dataset, it produces a confusion matrix that shows which instances were misclassified by the model, and finally, it produces the scatter plot shown in Figure 2 which visualises misclassified instances. Several graphical data mining and analytics tools similar to Orange are currently available, including Knime, Apache Nifi and Rapidminer, and are widely used across different scientific and industrial domains[2]. The main appeal of such **low-code tools**, compared to code-based data science toolkits such as Jupyter Notebooks, is that they enable an increasingly **digital-native and tech-savvy workforce who may lack a computer programming background**, to develop, reason about, and execute pipelines for scientific and decision-making purposes. If custom programming logic is needed, such tools also provide support for injecting scripts in languages such as Python and R in pipelines.

Figure 3 shows a high-level overview of data pipelines, which typically consume heterogeneous data that can be (not necessarily alternatives) real-time, already available in different formats from production systems or even locally stored. The processes by which data moves from one data pipeline component into another are typically and fundamentally based on DAGs (Directed Acyclic Graphs) that specify the flows of data operator applications to produce operational reports, graphical visualizations, or support decision-making processes based on machine learning models. A significant shortcoming of existing tools for developing data pipelines is that they **require all the components of a pipeline to be implemented in the same programming language** (e.g. Python for Orange, Java for Knime and Apache Nifi) instead of allowing pipelines to mix and match polyglot best-of-breed components (e.g. Python for Machine Learning, .NET for access to the native API of Excel spreadsheets). Regarding scalability, many tools only support **single-machine execution** (e.g. Orange), while others, such as Apache Nifi, provide **custom mechanisms for executing multiple parallel copies of a pipeline**. Such custom mechanisms are significantly **underdeveloped and restricted compared to modern container orchestration technologies such as Kubernetes** where individual containers (as

[1] https://orangedatamining.com [2] There are more than 3,000 Orange pipelines on GitHub alone: https://github.com/search?q=scheme+extension%3Aows

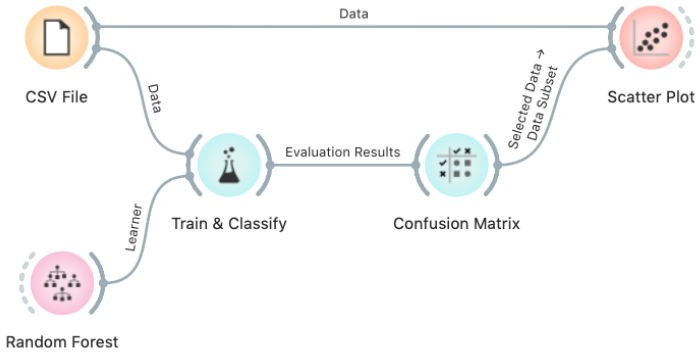

Figure 1: Example Data Analytics Pipeline

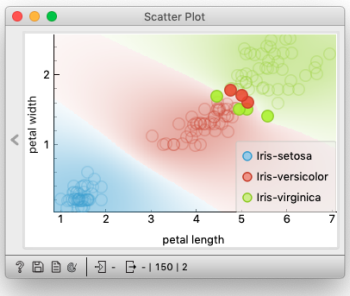

Figure 2: Scatter Plot produced by the Pipeline in Figure 1

opposed to entire instances of the pipeline) can be scaled up/down to best use the available computational resources. Additional limitations include **rudimentary editing support for embedded Python and R scripts** and **little support for intelligent context-aware recommendations** that can help users select and configure data transformation and visualisation components.

Recent developments in technologies for container orchestration (Docker and Kubernetes), program analysis (Language Server Protocol), and distributed systems development (microservice architectures) mandate a fresh rethink that will facilitate the **development and deployment of cloud-native, polyglot and scalable data mining and analytics pipelines**.

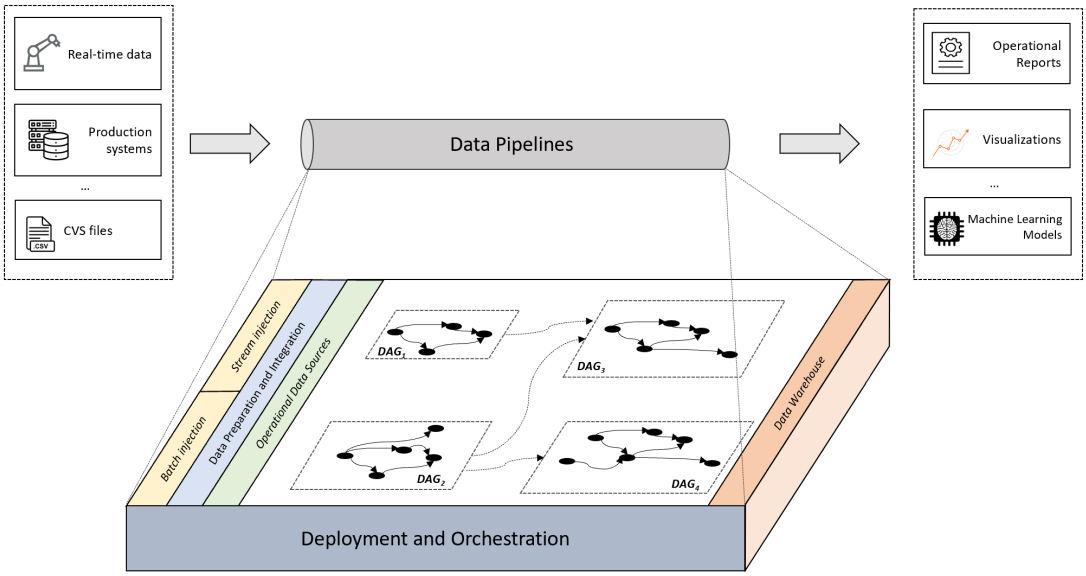

Figure 3: Typical Data Pipelines Architecture

Extreme data makes the context even more complex because it is *"an incarnation of Big Data concept distinguished by the massive amounts of data that must be queried, communicated, and analyzed in near real-time by using a very large number of memory or storage elements and exascale computing systems"* [2]. Examples of extreme data are millions of images per day produced by satellite systems at sub-kilometer resolutions that must be examined to understand better how the Earth's surface is changing over time, billions of social data posts queried in real-time, or hundreds of gigabits-per-second of scientific data that must be stored, filtered, and analyzed [2]. Thus, extreme data is characterized by three main dimensions i.e., *Volume*, *Variety* and *Velocity*. The *Volume* dimension is explained by the Big Data concept related to data size or weight. *Variety* refers to the need to process data from various streaming sources, which can frequently change. The *Velocity* dimension refers to how rapidly data can move from a source to the final user in terms of real-time data visualization and analytics to provide added value and knowledge extracted from the analyzed data. Consequently, it is not surprising that there is increasing interest in supporting the efficient mining of extreme data even though significant challenges are far to be solved, as discussed below:

- **C1 - Ensuring high-quality data**: as previously mentioned, data under analysis can come from different sources and have different formats. Moreover, sorting out missing data entries, fixing outlier values, and removing duplications, are example of tasks that need to be properly decided and applied to ensure data quality [3];
- **C2 - Enabling platform-independent data-mining:** in recent years, different approaches have emerged to support data science processes. However, *"people working with data are less interested in learning new technologies and tools"* [4]. Consequently, the possibility of reusing or enhancing solutions developed in different ecosystems is limited. Current practices demand programming models facilitating the development of data mining processes, improving the portability of developed solutions, and increasing overall productivity [2];
- **C3 - Facilitating the application of advanced data analytics components**: advanced multidimensional visualizations and deep data mining techniques are needed to support users while distilling meaningful patterns from large, heterogeneous, and dispersed data sources. Moreover, ML and data-intensive systems currently use bespoke components with limited reuse capabilities. Generic ML packages are merged with the final system through glue code, which makes the detection of errors and their repair difficult and costly [1];

- **C4 - Providing assistance during data mining processes**: to cope with the complexity of performing data mining processes, it is necessary to monitor what data scientists are doing, identify possible inappropriate choices, and provide recommendations during the entire data science process, including the specification and execution of data pipelines, the selection of algorithms to be applied, and the interpretation of obtained results [1];
- **C5 - Optimizing the execution of data pipelines**: when data is analyzed mainly by centralized solutions, the whole system's performance might be affected due to the unavoidable time needed to transmit the huge amount of data from the related sources to the remote cloud. This calls for distributed environments that can support scalable and efficient model training and usage [5]. Moreover, the application of data pipelines has to be smart by autonomously learning from past executions to optimize how new sparse and heterogeneous data have to be retrieved and mined. Furthermore, the execution of data pipelines has to be optimized with respect to data locality, i.e., the proximity of data to the processing location.

Data pipelines are typically **bespoke** and **developed without properly adhering to software engineering principles** like modularity, reusability, and separation of concerns. To make things even more complex, data mining processes are *"driven by the character of the data being analyzed and by the questions being asked and are often highly exploratory and iterative in nature"* [1]. Thus, **designing, deploying, and executing extreme data pipelines is a complex, technically challenging and error-prone task, requiring broad knowledge and experience across many topics and technologies**. Also with the aim of democratizing data science, more scientific and community innovations are needed to bridge the gap between how data scientists conduct their work and the level of automation that existing approaches can provide, as recognized in [1]. Thus, **automation** has to be fostered to support data science processes given the complexity of data science projects and related **demands for human expertise.** Automated machine learning (AutoML) is an example of successful automations supporting modeling stages where e.g., hyperparameter settings have to be defined, the class of machine learning models have to be chosen, and training algorithms have to be defined. AutoML can fully automate such tasks so that the performance of the ML-based system under development is optimised for the given use case.

The aim of CLOUDMINER is to deliver an integrated open-source platform to support the development and execution of extreme data mining processes by providing automations *i)* for **assisting users while specifying data pipelines** by providing them with relevant recommendations *ii)* for **optimizing the execution of designed pipelines** by properly distributing their components, and *iii)* for **supporting the exploration of analyzed data by means of smart and multidimensional visualizations**. The developed solutions will be integrated in a **scalable and dynamically reconfigurable micro-service architecture** and a **low-code development environment** that will complement and enhance the work of human experts while performing their extreme data mining activities.

Scientific, Technological, Market and Outreach Objectives

CLOUDMINER will provide a comprehensive open-source technical offering for developing and executing extreme data pipelines. The envisioned offering will be an amalgamation of existing best-of-breed technologies and of novel tools and techniques developed within the project. The scientific and market objectives of CLOUDMINER have been elicited through **analysis and synthesis of the business needs of four use cases** from the domains of traffic information and road maintenance management, weather forecasting, smart cities, and automotive, which will be used to evaluate the outcomes of the project (discussed in detail in Section 1.2.6).

The **scientific and technological (S&T) objectives** of CLOUDMINER are:

- STO1: *Development of a data quality management system* to support and optimize preprocessing of high-volume, high-variety, and high-velocity data;
- STO2: *Design and implementation of a polyglot data pipeline development environment* to enable data scientists to model data pipelines in a platform-independent manner and to facilitate the specification of data mining processes by means of high-level abstractions by providing users with sets of predefined patterns limiting errors, reducing programming time, and facilitating resource exploitation;
- STO3: *Development of advanced analytics techniques for deep data mining* to facilitate, by means of smart data visualization and chatbot facilities, the discovery of meaningful, reliable and useful patterns from large, heterogeneous, and dispersed data sources;
- STO4: *Development of intelligent recommenders for data mining assistance* to provide assistance during the whole data mining process by providing users with recommendations that are relevant for the modeling and execution tasks at hand;
- STO5: *Development of an optimised data mining execution environment* to facilitate the deployment and optimized execution of modeled data mining and analytics pipelines by means of a scalable microservice-based architecture.

The **market and outreach (M&O) objectives** of CLOUDMINER are:

- MOO1: Provision of a comprehensive methodology and open source toolkit for the development and optimized execution of extreme data pipelines;
- MOO2: Validation of the CLOUDMINER results through four case studies from the traffic information and road maintenance management, weather forecasting, smart cities, and automotive domains;
- MOO3: High industrial acceptance of the proposed methodology and technical solution;
- MOO4: Liaison with the OMG to influence standardisation activities of future extreme data mining languages;
- MOO5: Contribution of the innovative technical solutions of CLOUDMINER to existing or new projects of the Eclipse/Apache Foundations.

Table 2 provides an overview of performance indicators that will be used to measure the extent to which CLOUDMINER has met its objectives.

Table 2: CLOUDMINER Performance Indicators

Objective	Indicator type and indicative example
STO1: Development of a data quality management system	**Performance:** e.g., automatically and correctly detect at least 30% more data quality errors as compared to state-of-the-art. **Accuracy:** e.g., reduce the number of false positives in automated detection of data quality problems by at least 20% as compared to current practice. **Time-lag-to-analyse:** e.g., reduce the time between the obtainment of data and when it is available for analysis by at least 30% as compared to current practice.
STO2: Design and implementation of a polyglot data pipeline development environment	**Availability:** e.g. the produced open-source tools for polyglot data pipeline development and monitoring are publicly available as open-source software. **Extensibility:** e.g. new data processing components can be integrated in the pipeline development environment with no changes to the environment itself. **Dissemination:** e.g. number of papers published on the design and implementation of the polyglot data pipeline development environment in leading software engineering conferences and journals.
STO3: Development of advanced analytics techniques for deep data mining	**Availability:** e.g. the produced advanced data mining tools are publicly available as open-source software. **Integration:** e.g. the produced advanced data mining tools are integrated in the CLOUDMINER development environment. **Performance:** e.g. the time to compose complex workflows with challenging data is reduced by at least 25% as compared to current practice.
STO4: Development of intelligent recommenders for data mining assistance	**Availability:** e.g., the produced recommender systems and the underpinning training data are publicly available as open-source software and open datasets. **Dissemination:** e.g., number of papers published on the design, development, and accuracy of novel recommenders in leading data science and software engineering conferences and journals. **Integration:** e.g., the produced recommender systems are integrated in the CLOUDMINER development environment
STO5: Development of an optimised data mining execution environment	**Availability:** e.g., the developed pipeline deployment tools and pipeline deployment dashboard are publicly available as open-source software. The underlying deployment language is publicly available as an open specification. **Dissemination:** e.g., number of papers published on the automatic deployment of analytics pipelines; number of talks and demonstrations in industrial conferences. **Integration:** e.g., the pipeline deployment tools and the deployment dashboard are integrated in the CLOUDMINER development environment. The deployment dashboard will be also made available as stand-alone web-based application.
MOO1: Provision of a comprehensive methodology and open source toolkit for the development and optimized execution of extreme data pipelines	**Availability:** e.g. the produced platform is available as open-source software distributed under an industry-friendly BSD-family license (e.g. Eclipse Public License, Apache License), which enables third parties to develop proprietary (closed-source) extensions to the platform. **Usability:** e.g. the components of the platform will be seamlessly integrated and will deliver a consistent and high-quality user experience to engineers.
MOO2: Validation of the CLOUDMINER results through four case studies from the traffic information and road maintenance management, weather forecasting, smart cities, and automotive domains	**Productivity:** e.g. the case studies demonstrate increased productivity of data scientists and engineers (see Table 3 for detailed measures). **Innovation:** e.g. the case studies demonstrate that the CLOUDMINER platform is technically superior to existing data mining platforms for the domains of interest. **Integration:** e.g. the case studies demonstrate that the CLOUDMINER platform can be integrated with existing tools and processes.
MOO3: High industrial acceptance of the proposed methodology and technical solution	**Exploitation:** e.g. the industrial partners of CLOUDMINER use the produced platform in other projects, beyond the CLOUDMINER case studies. **Dissemination:** e.g. the CLOUDMINER platform attracts users and contributors external to the consortium.
MOO4: Liaison with the OMG to influence standardisation activities of future extreme data mining languages	**New standards:** e.g. new OMG standards based on the modeling and execution languages of data pipelines developed in CLOUDMINER.

Table 2: CLOUDMINER Performance Indicators

Objective	Indicator type and indicative example
MOO5: Contribution of the innovative technical solutions of CLOUDMINER to existing or new projects of the Eclipse/Apache Foundations	**New projects:** e.g. a new Eclipse/Apache incubating project is established to provide a reference implementation and supporting tooling for the methodology proposed by CLOUDMINER. **Existing projects:** e.g. contributions (e.g. patches, new modules) are made to existing widely-used projects that the CLOUDMINER platform will build atop – such as Sirius, Epsilon, and Apache Nifi.

Table 3 indicates the measures that will be evaluated through the CLOUDMINER case studies.

Table 3: CLOUDMINER Measures

Measure	CLOUDMINER Target	Justification and Measurement
Diversity of supported data sources	The CLOUDMINER platform will not place any restrictions on the kinds of data that can be minded by the provided facilities.	CLOUDMINER will leverage existing technologies to provide support for the most widely used types of data sources but also provide extensibility mechanisms to support new ones.
Development and deployment effort	A 50% effort decrease in the development and deployment of data pipelines.	By providing users with the CLOUDMINER low-code environment, the time and effort that the development team needs to spend on mining the data at hand should be significantly reduced. In addition, the polyglot microservice-based architecture of CLOUDMINER will enable the reuse and orchestration of existing components, thus reducing errors in the whole data management processes.
Discovery effort	60% decrease of the effort needed to discover and distil patterns from large, heterogeneous, and dispersed data sources.	CLOUDMINER will provide users with intelligent facilities for multidimensional visualizations and a chatbot to answer questions against the data under investigation. Natural language components will simplify the interaction of users with the system.
Usefulness	More than 80% of the produced recommendations will be useful.	Producing irrelevant or incorrect recommendations can be distracting and frustrating for users. In CLOUDMINER, we aim for at least 80% of the produced recommendations to be useful to users. We will measure this through unobtrusive mechanisms (i.e. keeping track of the number of times when recommendations were accepted/ignored by the users). These feedback mechanisms will consequently be used to continuously improve the accuracy of the overall recommendation system.
Standardisation	At least one proposal for creating standards related to specification and execution of data pipelines.	CLOUDMINER aims to promote the standardisation of the envisioned mechanisms with the Object Management Group, and their implementation within new or existing projects of the Eclipse/Apache Foundations. Project partners have a strong track record of engaging in such activities.

1.1.1 Relation to the work programme

CLOUDMINER is strongly aligned with the topic **HORIZON-CL4-2022-DATA-01-05 - Extreme data mining, aggregation and analytics technologies and solutions (RIA)**. The project aims to develop a novel methodology and supporting languages, algorithms, and tools for designing, analyzing, and efficiently executing data pipelines tailored for mining heterogeneous, sparse, and constantly growing amounts of data. CLOUDMINER places particular focus on data that is not necessarily structured and aims to develop facilities that will manage the quality of the data consumed by produced data pipelines. CLOUDMINER will assist the design, deployment, and execution of data pipelines through intelligent recommender systems that will provide users with recommendations relevant to the data science process by relying on the experiences gained by previous data pipeline definitions and executions. To this end, monitoring facilities will be developed to feed a knowledge base used to train the wanted recommender systems. Smart facilities for multidimensional visualizations and a chatbot to answer questions against the data under investigation will also be developed. The contributions of CLOUDMINER will be validated through 4 use cases that will involve datasets with different structures, volume, variety, and velocity (3V) from the traffic information and road maintenance management, weather forecasting, smart cities, and automotive domains.

1.1.2 Ambition

The CLOUDMINER project goes beyond the state of the art by developing novel and integrating existing techniques and tools within the following main technical areas:

- Extraction and Quality Management for Extreme Data;
- Polyglot Data Pipeline Development;
- Advanced Analytics Techniques for Deep Data Mining;
- Intelligent Recommenders for Data Mining Assistance;
- Optimised Data Pipeline Execution.

The following sections present an overview of the state of the art in these areas, which will act as the baseline for CLOUDMINER, as well as the envisioned innovation that the project will deliver.

1.1.2.1 Extraction and Quality Management for Extreme Data

Baseline: Data extraction, cleaning, and transformation (hereafter referred to as data preprocessing) is a time-consuming and effort-intensive activity for data mining. There has been a substantial amount of work on these topics in the last 50 years of database-related research (e.g. [6]). Existing work mainly proposes monolithic solutions to data problems and assumes a one-solution-fits-all approach. Another method, which is becoming more prevalent lately with the increased importance of extreme data, is the use of pipelines for data preprocessing. Data engineering pipelines combine different data extraction, cleaning and transformation algorithms and tools that are automatically executed as needed (e.g. [7, 8, 9]).

Data quality assessment and management are among the most time-consuming tasks in a data engineering pipeline; several proposed solutions exist from large tech companies and organizations. These solutions develop dedicated languages or APIs for specifying data constraints. These constraints describe how data should look like, and the tools can identify instances that violate these constraints. Such solutions include Google's TensorFlow Data Validation (TFDV) [10], Amazon's Deequ [11], and Apache Griffin[3]. Such tools have different features and characteristics, and choosing one over the other requires some kind of trade-off analysis. Moreover, the tools mentioned above suffer some design limitations. For example, most of them come with predefined constraints, or they only allow the user to combine existing constraints to generate new ones. Moreover, although these tools provide usable user interfaces, users must manually implement how these systems can be integrated into a data engineering pipeline. For example, TFDV is configured to run on an Apache Beam[4] cluster, but the integration with popular big data systems, such as Spark, Hadoop, and HDFS, is not straightforward, and it does not come out of the box. Finally, existing data quality assessment and management technologies hinder the automation of data quality tasks, which is of paramount importance for data of high volume and velocity. For example, Deequ support for accessing data checking results is limited; users have to write code to extract and persist these results manually. Similarly, TFDV and Apache Griffin persist the quality properties checked by the tools and the checking results to a file. However, these files do not have a schema, and their semantics is informally defined. Consequently, it is difficult for other tasks in a data mining pipeline to parse and use these results in an automated manner.

To summarise, there are many approaches and tools for data extraction, cleaning, and transformation in data engineering pipelines. However, there is **no support for data scientists on when to use a particular tool or algorithm**; it is **not easy to integrate** many of them in a data preprocessing pipeline, and most of these algorithms and tools **do not support automation** resulting to in high amounts of time and effort spent on these activities.

Innovation: CLOUDMINER will develop novel tools for extracting, cleaning, and transforming extreme data. The proposed solutions will rely on **caching, incrementality, and parallelism to enable high-volume and high-velocity data preprocessing**. Moreover, metadata inference techniques will be used to enrich datasets with semantic information that will be used to integrate high variety of data. The envisaged solution will provide **intelligent guidance to data scientists on which alternative tools or algorithms to use**. To this end, we plan to extend the concept of care label [12] and apply it to data preprocessing tasks. Care labels are currently used for ML algorithms and provide an easy-to-understand method for providing metadata on static properties of ML algorithms. In CLOUDMINER we plan to use data engineering care labels to capture metadata on data preprocessing tools and algorithms, including performance and accuracy of algorithms, computational constraints, and dataset-related constraints. These care labels will be machine-readable and will guide the data scientists on the use of alternative data preprocessing solutions. The solution proposed by CLOUDMINER will support easy integration of pipeline tasks. To support integration, **data preprocessing tasks will be accompanied by sets of pre- and post-conditions and invariants**. They will then be integrated into the pipelines following a design-by-contract approach. Moreover, the input and output of these tasks will be described by a schema that will be provided as part of their care labels.

Automation will be a primary concern of CLOUDMINER. Three different types of automation will be considered: mechanization of relevant tasks, the composition of tasks, and recommendations of tasks. For example, for data cleaning CLOUDMINER will devise a data validation and repair imperative language (as opposed to declarative languages proposed by existing tools such as Deequ) that will support the specification of complex validation patterns and repair behaviors. Machine-Learning algorithms will enable the inference of validation patterns from existing data. **Incremental evaluation of constraints** and their associated metrics will be supported to account for high velocity and volume data. **Distribution of data validation tasks** will be provided to the user transparently. Finally, an evolutionary process will be used to **search for data repairs and recommend the most probable ones** to the user.

1.1.2.2 Polyglot Data Pipeline Development

Baseline: Several low-code platforms such as Knime[5], Orange[6], Apache NiFi[7] and RapidMiner[8] are available for graphically assembling data mining and analytics pipelines from existing data processing and visualisation components. Such environments typically include:

- a library of built-in data transformation and visualisation components;
- extensibility mechanisms for contributing custom components;
- support for integrating custom scripts (e.g. in Java, Python, R) in pipelines;
- facilities for monitoring the execution of pipelines.

Existing platforms are typically monoglot and require custom components to be implemented using a specific programming language (e.g., Knime and Nifi components are implemented in Java, whereas Orange components are implemented in Python). While some of them also provide support for calling web-services (that can wrap around functionality expressed in different programming languages) in pipelines, this is largely an afterthought, and the relevant tooling they provide is rather rudimentary. By effectively **constraining extenders to a single programming language**, the opportunity to use heterogeneous technologies for different parts of the pipeline remains unexploited. As shown in [13], the **selection of programming languages** for different

[3] https://griffin.apache.org/ [4] https://beam.apache.org/ [5] https://www.knime.com [6] https://orangedatamining.com [7] https://nifi.apache.org
[8] https://rapidminer.com

parts of a pipeline can have a **significant impact on end-to-end performance** (a combination of Go and Java for different parts of the pipeline was found to be the most performant solution in that work). Regarding scalability, some platforms only support **single-machine execution** (e.g., Orange) while others, such as Apache Nifi, provide **custom mechanisms** for executing multiple parallel copies of a pipeline. Such custom mechanisms are **significantly restricted** compared to modern container orchestration technologies such as Kubernetes, where individual containers (as opposed to entire instances of the pipeline) can be scaled up/down to make the best use of the available computational resources.

While low-code data mining and analytics platforms aspire to support most recurring data transformation and visualisation tasks via reusable components, sometimes, users **need the expressive power of textual programming languages** such as Python or R to implement custom data processing logic. Existing platforms recognise this need and invariably offer some form of "script" component that allows users to specify a custom program in the context of an otherwise graphical pipeline. However, the support for different programming/scripting languages varies significantly. At best, existing platforms offer advanced editing support for their primary programming language (e.g., Knime provides tool support such as context-aware code completion for embedding custom Java-based scripts in pipelines), but **very basic tooling** for writing such scripts in other languages. With the advent of the Language Server Protocol[9] (LSP) and the availability of high-quality Language Servers that implement its JSON-RPC protocol[10] for most programming languages, there is an **unexploited opportunity to provide high-quality editing support** (including code completion, reference navigation, and refactoring) for a wide range of programming languages.

Innovation: The ambition of CLOUDMINER is to develop an environment where data processing and visualisation components will be truly polyglot and scalable by leveraging containers, container orchestration, and a microservice-based architecture. More specifically, in CLOUDMINER, every data processing and visualisation component will be encapsulated as a **REST API in a Docker container**. The API will provide methods that will specify the **graphical interface** of the component using open standards such as HTML and SVG, and that will **trigger, monitor, and manage the execution** of the component. A web-based front-end will discover such components from a set of configured Docker containers and provide facilities for **assembling them into pipelines** and for coordinating their execution. In this way, developers of data processing and visualisation components will be **free to choose best-of-breed technologies** (e.g., Python for Machine Learning, C++/Cuda for high-performance numerical computations, .NET for access to the native API of Excel spreadsheets) instead of being constrained to the pipeline framework's preferred programming language. Furthermore, through the use of Language Servers and LSP, developers will be offered **rich editing capabilities** for any programming language they implement their components in, as if they were coding in an IDE supporting that language.

In addition, the platform will support **versioning of its data pipelines**. Even though their specifications will be stored in a structured format like JSON, performing common versioning activities on them using standard text-based facilities is bound to be error-prone. Therefore, we envision implementing custom ways to identify changes in pipelines and consequently resolving any conflicts between their different versions.

Finally, pipeline developers will also be able to use **natural language descriptions of their intent** to automatically identify suitable components as an aid to improve efficiency and productivity. The ambition is to design and develop an add-on for CLOUDMINER that will process user text about the desired data analysis process and make suggestions for pipeline components and sequences of components, in an attempt to lower the entry barrier for novice users.

1.1.2.3 Advanced Analytics Techniques for Deep Data Mining

Baseline: To conceive novel techniques to support smart and scalable identification and analysis of meaningful patterns from large, heterogeneous, and dispersed data sources, three different baselines will be considered as main starting points of the work, i.e., research done in the fields of *data mining*, *federated learning*, and *chatbots in software engineering*.

Data mining (DM): DM aims to uncover novel, interesting, and understandable patterns that provide useful knowledge [14]. DM methods such as Frequent Pattern Mining, Pattern Discovery, and Decision Trees are typical, but they often produce an overwhelming number of overlapping or redundant patterns, are challenged performance-wise by large datasets [15], and can be difficult to interpret [16] or explain [17]. In contrast, semantic data mining (SDM) is a form of relational data mining that uses annotated data together with complex semantic background knowledge, such as ontologies and knowledge graphs/bases, to learn more informative rules that are also easier to interpret and explain [18, 19, 20]. However, the drawback of SDM methods lies in the high computational complexity of existing SDM algorithms, resulting in long run times even when applied to relatively small data sets [21]. Prior approaches to improving performance include using network analysis, which prunes background knowledge for more efficient processing using an existing SDM method [22]. NetSDM [23] builds on this approach and computes term significance with a personalized PageRank. While background knowledge has been utilized in various parts of the data mining process, including data pre-processing, model construction, and post-processing [24], **these are not uniformly used**. Pattern Discovery and Disentanglement System (PDD) uses machine learning. It can discover explicit patterns from the data with various sizes and imbalanced groups, and screen out anomalies [25].

Federated learning: Traditional machine learning algorithms, such as Support Vector Machines (SVMs) and Random Forests (RFs), are typically trained and executed on a single node/CPU core. Bespoke parallelisation methods have been proposed for particular algorithms, such as SVMs [26] and RFs [27]. Libraries for deep learning, such as PyTorch and TensorFlow, mainly offer two distributed neural network training methods: data parallelism and model parallelism [28]. In data parallelism, data is partitioned, and each partition is processed in a separate node. A copy of the same model is trained on each node with its subset of data. Each node computes training errors independently, updates its model based on the errors, and communicates all changes to other nodes to update their models. In model parallelism, different model parts are trained on the same data on different nodes. Therefore, **nodes need to synchronise their common parameters in every training step**. Following most of the distributed/federated learning approaches above, data is distributed on a network of inter-linked resources, raising privacy

[9] https://microsoft.github.io/language-server-protocol [10] https://www.jsonrpc.org

concerns when the data is sensitive [29]. In this context, key privacy components are transparency and consent, data minimisation, and anonymisation of released aggregates [30].

Chatbots in software engineering: Recent advances in Natural Language Processing (NLP) have fostered the proliferation of conversational agents or chatbots, i.e. software programs with a Natural Language (NL) user interface. These can be embedded within social networks, e.g. Slack, version control systems, e.g. Github, or be deployed into intelligent speakers for voice-driven conversation. Chatbots are used to facilitate access to all kinds of services (e.g., booking flights, shopping, checking weather conditions), and they are increasingly used to facilitate software engineering tasks [31, 32, 33]. For example, developers use bots to automate deployment tasks, assign software bugs and issues, repair build failures, schedule tasks like sending reminders, or for customer support. Recent literature is concerned with instructing and discussing with robots, querying knowledge graphs or data collections, and modelling business processes. NL interfaces for instructing a robot map free language phrases to the most similar concept known to the robot, e.g., an object, a region, or a goal [34, 35]. Robots can verify or clarify ambiguous instructions by generating clarification questions [36]. The Ontology-based NL Interface (ONLI) maps natural language questions to an ontology model about DBpedia. In general, all applications above are different realisations of the concept of mapping free NL to a controlled language. Controlled languages are subsets of natural language that use a limited vocabulary and simpler grammar rules. As a result, **controlled languages are simpler and more constrained than NL** [37, 38], and can be adjusted to only cover language structures that map directly to user intents. Integrated chatbots have been developed by software companies, e.g. Google's DialogFlow[11], Amazon's Lex[12], IBM's Watson Assistant[13] or the Microsoft Bot Framework[14]. Independent platforms (e.g., FlowXO[15], Landbot[16] or ChatFuel[17]), frameworks (e.g., Rasa[18]), services (e.g, LUIS[19]) and libraries (e.g., ChatterBot[20]) have also emerged.

Innovation: CLOUDMINER aims to create a new suite of advanced data mining technologies to bring insight into the results of complex data transformations generated throughout developed data pipeline. CLOUDMINER will advance the state-of-the-art in four areas: *i)* **smart data visualization**, *ii)* **privacy-preserving data mining**, *iii)* **scalable semantic data mining**, and *iv)* **interactive semantic query answering**.

CLOUDMINER will support users in constructing timely and effective visualizations with a novel smart data visualization technology. The proposed solution will combine performant data profiling and summarization techniques, logic-based and latent semantics, scalable data visualization techniques, and machine learning methods to suggest and automatically generate **interactive visualizations based on the type and relations in the data** generated throughout data pipelines. The proposed solution will **reduce the effort to examine the impact of preceding data transformations** and **guide the development** of downstream data transformations. The visualizations will be **FAIR-by-design**, in that they may be published to connected storage solutions with automatically generated, user-augmented, standardized, and machine-readable metadata detailing their context and provenance in the workflow. Published visualizations will be directly embeddable as intelligent objects in research publications, providing a clear path back to the source pipeline.

CLOUDMINER will advance state of the art in semantic data mining to **identify both well-known patterns and non-obvious but well-supported patterns** in pipeline-generated data. The proposed solution will combine background knowledge in the form of ontologies, deductive reasoning, statistical learning, and network analysis to identify salient parts of a knowledge graph from which logical patterns can be scalably mined using existing software. CLOUDMINER will enable pipeline engineers to examine and validate these patterns as consistent or inconsistent with their expectations. This knowledge will be stored and linked as evidence of correctness of the data transformation for future examination.

Privacy-preserving federated learning is fairly new, both as a research field and as a set of tools and methods that allow to apply it to a variety of data types and application domains [30]. CLOUDMINER provides an excellent opportunity to propose innovative methods for federated learning and associated **privacy preservation** due to the variety of data made available by the use case partners and the size of this data at an extreme scale. Novel federated learning methods will be packaged as pipeline components. Depending on the resources available for each pipeline execution, they will be made available to the user to be selected for deployment.

CLOUDMINER will advance the state of the art in **using natural language for building data pipelines**. We will propose advanced NLP techniques to recognise relevant concepts and relations and map them to existing analysis components.

1.1.2.4 Intelligent Recommenders for Data Mining Assistance

Baseline: Even though data-intensive scientific workflows facilitate the definition of a data pipeline, they have to cope with several issues compared with traditional scientific workflows e.g., diversity of data sources, the deployment of the actual code, and constraints on storage capacity. Thus, such kinds of systems should be carefully conceived by exploiting different techniques that cope with several challenges. PyTerrier [39] is an information retrieval (IR) framework that expresses complex data analysis operations in terms of composable pipelines. By relying on a set of transformers, the proposed framework builds a directed acyclic graph (DAG) that represents essential IR components, e.g., retrieve, re-rank, combine results, rewrite the query, to name a few. Such a pipeline can be customized or even optimized using the graph rewriting patterns technique. PyTerrier eventually generates Python code that implements the specified pipeline and can be integrated into Jupyter notebooks as well as well-known IR search engines, i.e., Anserini [40] and Terrier [41]. Similarly, PyCaret [42] is an open-source machine learning library written in Python that aims at supporting the so-called citizen data scientists who can perform both simple and moderately sophisticated analytical tasks. The library essentially includes several well-known ML modules such as scikit-learn, XGBoost, LightGBM, to name a few. Even though it is still under construction, the goal of PyCaret is to speed up the development of machine learning workflows by providing users with reusable modules, each encapsulating machine learning tasks.

Despite the recent efforts in introducing mechanisms to increase the reuse of well-defined ML tasks and to support developers with fragments of already developed workflows (as in the case of Apache Nifi), producing data pipelines is still a complex process. It involves heterogeneous data sources, different processing algorithms, broad knowledge across many topics, and manual searches for every resource to find the most suitable ones.

[11] dialogflow.com [12] aws.amazon.com/lex [13] www.ibm.com/cloud/watson-assistant [14] dev.botframework.com [15] flowxo.com [16] landbot.io [17] chatfuel.com [18] rasa.com [19] www.luis.ai [20] chatterbot.readthedocs.io

The past decades witnessed the adoption of *recommender systems* as the major technology to help customers navigate the abundant product offerings of online retail platforms. Even the software engineering community is investigating the adoption of recommender systems to assist developers in navigating large information spaces and getting instant recommendations that might be helpful to solve the particular development problem at hand. In particular, a **recommendation system in software engineering (RSSE)** has been defined as "*... a software application that provides information items estimated to be valuable for a software engineering task in a given context*" [43]. When developers join a new project, they have to master many information sources [44] (often in a short time). In such a context, the problem is not the lack of information but an information overload coming from heterogeneous and rapidly evolving sources. Thus, RSSEs aim at giving developers **recommendations, which can consist of different items including code examples, issue reports, reusable source code, and third-party components, documentation**.

The application of recommender systems technology has been only recently investigated to support the development of scientific pipelines. In particular, to address the problem of identifying which pipelines and datasets can appropriately be used together, Mazaheri *et al.* [45] propose a collaborative filtering system to recommend pipelines and datasets based on provenance records from previous executions. The developed recommender engine uses latent factor models to encode the retrieved information as matrixes where the pipelines, datasets, and execution outcomes represent users, items, and ratings, respectively. The conducted evaluation shows that the proposed system can recommend the proper pipeline elements with usable accuracy on the tested datasets and pipelines. The Galaxy tool recommender [46], the PROPHETS system [47], and the WINGS approach [48] are very recent approaches to assist workflow developers to identify candidate software components for a given workflow. Similarly, other recommender systems (like the Oracle machine-learning toolkit [49], PennAI [50], and Dyad ranking [51]) have been conceived to help developers in selecting specific algorithms for a given problem with respect to an objective function such as F1 score, accuracy or mean average error. Most of the proposed recommender systems above have been developed and evaluated in the neuroscience and biomedical fields.

Innovation: CLOUDMINER aspires to advance state of the art in the context of recommender systems to support the development and execution of data pipelines. Advanced machine learning algorithms will enable **real-time recommendations while modeling data pipelines** and thus help select reusable components and apply best practices. Furthermore, the execution of data pipelines will be monitored to feed a **knowledge base** (KB) that will be purposely developed in the project. The KB will be used to train the envisioned recommender systems and provide developers with **suggestions about how to optimize the execution of the pipeline under development**. In particular, the KB will store performance issues and corresponding countermeasures that have been detected and resolved during past executions of similar pipelines.

1.1.2.5 Optimised Data Pipeline Execution

Baseline: One of the most challenging aspects of big data projects has been **turning an experimental pipeline into a viable integrated system** capable of continuous operation that **processes vast amounts of data with the required throughput**. Therefore, when discussing state-of-the-art data processing, several dimensions have to be considered:

- workflow management (resource management, scheduling, reliability, consistency),
- architecture with regards to scalability and performance,
- compatibility, connectivity, integration of existing components,
- ease of use.

Hadoop [21] can be considered the first big data framework that enjoyed widespread use and acclaim. It introduced the MapReduce programming model and the distributed file system HDFS (Hadoop Distributed File System), which allowed large amounts of data processing with a distributed infrastructure. Frameworks such as Tez [22] introduced optimizations to the MapReduce model based on directed acyclic graphs and provided interactive programming. The emphasis on data mining pipelines has shifted from a batch-processed approach to real-time processing in recent years. This development is also reflected in the evolution of execution environments, where batch-processing systems, e.g., based on MapReduce, were extended or replaced by real-time data processing based on more flexible processing engines such as Apache Spark. At its core, Spark is based on the RDDs (Resilient Distributed Dataset) concept, enabling in-memory processing, unlike HDFS, which writes the results to disk. This micro-batch approach allows for near real-time processing of input streams. For some time, the so-called Lambda Architecture received a lot of attention. The three-layered architecture consists of a batch layer, a speed (or real-time) layer, and a serving layer. A stream processing system like Apache Storm[23] is utilised for real-time insights, while simultaneously, a batch-processing system such as Hadoop is generating high-quality results based on all the available data [52]. Apache Flink [24] can be considered the next step of the evolution, as it supports both batch and stream processing and is designed for true stream processing natively, leading to lower latencies. Downsides of Flink are its Java and Scala APIs, leaving support for other languages to be desired [53].

Processing can be divided into distributed and parallel processing. Parallel computing refers to the code running on a single physical machine with several processors working simultaneously on shared memory. In distributed computing, the tasks are shared among many computers, each using its memory. To facilitate this on a cluster or grid of computers, a data-processing engine (for example, Hadoop or Storm) and a persistence service (for example, HDFS or Cassandra) are needed. Hybrid infrastructures have been proposed to enable a common execution environment to work across heterogeneous platforms (e.g., cloud, private cloud, and grids) [54]. An increasingly important requirement for modern execution environments is flexible scalability of the pipeline as a whole and precisely targeting the most demanding components. Container orchestration solutions such as Docker Swarm and Kubernetes are well suited for this purpose. A growing community of practitioners is starting to utilize Kubernetes in the context of data processing pipelines. Open Source projects like Kubeflow[25] take advantage of Kubernetes to offer a cloud-native end-to-end workflow but still lack polish for productive use. Furthermore, since Google maintains the project, it is not working as well on other cloud platforms.

Today, **a broad spectrum of technologies and frameworks is used for data mining tasks**. Users like to have the option to stick with the languages they know or even integrate existing code into new pipelines. Modern data-processing frameworks like

[21] https://hadoop.apache.org/ [22] https://tez.apache.org/ [23] https://storm.apache.org/ [24] https://flink.apache.org/ [25] https://www.kubeflow.org/

Apache Storm and Spark support a variety of popular languages and are compatible with many persistence providers like Cassandra [26] and Hive[27]. Many attempts have been made to simplify setting up a data processing pipeline for the end-user, e.g., providing powerful configuration options instead of requiring code. Some Frameworks such as Apache Pig [28] and Cascading [29], are trying to simplify development of MapReduce algorithms by providing high-level abstractions. But those frameworks come with limitations, as they require the user to learn Pig's custom scripting language or the operator-based programming of Cascading [55].

Innovation: CLOUDMINER aims to mitigate the limitations and downsides of the frameworks mentioned above by supporting pipelining and integrating **heterogeneous and distributed data-processing jobs** (using MapReduce, Spark, and scripts). Also, it will allow the reuse of existing programming paradigms rather than requiring developers to learn new ones for writing analytic algorithms. The ambition of CLOUDMINER is to develop an execution environment that **combines methods from DevOps with methods from DataOps**. Especially some concepts from DevOps, such as continuous integration and continuous delivery (CI/CD), are required for innovative pipeline execution environments, especially when considering the whole lifecycle of data pipelines (from design to maintenance). The efficiency throughout the development of data pipelines will be improved by providing **automated version control**, **dependency management**, and **test automation**. This will facilitate historical traceability and reproducibility, crucial for **continuously developing pipelines**. Secondly, CLOUDMINER envisions using a **microservice-based architecture for the execution environment**. It will encapsulate each data processing component into its own container. Furthermore, the execution environment will **monitor** and deploy such data processing components and **provide APIs** for other CLOUDMINERcomponents. Thus the execution environment will shorten the data pipeline development life cycle and provide continuous delivery with high-quality data pipelines.

1.2 Methodology

This section highlights the main concepts underpinning CLOUDMINER and describes the overall approach that we intend to follow in order to reach the CLOUDMINER objectives. The following sections detail the proposed approach to achieve the Scientific and Technical (S&T) objectives of CLOUDMINER .

1.2.1 STO1: Development of a data quality management system

CLOUDMINER will develop next-generation **reusable components for data preprocessing** that are suitable for extreme data. CLOUDMINER will tackle the challenges of **velocity**, **variability**, and **volume** of data by enabling automation of data preprocessing tasks and by facilitating integration of such tasks in data mining pipelines.

CLOUDMINER will focus on three different types of **automation**, namely mechanization of tasks (e.g., automatic repair of data errors), composition of tasks (e.g., composition of data validation rules), and user assistance in the form of recommendations (e.g., recommendations of data cleaning algorithms based on the type of data used). AI techniques will be used to support automation. Machine Learning will be used to infer data quality validation rules and user recommendations, whereas evolutionary algorithms will be used for automatic error repairs of datasets. Furthermore, to enable **intelligent integration** of datasets and of data preprocessing tasks, CLOUDMINER will devise sophisticated metadata management infrastructure that will adapt and extend the concept of care labels from the Machine Learning domain. In addition, metadata inference techniques will be considered for datasets without metadata. To account for the **scalability** of CLOUDMINER's data preprocessing components, the proposed technologies will support **incrementality** for repeated computations on different versions of the data, **caching** to avoid recomputation of previously computed values, and **parallelism** to enable distribution and acceleration of tasks.

1.2.2 STO2: Design and implementation of a polyglot data pipeline development environment

CLOUDMINER will develop a next-generation cloud-native low-code platform for developing, configuring, executing, and monitoring polyglot data mining and analytics pipelines. Unlike existing low-code platforms such as Apache NiFi, Knime, and Biolabs Orange, which require all components of a pipeline to be implemented in the same programming language (Java for Nifi and KNIME and Python for Orange), CLOUDMINER will provide a **cloud-native microservice-based architecture that will leverage containers and container orchestration** solutions (e.g., Docker and Kubernetes) to facilitate the assembly of polyglot pipelines the components of which can be implemented in different programming languages and technologies.

CLOUDMINER will develop a web-based graphical pipeline editor that will be able to connect to containers that provide data processing and transformation capabilities in the form of microservices, and allow developers to create, configure and link instances of the respective microservices in the context of complex pipelines. To achieve this, CLOUDMINER will define standard interfaces that such microservices will need to conform to for design-time and run-time configuration using technologies such as OpenAPI. To lower the entry barrier to designing data analysis pipelines, CLOUDMINER will use natural language processing as an aid for pipeline design. Data analysts will be able to locate existing data analysis components using an assistant based on natural language. The assistant will process natural language sentences, keyed in by the user, by applying state-of-the-art technical term and relation extraction methods to identify and **suggest existing pipeline components** that match the extracted terms and relations to the user.

As experience with similar platforms has demonstrated, glue scripts in languages such as Python, R, and Julia are often necessary in the context of complex data mining and analytics pipelines. At best, existing platforms offer rich tool support for their primary programming language (e.g., Knime provides tool support such as context-aware code completion for embedding custom Java-based scripts in pipelines) but very basic tooling for writing such scripts in other languages. CLOUDMINER will make use of **existing language servers** for popular languages such as Python, R and Julia to provide **intelligent script editing capabilities** (e.g. context-aware auto-completion, code folding, refactoring) within the graphical editor to improve developer experience and to identify and report errors before a pipeline is executed.

The pipeline development environment of CLOUDMINER will also be augmented with monitoring, runtime-configuration and debugging capabilities. While the development environment will not be necessary for executing pipelines, it will be able to attach

[26] https://cassandra.apache.org/ [27] https://hive.apache.org/ [28] https://pig.apache.org/ [29] https://www.cascading.org/

itself to the execution of local and remote pipelines and **monitor the execution of pipeline components** (e.g., execution time, incoming/outgoing data, backlog), **scale up/down individual components** at runtime by instructing the container orchestrator to adjust the number of container instances respectively, and place breakpoints and input/output watches in individual components to **enable debugging**. Finally, pipelines need to be versioned like any other software. As pipeline specifications will be stored in a structured format (e.g. XMI, XML, JSON), performing differencing, conflict detection, and merging on them using standard text-based facilities is error-prone. Therefore, CLOUDMINER will implement custom facilities for **identifying changes between different versions of a pipeline** and for **detecting and resolving conflicts** in a semi-automated way (some conflicts will require user intervention to resolve).

1.2.3 STO3: Development of advanced analytics techniques for deep data mining

CLOUDMINER will develop semantic data mining (SDM) tools to enable pipeline engineers to **identify correct, meaningful, and problematic patterns** in pipeline-generated data so as to support the development of sound and complete data pipelines. Our focus is twofold: *i)* to improve the overall performance of finding patterns by developing new semantic indexing methods, and *ii)* to improve the quality of the patterns by applying background knowledge (e.g ontologies mapped via the data semantics pipeline in WP2), across data preprocessing, model building, and post-processing. We will leverage existing state of the art semantic data mining approaches (e.g. NetSDM), particularly on large and complex relational data/graphs, which are needed to uncover statistically enriched subgroups in large data. SDM components will be evaluated in their speed to produce patterns and the quality of patterns in relation to tasks identified from CLOUDMINER use cases.

CLOUDMINER will create **smart and effective visualization components** to support data pipeline engineers in the development and execution of their pipelines. Data profiling and summarization techniques will be combined with lightweight NoSql database technologies and high level declarative plotting libraries (e.g., plotly[30]) to power interactive and customizable data dashboards. More powerful faceted visualizations that exploit semantic structures (e.g. type and property hierarchies) will be developed by exploiting data semantics identified via WP2's semantic data preprocessing pipeline. Advanced machine learning techniques, including recommender systems arising from WP5, will be exploited to predict and suggest data visualizations based on the (semantic) type of data selected and user activity. Data engineers will be able to (openly) publish embeddable visualizations with structured, standardized metadata including workflow provenance and visualization settings using data storage connectors developed in WP3. Visualization components will be evaluated using a number of criteria including utility, usability, and performance on tasks identified from CLOUDMINER use cases.

CLOUDMINER will design and implement pipeline components that will allow **machine learning methods to be trained and executed in a distributed manner**, taking **privacy for sensitive data** into account. We will focus separately on federated training for machine learning models, federated training for deep learning models, and federated execution of pre-trained models. We will consider traditional machine learning algorithms, such as Random Forests, Decision Trees, and Support Vector Machines and popular deep learning models such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Attention models. Following centralized, decentralized, and heterogeneous learning settings, training over multiple GPUs, CPUs and/or nodes of a computer cluster will be considered. In terms of federated execution of traditional machine learning and deep learning models, we will consider multiple scenarios, ranging from voting ensemble prediction to cases in which different neural network layers execute on different nodes of a computation network. Federated methods will be evaluated on tasks relevant to the extreme datasets CLOUDMINER use case partners provide.

CLOUDMINER will develop a **chatbot** that will allow users **to explore data before and after processing using Natural Language (NL) conversations**. To query data and the processing results, we need to bridge free NL to a language of controlled concepts, queries, data descriptions, etc. For these purposes, we plan to adapt to this task and build on top of methods for term extraction, named entity recognition, synonym identification, text simplification, and co-reference resolution. For most of these methods, we will need to employ suitable word and sequence representation paradigms, such as embeddings, on top of which we will then define similarity.

1.2.4 STO4: Development of intelligent recommenders for data mining assistance

CLOUDMINER will devise recommender systems to assist users while defining and executing data pipelines. In particular, concerning the development phase, users will be provided with **suggestions** of different items consisting of **reusable fragments** of already specified data pipelines, which are most appropriate for the data at hand, **design best practices**, etc. Existing platforms offer limited support for assisting users while defining data pipelines and reusing existing components. For instance, Figure 4 shows a screenshot of Apache Nifi at work when the user has to select some of the available controller services to be added to the data pipeline under development. The user is supposed to go through the complete list by possibly filtering it using the available tag cloud. However, the shown inventory is independent of the current context, and consequently, even unnecessary components are given.

By relying on the similarity of the project under development with already defined ones, the intelligent assistant of CLOUDMINER will provide **context-aware recommendations** to help developers identify the components that can be employed for the pipeline under development. Similarly, concerning data pipeline executions, the CLOUDMINER recommender systems will suggest **how to distribute the modeled pipelines into the available execution environments** to get the best performance.

Advanced machine learning techniques will be exploited to enable the automated identification of **reusable** data mining components that users can adopt. To retrieve recommendations that are highly relevant for the current context, the envisioned machine learning algorithms will be trained with a **knowledge base** that also needs to be curated in the context of this CLOUDMINER objective. In particular, a monitoring infrastructure will be employed to feed a knowledge base that will be exploited to **optimize** new pipelines by relying on data collected during past execution of other ones.

[30] https://plotly.com/

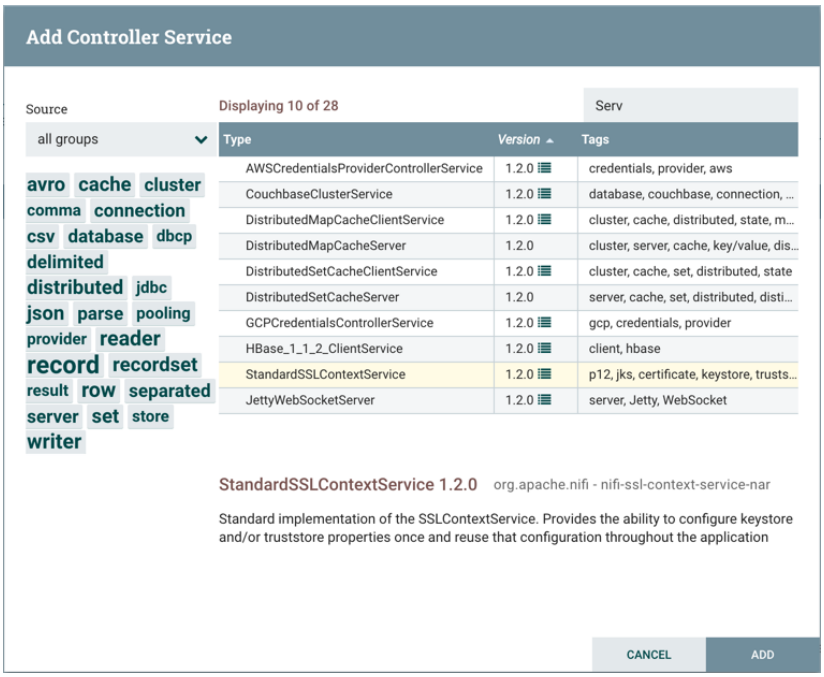

Figure 4: Choosing Controller Services in Apache Nifi

1.2.5 STO5: Development of an optimised data mining execution environment

To support the optimised execution of data pipelines, CLOUDMINER will develop a **pipeline execution environment** that provides a **cloud-native microservice-based architecture**. This execution environment will make use of containers and container orchestration solutions (e.g., Docker and Kubernetes) to facilitate the assembly of polyglot pipelines. CLOUDMINER will also develop **pipeline deployment tools** that will support the underlying definition of deployments. By means of the pipeline deployment tools users will be able to define deployments for the data pipelines created by the CLOUDMINER development environment. Each pipeline will be built as its own image to be deployed by the pipeline execution environment. The deployment tools will also use inputs from CLOUDMINER recommenders to support data engineers in creating optimised pipeline deployments.

To develop the CLOUDMINER execution environment, state-of-the-art technologies will be considered (such as Apache Airflow[31]), and extended to pipeline executions and to support DataOps based on distributed micro-services. Through scalable orchestration, CLOUDMINER will enable the execution of pipelines on a wide range of hardware from local laptops for development purposes to high-performance cloud environments, matching the needs of the user and the stage of the development lifecycle. The CLOUDMINER execution environment will also support DevOps concepts, such as continuous integration and delivery, to automate the testing and deployment of data pipelines. A **web-based deployment dashboard** will be developed as part of the execution environment to visualise the current status of deployed data pipelines as well as monitor them. Furthermore, the pipeline execution environment will provide APIs to enable **remote monitoring and debugging capabilities** from external tools, e.g., the pipeline development environment (see STO2).

1.2.6 Use Cases

The scientific and market objectives of CLOUDMINER have been elicited through **analysis and synthesis of the business needs of four use cases** from the domains of traffic information and road maintenance management, weather forecasting, smart cities, and automotive. The following sections present the domain and ambition of each use case and a discussion on the evaluation methodology and metrics.

1.2.6.1 Use Case 1: Deep Data Mining for Traffic Statistics and Modal Share Monitoring (INFT)

Infotripla is a Finnish company working as a data integrator and service provider in the Smart Traffic, Smart City, and Smart Infrastructure Treating business areas. The company's main customers are cities, transport agencies, infrastructure maintenance contractors, and service providers. Infotripla's solutions are mainly cloud-based information services that consume diverse big data from various data sources (own sensors, open data, and commercial data from business partners). Infotripla is part of Aebi Schmidt Group, a global actor for smart product systems and services for the treatment of mission-critical infrastructural and agricultural areas. Infotripla's main objective in CLOUDMINER is to **develop and improve its information service products** with the help of the results developed during the project. Infotripla will also provide concrete support and contribution to the other project partners by applying and piloting the project results to the Infotripla use case.
Infotripla's use case is related to the cities' and transport agencies' needs to

- create and maintain information about the development of the traffic system's realized service level from a several years perspective (traffic statistics);
- monitor and predict the effects of different measures taken on the transport system, concentrating especially on the effects on modal share (percentage of travellers using a particular type of transportation or number of trips using said type);

[31] Apache Airflow: https://airflow.apache.org/

- The digital solutions related to traffic statistics and modal share monitoring need continuous, high-quality measurement data, e.g., amounts, velocities, timing, locations of vehicles, pedestrians, and bikers. This data is already available and used by Infotripla. However, the data sources are decentralized, and the data quality varies;
- The relevance and customers' interest in the issues mentioned above are constantly growing because customers are facing major challenges due to urbanization, climate change, and energy availability. Therefore, the issues are at the core of Infotripla's service provision also – and especially – in the future to provide information to customers' decision making.

High-quality information for decision making requires **high quality and integrity of data**. Because of the challenges related to the currently available data, there is a significant **need for efficient methods and technologies** e.g., for sorting out **missing data entries, fixing outlier values, and removing duplication**. Infotripla sees a great opportunity to collaborate with the project partners' specialists and get relevant recommendations for the modelling and execution tasks related to the company's objectives. By using the CLOUDMINER's novel data quality management, mining, and analytics technologies, the company anticipates the following benefits:

- to be able to provide **better information for the customers decision making** and so help to meet the business requirements better and to enable company growth;
- **improve the efficiency and quality of the company's internal data management process** with the help of the project's technologies and the project partners' support.

1.2.6.2 Use Case 2: Deep Data Mining for Improved Meteorological Data and Forecasting Services (MLGX)

Kachelmann GmbH currently serves B2B customers of various backgrounds (e.g., agriculture, logistics, insurance, energy sector) and governmental agencies (e.g., flash-flood forecasts via calibrated radar-derived precipitation signals, winter weather services, etc.). It is also developing solutions to make high-quality meteorological data accessible for SMART solutions integration: the currently available weather products on this market are of relatively low forecasting quality. The Kachelmann GmbH has established itself as a high-quality data warehouse and service provider of sophisticated meteorological products that feature high accuracy and high service availability. The Kachelmann Group is interested in expanding its derived product repertoire and is eager to broaden its knowledge by working with interdisciplinary research teams and industries.

The Kachelmann GmbH is part of the Kachelmann Group, and besides Meteologix AG, it is responsible for the group's business with companies and official authorities and produces R&D driven data-based solutions for both B2B and B2C markets. Its main business is developing scientific and innovative weather forecasting tools that are IT powered and research-based, e.g., the invention and operationally running of its full-physics 1x1km weather model (SuperHD) and an urban weather model (UltraHD), radar signal post-processing, storm-tracking tools and the integration of machine learning, and state-of-the-art technologies to solve weather-related problems for their customers.

High-accuracy weather forecasting is important in an increasingly connected and automated world. It can also help raise awareness and initiate safety and preventive actions to mitigate adverse meteorological events, especially within a changing climate. Besides that, Kachelmann Group also operates its weather station network of more than 600 weather stations in central Europe. Furthermore, it runs several freely accessible international B2C weather portals (meteologix.com / kachelmannwetter.com) that aim to gather all high-quality available weather information in one place and make them available and understandable to all citizens by providing access to a vast easily visualized amount of meteorological data, glossaries, info-texts, and explanatory tutorial videos and had more than 70 million visits in 2020. Kachelmann GmbH believes that small active teams help foster innovation and the creative process as less overhead and management is needed, which has often hindered quick implementation in the past, which is why Kachelmann GmbH aims to stay small and flexible with no more than 20 people in its core innovation team.

When it comes to weather forecasting modelling, the concept of ground truth, which is valid meteorological in-situ measurements, is of high importance. Every weather model needs **as much ground-truth as possible** for its so-called initializing fields that build the bases of the 'now' that influences the deterministic forecast modelling of the future hours. Kachelmann GmbH and its team have decades of experience setting up, operating, and quality testing weather stations. During the last 30 years, the weather measurement hardware market has much improved; however, some **important problems do still persist**. One of these problems is **detecting a station whose measurements slowly become unreliable** and out of order due to a plethora of reasons (technical, but also human intervention like new structures nearby). Another issue is **short-term unnatural 'outlier' measurements** due to some usually not weather-related events, e.g., watering sprinklers, wildlife. Both problems significantly impact the weather forecasting quality as they can seriously deteriorate a forecast model's initializing fields and the resulting forecast. Still, these data points themselves are a problem when aggregated in weekly, monthly, and yearly averages, thus significantly biasing the historical database. If these measurement errors have not been detected and removed when entering a meteorological archive, they will remain undetectable. Sometimes this removal of outliers – especially if it is done with the help of too simple algorithms – results in the **removal of true (but special and unusual) signals**, which results in an equally sized data quality problem. Thus, the most reliable state-of-the-art is still to either let the data as is or use automated simple algorithms to 'flag' stations and then in a second step **review them by a meteorologist, which is very costly**, especially for big grid data owners as governmental agencies or Kachelmann Group that has access to 14K weather stations worldwide. Models that can 'learn' patterns and take into account a variety of variables that are both specific and sensitive and thus produce a **minimal false classification** rate would significantly enhance weather in-situ data quality. In a next step, such models could also be used to **predict when deterioration of data quality** from certain sensors is likely to be expected, so maintenance and replacement could be planned proactively. By using the CLOUDMINER technologies, the company anticipates the following benefits:

- **Improved data quality management procedures** for filtering out unreliable weather data without sacrificing unusual but valid data;
- **Efficient and scalable** big weather data processing on-premises and on the public cloud with minimal effort needed to move from on-prem to the cloud due to the standard foundational technologies of CLOUDMINER (e.g., Docker and Kubernetes);
- **Advanced data mining and analytics capabilities** for predictive maintenance/replacement of the weather monitoring stations.

1.2.6.3 Use Case 3: Energy Data for Maia City (EDP)

Municipalities own and manage several infrastructures, such as public offices, schools, social housing, swimming pools, distributed energy power plants, public lighting, electric vehicle (EV) charging stations, and the EV fleet. All these infrastructures house equipment and systems that collect and generate large amounts of data. Unfortunately, in most cases, little use is given to this data due to the disparity between the different formats, sources, and qualities of the data. The Municipality of Maia, located in the north of Portugal, faces these same challenges. Maia is one of the Porto Metropolitan Area municipalities. It is located in the north region of Portugal, in Southwestern Europe. Maia has 135K inhabitants (61K active residents). The City is home of 14 business and industry districts, 1 science park, and more than 15K companies. It is also one of the most industrialized municipalities of Portugal and an important transportation hub. The productive structure of the City is 74% services, 25% industry, and less than 1% agriculture. Maia began seriously paving the way to be a sustainable city in 2012, first by tackling energy issues and in 2014 by creating the Sustainable Energy Action Plan addressing the RES penetration, energy efficiency, CO2 emissions, mobility (including soft mobility, promotion of public transport, e-mobility, among others), citizen's engagement, among others.

Aiming towards the concept of City as a Service, that is, a digital place where stakeholders can interact, gather data, create digital twins, and test businesses, the City of Maia has been developing a data platform. The data on this platform comes from many sources, which then feed a data lake: traffic data systems, RES systems, heating systems in municipal infrastructures such as pools and schools, and parking lots, among many others. Consequently, the source data is heterogeneous and has a high volume and many quality levels. The pre-processing and integration of this data into the data lake is challenging and presents a real opportunity for CLOUDMINER technologies. Besides the heterogeneity of its data sources and formats, Maia's Municipality is also **limited in IT human resources** to carry out the data mining necessary to extract the utility of the data fully.
By using the **low-code distributed data mining facilities** offered by CLOUDMINER, the City anticipates the following benefits:

- Streamline and automate the pre-processing of all data.
- Gain insights into the data by easily feeding it to data visualization tools.
- Improve the environmental performance of the municipality by:
  - Managing and reducing the municipality's energy consumption by identifying possible energy saving and flexibility solutions through the data.
  - Managing and optimising the traffic system lowering fossil fuel consumption, and reducing GHG emissions.
  - Processing information on environmental parameters, such as air quality, pollution, traffic congestion, etc.
  - Enabling innovative energy structures such as energy communities (using P2P) and Virtual Power Plants.

1.2.6.4 Use Case 4: Process Data Mining and Analytics to Achieve Early Fail Detection, Reduce Customer Returns and Reduce Process Cycle Time (CONTI)

Continental develops pioneering technologies and services for sustainable and connected mobility of people and their goods. Founded in 1871, the technology company offers safe, efficient, intelligent, and affordable solutions for vehicles, machines, traffic, and transportation. Currently, Continental employs more than 192K people in 58 countries and markets. With the new Automotive organization of Continental, there are five dynamic, powerful, and flexible action fields Business Areas that consistently take their bearings from Continental's strategic action fields and the development of the global automotive market. Continental Automotive products are produced in more than 45 manufacturing facilities running electronics and mechanics production lines. To ensure product quality, a high number of 100% quality checks are implemented as part of the manufacturing processes. As product and manufacturing traceability is mandatory, all related process results and process details are stored identifiably and assigned to the final single part product. Process details are related to these manufacturing & logistic processes:

- electronics production: solder paste printing, component assembly, in-circuit tests, reflow and wave soldering, functional tests, final tests;
- mechanics production: machining, component assembly, housing, dispensing, leakage test;
- analysis & rework: defect analysis, component replacement, disassembly/assembly;
- logistics: incoming, replenishment, packing, customer delivery, customer return.

For each of the manufacturing locations, 50-200 million process details are collected per day, summing up to 5 billion manufacturing records per day for all Continental Automotive's Business Areas.

At Continental Automotive, manufacturing data is collected on-site in real-time and transferred to a Manufacturing Data Lake using an Edge2Cloud architecture and the services of a leading cloud infrastructure service provider. All manufacturing data is accessible online and has more than 20 years of historical data. Currently, most of the data is used for traceability and KPI reporting. CLOUDMINER technologies will allow to **develop and deploy experimental pipelines quickly**. In an iterative process, improvements can be made based on evaluating the pipeline results within Continental. When the benefit for Continental is proven, the scalable infrastructure of CLOUDMINER can **facilitate an easier transition of the experimental pipeline into productive use**. Additionally, the graphical UI and recommender system of CLOUDMINER allows **participation of non-technical experts**, requiring less coordination with the IT department. Finally, the polyglot microservice-based architecture of CLOUDMINER will enable **reuse and orchestration of components** that were implemented in the scope of other research projects and activities.

By using Cloud technologies and resources there are no technical limitations in using this data for extreme data mining, aggregation and analytics use cases. By means of CLOUDMINER technologies the benefits described below are anticipated.
Reduce number of customer returns: the number of customer return rates at Continental Automotive is already very low due to 100% quality checks and process optimization. Further improvement is a big challenge, but it also has the opportunity for significant savings as well as increasing customer satisfaction. Considering customer return data, including return reasons, serial numbers, combined with all related manufacturing data assigned to all single parts produced and delivered, there is a

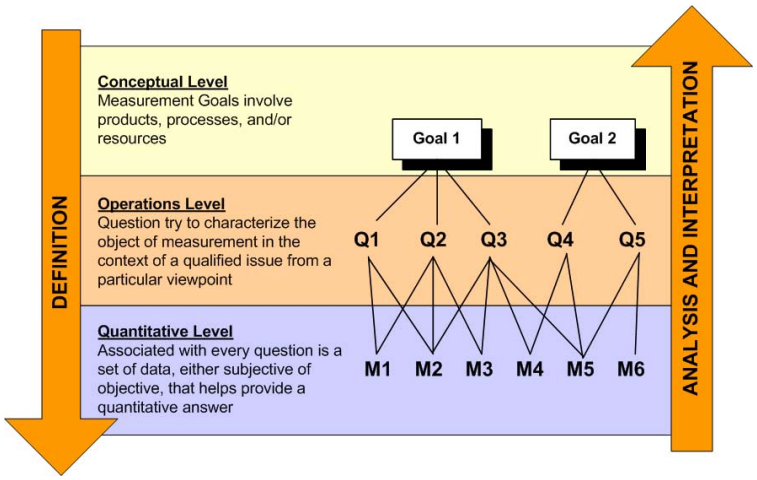

Figure 5: Overview of GQM methodology

very good opportunity to find out the dependencies between collected process details and the probability of a customer return risk for a single individual part, even if all process result values are within the given specification limits.

Reduce process cycle times: As mentioned above, 100% quality checks are implemented to ensure high product and manufacturing quality. Processes with implemented quality checks measure and collect 10-5000 different process or product details for every single part. Physical measurement data acquisition consumes production equipment resources and increases process cycle times. There is a very good opportunity to find out and eliminate redundant data acquisition parameters without reducing the quality check coverage by developing and implementing appropriate algorithms and technologies.

Early failure detection: Similar to the use case "reducing customer returns", there is a very good opportunity to find out dependencies between already collected process details and, later in the production, rejected products. Implementing appropriate algorithms and services is highly possible to prevent further processing for early detected single units. This generates savings by not consuming human, equipment, and material resources. In many cases, this eliminates additional scrap if early detected single units can be reworked instead of the need to scrap them later because rework is not possible anymore.

All data of all CONTI production locations is available via the manufacturing data lake, which can be accessed in the same way for all use cases. Therefore, the three use cases could be addressed within one processing pipeline, with each use case focusing on a different subset of data. To support the project - in addition to the use cases and the huge amount of valuable manufacturing data - Continental Automotive will provide knowledge and resources in manufacturing processes, data collection and Industrial IoT, data ingestion and event sourcing, data engineering and transformation.

1.2.6.5 Use Case Evaluations

The CLOUDMINER project has utilised the approach of goal-oriented measurement as the basis for developing the assessment criteria measures to validate the project technologies in each of the four Use Cases. In particular, the project utilised an adaptation of the GQM method developed by Basili and colleagues[32], which serves as a popular foundation framework for many measurement initiatives. The GQM method was used to define measurements in such a way that:

- Resulting metrics are tailored to the project and its goals.
- Resulting measurement data play a constructive and instructive role in the project and towards the market.
- Metrics and their interpretation reflect the values and the viewpoints of the different groups affected (e.g. developers, users, researchers).

As illustrated in Figure 5, GQM begins by identifying measurement goals (conceptual level) that support (are aligned with) overall project goals. The consortium then poses questions (operational level) to further clarify and refine the goals as well as capture the variation of understanding of the goals that exists among the partners. The consortium then identifies metrics that will provide answers to the questions (quantitative level).

What distinguishes GQM from other measurement paradigms is the hierarchical tree structure used to maintain the relationships amongst goals, questions and metrics. GQM uses a six-step process where the first three steps are about using the overarching project goals and desired industry impact to drive the identification of the right metrics, and the last three steps are about gathering the measurement data and making effective use of the measurement results to drive decision making and improvements. The adaptation of Basili's GQM process utilised by the CLOUDMINER project is as follows:

1. Generate questions that define each of the five primary scientific and technological goals (see STO1-STO5 in Table 2) in a quantifiable way
2. Specify the measures needed to be collected to answer those questions and track process and project conformance to the goals
3. Develop mechanisms for data collection for each of the measures
4. Collect, validate and analyse the data to provide feedback for corrective actions
5. Analyse the data periodically to assess conformance to the goals and to make recommendations for future improvements

[32] Basili, Victor R.; Caldiera, Gianluigi; Rombach, H. Dieter, "The Goal Question Metric Paradigm, Encyclopaedia of Software Engineering (Marciniak, J.J., editor)", Volume 1, John Wiley & Sons

The CLOUDMINER consortium has completed the first steps of the GQM process to arrive at a first set of five quantitative measures for evaluation (see Table 3). Based on previous experience applying GQM for industrial evaluations, the project partners estimate approximately 16-18 measures in total will be needed. For example, for objective *STO2: Design and implementation of a polyglot data pipeline development environment*, GQM output will be a set of measures related to developer productivity and effort reduction. The precise measures and data collection methods for each Use Case will be specified in deliverable D7.3 Evaluation Methodology (M18).

1.2.6.6 Comparative Measures

The evaluation measures the CLOUDMINER project will need to be considered with regard to the relative improvements that were achieved when compared with already existing technologies. Procedures for calculating comparative measures for the CLOUDMINER project evaluations will be structured in one of two different ways:

- *Before and after* – the measures are calculated by undertaking a specific task related to CLOUDMINER capabilities using current technologies and practices, followed by the same task being undertaken using the CLOUDMINER technologies.
- *Side by side* – the measures are calculated by carrying out tasks related to CLOUDMINER capabilities in parallel where one individual or group uses current technologies and practices, while another individual or group uses CLOUDMINER technologies to carry out the task.

The "before and after" approach for comparative measures typically requires fewer resources to carry out the measurement procedure. There is also potentially less variability in the resulting values due to differences in skills and experiences compared to the two individuals or groups used for the "side by side" approach. However, in some cases, the side by side approach is preferred when evaluating a new capability or carrying out a task that is not already familiar or in place within the industrial development organisations as it can provide a more scientific measurement using a control group and experiment group to quantify benefits and advances provided by the CLOUDMINER technologies.

1.2.6.7 Qualitative Measures

There are several improvements provided by the CLOUDMINER technologies that are difficult or complex to quantify, or that have a significant subjective element affected by the industrial domain, capabilities, or previous experience of the individuals carrying out the evaluations (e.g., usability). These improvements are nonetheless important to capture as they can have a substantial impact on whether CLOUDMINER technologies are broadly adopted by industry. Such improvements will be quantified using measures specified for the CLOUDMINER evaluations and through established techniques such as cognitive walkthroughs, think-aloud studies [56], and four point Likert scale questionnaires that employ the "forced choice" method so that evaluators are required to take a position either favourable or unfavourable towards the improvement or advancement provided by the CLOUDMINER technologies.

1.2.7 Links with National and International Research Activities

We have identified the following international research activities with which we are very keen on establishing mutually beneficial links in the context of CLOUDMINER.

DataCloud (https://cordis.europa.eu/project/id/101016835) aims at supporting the development of BigData pipelines to interconnect end-to-end industrial operations of collecting, pre-processing and filtering data, transforming and delivering insights, training simulation models, and applying them in the cloud. The CLOUDMINER advanced analytics techniques for deep data mining (including the multidimensional visualizations, and the chatbots), the envisaged recommender systems, and the scalable micro-service architecture can be of interest to DataCloud; CLOUDMINER can benefit from the work done in DataCloud for developing the low-code development environment;

Confiance.ai (https://www.confiance.ai) is a French national project on trustworthy AI systems. The CLOUDMINER techniques for the extraction and quality management of extreme data can be of interest to this project;

Gaia-X (https://gaia-x.eu) is a software federation system that can connect several cloud service providers and data owners to ensure data exchange in a trusted environment and boost the creation of new common data spaces to create digital economy. Gaia-X can benefit from the micro-service architecture of CLOUDMINER to provide users with data analytics features available on the cloud. Furthermore, CLOUDMINER can benefit from the data available in Gaia-X to train the foreseen recommenders and refine the requirements of the envisioned data mining tools.

Qu4lity (https://qu4lity-project.eu/) is a Horizon Europe project dedicated to Autonomous Qu4lity (AQ) and Zero Defect Manufacturing (ZDM) in Industry 4.0. Among the main objectives, the project is developing technological solutions for the integration and seamless interworking of the digital enablers of the autonomous quality paradigm, including Big Data, AI, Blockchain, Edge/Fog computing devices, and 4G/5G networking technologies. Qu4lity can benefit from the technological offering of CLOUDMINER , especially to investigate the possibility of developing data pipelines to support integration goals promoted by Qu4lity. CLOUDMINER can be inspired by the ways different digital manufacturing components and functionalities are exposed as Open APIs in Qu4lity;

BRAINE (https://www.braine-project.eu/) is a Horizon Europe project on Big data pRocessing and Artificial Intelligence at the Network Edge. CLOUDMINER can be inspired by the BRAINE vision for utilizing edge resources and providing network-edge workload distribution schemes. BRAINE can benefit from the low-code environment of CLOUDMINER to lower the barriers to utilising edge computing for artificial intelligence applications.

1.2.8 Technology Readiness Level

CLOUDMINER aims to produce an integrated technical offering that will reach TRL 6 within the lifetime of the project, and will allow industrial use cases to demonstrate it outside the lab and in relevant industrial environments. Given the concrete objectives of the use cases and the expertise of the respective partners, we expect that the technologies will also be demonstrated within an operational environment (TRL 7) for at least one of the use cases of CLOUDMINER.

1.2.9 Interdisciplinarity Aspects

CLOUDMINER focusses on R&D activities in computer science, in particular in ML, AI, NLP, and data science. The CLOUDMINER end-user services will be designed such that also non-IT experts can efficiently make use of them. With the digital transformation, more and more data scientists and developers are directly employed and closely work together with business analysts on joint client-related R&D projects. In CLOUDMINER, we bring together different target groups addressing the needs of executives, non-IT analysts, data scientists, developers, and researchers. With our CLOUDMINER consortium, we also connect experts from different technical domains, e.g., text/language processing, graph experts, data mining experts together to jointly work on solutions that go beyond applicability solely in the area of computer science. Leveraging different perspectives during the development fuels innovation, fosters collaboration, strengthens relationships, and ensure the broader utility of project outcomes.

1.2.10 Gender Dimension

The consortium is aware of and is committed to actively enforcing the EU's gendered innovation policies and practices. In the context of the technical implementation, the data processing components of CLOUDMINER will be developed by considering gender-specific aspects, to minimise the possibility of spreading gender and racial biases due to lack of diversity in the training of these components. Actions to this direction include (but are not limited to): *i)* the reduction of gender bias in text and NLP analysis when processing gender-neutral terms, *ii)* the consideration of gender-related aspects when building the components for training recommender systems, *iii)* the consideration of gender-related preferences with regards to different code development paradigms, communities, etc. In the context of the use case execution, the consortium will ensure good gender balance in the requirements elicitation activities, design of the pilots, and various pilots and tests of the use case pilots; this will be specifically addressed in WP1 (all tasks). In this way, we will ensure the existence of no gender bias in the gathered requirements and feedback about the pilot test of the CLOUDMINER technologies.

1.2.11 Open Science

The main issue of open science is to make it easier to publish and communicate scientific knowledge. Practices of open science include campaigning for open access to scientific publications and data, providing free and open-source software, and encouraging scientists to practice open-notebook science with the aim of making the entire primary record of a research project publicly available online as it is recorded. These practices should be standard for research completely funded by the public.

In joint research projects with use-case partners, these practices can be in conflict with private sector interests. The use-case partners in this project aspire to gain knowledge that they can use as a competitive advantage over others and to develop new markets. This argues against a full implementation of open science practices. In this project, we will balance the different interests. The scientific results will be published in open access venues. To the extent necessary and appropriate, we will provide the data used to substantiate the scientific results in anonymized form. Data and software collected and developed to strengthen the position of use case partners will remain confidential for an appropriate period of time. Corresponding regulations will be recorded in the consortium agreement. In the interaction between industry and other stakeholders through open science practices, however, it is also important to ensure that the other stakeholders themselves benefit from their creativity and work. The project workplan includes specific actions for implementing open science practices including the following:

Open-source access to results - most of the technology results (to be defined in the scope of the project) will be made available as open-source software and accessible from a public repository (e.g., GitHub). The open-source strategy includes licensing that supports commercial exploitation using well established license terms (e.g., BSD-style licenses such as Apache 2.0, or Eclipse EPL).

Open-access publications - all of the peer reviewed scientific publications of the project will be provided in open access and a specific central budget has been established in the project for any associated fees. Publications will be made available in the project's website, as well as public repositories e.g., OpenAIRE's Zenodo.org European repository.

Open Research Europe Publishing Platform (OREPP) articles - the project will publish a minimum of 3 articles in the OREPP, including one project overview article at the beginning of the project and a concluding article providing a summary of the project's main research achievements at the end of the project.

Open Data Sets - the project includes tasks for Data Management planning and will ensure that research datasets used for benchmarking, testing and other development activities are included as part of the open-source distribution of project results. Selected datasets from the demonstrators will also be anonymised and made available.

Early Scientific Access - the project will utilise online repositories, including e.g., the Computing Research Repository (CoRR) (arXiv.org and others), and the Electronic Colloquium on Computational Complexit (ECCC) to provide the scientific community with early access to the project's results as informal publications.

The partners fully support an open science policy and will ensure the processes and communities supporting the open evolution of the open-source project results continue embrace open science policies as further contributions and innovations are included in the open-source base.

1.2.12 Data and Software Management

To cope with all issues regarding the use of data within the consortium and the potential public release of selected project data and results and based on the guidelines of the EC for FAIR (findable, accessible, interoperable and reusable) data management

in Horizon Europe, a detailed Data Management Plan (DMP) will be formed and delivered as part of the deliverable D9.1 due at M6. Among others, the DMP of the CLOUDMINER project will list the datasets that will be collected, processed or generated by the project, and will specify whether and how each of them will be exploited or made publicly accessible for re-use, and also how it will be curated and preserved during and after the project's lifetime. The DMP will be a living document that evolves during the lifespan of the project, and will be updated as the project progresses with additional datasets. Further details about the data management plan of CLOUDMINER are:
Types of data/research outputs: The work in CLOUDMINER will be associated with: data sources mined during the development of the use cases, datasets for training the recommender systems, the source code of the CLOUDMINER technologies, publications, presentations, video tutorials, interviews and promotional materials.
Findability of data/research outputs: The produced publications, presentations and datasets will be stored in repositories that enable the association of the research results with digital object identifiers (DOIs). In example, the arXiv repository can be used for increasing the findability of preprints of the CLOUDMINER publications, and Zenodo for hosting publications, presentations and datasets. Software and data associated to CLOUDMINER methods will be hosted in trusted repositories, such as GitHub, and links to the webpages of these research outputs will be included in the relevant publications and presentations as well as in selected promotional materials of the project.
Accessibility of data/research outputs: We will follow the "green model" for open access by publishing our work in prestigious venues and archiving the full text in preprint repositories (e.g., arXiv), EU-supported repositories (Zenodo), as well as in our institutional online repositories and the project's website.
Interoperability of data/research outputs: The CLOUDMINER technical components will interchange data using REST APIs and lightweight serialisations, e.g., JSON. Data formats and vocabularies used by the components will be made public and, within the project, data interoperability will be ensured by either using common formats and vocabularies or ensuring mappings are defined between different formats and vocabularies. Software developed will be based on frameworks and programming languages that ensure compatibility with different operating systems.
Reusability of data/research outputs: Datasets and other research outputs will be published with complete documentation and clear usage guidelines. For AI models, we will promote the reporting of Green AI evaluation measures, as in [57]. That is, we will also training/inference times and used infrastructure. Permissive (BSD-style) open-source licenses will be preferred for the released software and data.
Curation and storage/preservation costs: They will be covered by the corresponding consortium members capturing and using the data. The consortium members have experience in data and software management and protection, and will capitalise this experience in the work in CLOUDMINER . The institutional Data Protection Officers (DPO) of the consortium members will support the consortium in these activities. All processing of data in CLOUDMINER will be based on a legitimate ground and informed consent. Last but not least, CLOUDMINER will minimise the processing of personal data to what is absolutely necessary, will anonymise personal data as soon as possible, and will ensure transparency with regard to the functions and processing of personal data.

2 Impact

2.1 Project's Pathways towards Impact

2.1.1 Outcomes and Wider Impacts

Existing graphical data analytics and mining platforms such as Orange, Knime and Apache Nifi **predate container-oriented technologies**, which have arguably revolutionised the development of distributed data-intensive systems in recent years. Therefore they lack the flexibility and scalability offered by container orchestrators such as Kubernetes and Docker Swarm.

CLOUDMINER will **natively embrace container-based technologies** to produce a next-generation cloud-native open-source platform for low-code development and scalable execution of polyglot data analytics and mining pipelines. The modular and polyglot architecture of the CLOUDMINER platform, will enable analysis and mining of large, dispersed and heterogeneous data using best-of-breed technologies both locally/in-house and on commodity cloud computing platforms.

By leveraging robust and widely-used container technologies such as Docker and Kubernetes, adopters will be able to deploy CLOUDMINER pipelines on infrastructure provided by any cloud provider with **minimal configuration effort and lock-in risk**. Furthermore, using the graphical pipeline development and monitoring tools and the intelligent recommendation tools produced in CLOUDMINER , organisations will be able to **involve domain experts** who lack a computer programming background in the design and implementation of data mining and analytics pipelines, thus **releasing precious IT resources** and making **better use of the experts' domain knowledge**. In particular, we foresee the following scientific, economic and societal contributions:

- *Scientific*: Methods and tools for systematic and disciplined development of polyglot container-based data analytics pipelines; a first-of-kind microservice-based graphical development and execution environment; novel caching, incrementality, and parallelism algorithms for high-volume and high-velocity data preprocessing; smart data visualization, scalable and privacy-preserving data mining methods, interactive semantic query answering facilities; real-time recommendations for developing and optimising data pipelines;
- *Technological*: A scalable open-source platform that organisations will be able to extend with domain-specific data preprocessing, mining, analytics and visualisation components implemented in any modern technology;
- *Economic*: Lower IT costs through direct participation of domain experts in the design and implementation of data mining and analytics pipelines; long-lived pipelines that require reduced effort to keep up to date with evolving technologies and frameworks; agility to move deployment to more cost-efficient commodity cloud providers;
- *Societal*: Wider adoption of automated data analytics and mining to support insight extraction, and decision-making processes in the domains of traffic information and road maintenance management, weather forecasting, smart cities, and automotive through the four use-cases of the project.

2.1.2 Scale and Significance of Contributions

Graphical data analytics and mining platforms such as Orange, Knime and Apache Nifi have a substantial user base across different domains and scientific disciplines and have **clearly demonstrated the value of low-code data mining pipelines**, especially for **non-computer scientists**. In CLOUDMINER we wish to **take graphical data analytics and mining platforms to the next level** by enabling polyglot development, intelligent recommendations, scalable execution and offering sophisticated built-in components for data quality management and deep mining.

The polyglot nature of CLOUDMINER aims to **reduce the need for implementing similar data analytics and mining components** in different programming languages and platforms and to enable data scientists and engineers to **focus on novel components** instead. The cloud-native scalability mechanisms that CLOUDMINER will offer on top of containers and container orchestration technologies, will also enable the **execution of low-code data analytics and mining pipelines at an extreme scale** - in stark contrast to current platforms which either support only local pipeline execution or implement basic distribution facilities (compared to state-of-the-art technologies such as Kubernetes). Finally, the intelligent recommendation mechanisms of CLOUDMINER will enable more **domain experts to contribute effectively** to the development of complex data analytics and mining pipelines by helping them discover and connect analytics components that suit the data and tasks at hand in meaningful ways.

2.1.3 Requirements and Potential Barriers

For CLOUDMINER to deliver its expected impact, the methods and technologies that it will develop must see significant adoption by practitioners or influence future open-source or commercial data mining and analytics frameworks. Prior experience suggests that practical adoption from a significant user community can take years to achieve after the end of a RIA project, as the produced software matures, documentation improves and accumulates, and early adopters start reporting successful applications of the project technologies.

To expedite adoption and impact delivery, CLOUDMINER partners will follow a "release early release often" delivery model with an emphasis on producing and publicising (see next section) high-quality public-facing documentation (including video tutorials) for the technologies of the project. Members of the consortium have a long track record of developing widely-used and extensively documented open-source software (e.g. the Epsilon[33] project of the Eclipse Foundation developed by the YORK team) and hands-on experience that they will leverage to this end.

2.2 Measures to Maximise Impact

2.2.1 Dissemination and exploitation of results

The CLOUDMINER partners have prepared an initial plan for the dissemination and exploitation that will maximise the European impact of the project actions. The plan has three primary and interdependent elements:

1. **Targeting of specific markets** – these are the groupings of organisations or individuals who will benefit most from the innovations developed in the CLOUDMINER project.
2. **Dissemination actions** – these initially create awareness amongst the targeted communities and are intended to eventually transform awareness into demand for CLOUDMINER results.
3. **Exploitation actions** – mechanisms that ensure the project innovations will be accessible for the targeted communities on a sustainable basis, providing substantial benefits, and with required supporting structures.

Supporting structures associated with results made available as commercial products includes services such as training, installation and customisation. For results made available as open source, supporting structures would also include communities of interested technology contributors and an open and fair evolution process.

A fundamental strategic decision for exploitation the partners have already taken is to make all of the CLOUDMINER technologies available as an open source distribution.

Each of the elements of the plan is described in the following sections. The CLOUDMINER work plan includes a formal Dissemination and Exploitation Plan deliverable, periodically updated during the project operation, that will provide greater details to the initial plans provided at the time of this proposal.

2.2.1.1 Target Markets

The CLOUDMINER partners have identified three communities for dissemination and exploitation actions, each of which will benefit from the project innovations:

- **Data Scientists and Analysts** will create demand for the project technologies as they become aware of the benefits they provide over existing low-code data mining and analytics platforms. The project will carry out a wide range of dissemination actions to create awareness of the project results amongst this audience, including participation in practitioner-oriented conferences (e.g. EclipseCon, ApacheCon) and frequent release of developer-targeted material (e.g. tutorials, examples, screencasts).
- **Vendors of Data Mining and Analytics Platforms** are an important part of the CLOUDMINER exploitation strategy because their adoption of the project technologies as updates to their existing data mining platforms, or for use in new platforms, is a key channel for the CLOUDMINER results to reach the data scientists and analysts as commercial offers with the associated supporting structures of training, installation, customisation etc. Once the prototypes are developed and validated within the project, the project partners will target dissemination actions specifically towards data mining platform vendors to make them aware of the benefits the project technologies will provide to their customers. The exploitation strategy towards platform providers includes making technologies available in open source so benefits can be readily evaluated.

[33] https://eclipse.org/epsilon

- **Academia and Researchers** communities are important for two reasons: *1)* they currently provide and maintain data mining frameworks that can exploit advances provided by the CLOUDMINER technologies; and *2)* they can contribute to the evolution of the CLOUDMINER technologies via a sustainable open source process that will be established by the project partners during the lifetime of the project. Dissemination actions will target technical journals, conferences and other academic and research events and communications channels. First demonstrations of project results will be made available and liaisons with related EU and nationally funded projects addressing related topics will be established so that the Academia and Research community can make technology contributions to the CLOUDMINER open-source distribution.

The partners have the capabilities and existing contacts to address each of the three major target markets for the project results.

2.2.1.2 Dissemination Actions

The CLOUDMINER project will establish a substantial framework for the dissemination of the project results amongst each of the targeted communities of data pipeline developers, system integrators, and academia and researchers that are involved in the development of or in the delivery of data mining and analytics solutions.

The project dissemination activities are planned to be fully aligned with the needs of the various constituencies that will benefit from the project results. The dissemination actions will include coordination with several important communities:

- **Industry Demonstrators** - on site demonstrators will be established at industrial organisations who are partners in the project and who will deploy within their own organisations the RTD results from CLOUDMINER within an industrial context of data mining and analytics. The experiences will be documented and benefits quantified to showcase the CLOUDMINER results within actual industrial environments in the form of Case Studies to motivate and assist other industrial organizations.
- **Network of Industrial and Government Users** - the project will work with existing international networks of SMEs and large organisations from the target markets for developing data mining and analytics pipelines. Examples of networks to be targeted are Digital Information Hubs, Big Data Value Association (BDVA), Humane-AI-Network, International Data Spaces Association (IDSA), and other data focused communities.
- **Networks of scientific researchers** - the project will establish links with existing formal and informal scientific networks, such as NoEs. In particular, the HiPEAC Network of Excellence is a key target for collaboration as it is organised as it includes more than 250 members from across Europe from a wide range of large Industrial and SME organisations, as well as from leading academic and research centres, and organising regular conferences and opportunities to disseminate results from CLOUDMINER and also gain further inputs from industry and academia.

The dissemination actions already planned within the work programme include the following;

- **Horizontal dissemination material** - Logo, brochures, templates, and other actions that create identity, consistency and awareness of the CLOUDMINER project;
- **Events Participation** - present the main RTD results at appropriate European and International events (outside of Europe) such as Conferences, Fairs, and Congresses; and
- **New University Courses** - the University partners will use the CLOUDMINER results in data science, data analytics, modelling and software engineering courses for students (Automated and Model-Based Software Engineering, Data Science, Cloud Computing), as well as prototype tools for advanced students to contribute to ongoing developments of the CLOUDMINER technologies.

The dissemination actions will create awareness of the project results amongst several different industrial segments both in Europe and internationally.

2.2.1.2.1 Dissemination Approach Wherever possible, research results will be communicated for the creation of external awareness and knowledge building within the targeted European software developer and scientific communities. The communication should guide and prepare potential users for the benefits and improvements that will be made possible by the expected outcomes of the CLOUDMINER project. In order for the dissemination to be effective, several means will be used covering academic and industrial domains: publications in journals and conferences, participation in national and international events, workshops and press releases. Use Cases will be performed that specifically highlight the benefits for the traffic information and road maintenance management, weather forecasting, smart cities, and automotive sectors and for the broader data mining and analytics community, however the dissemination actions will also target a much wider European audience.

2.2.1.2.2 Dissemination in Journals and Conferences There are several different communities that will be highly impacted by the project results, including general communities of practice in data science and analytics. The relevant research communities maintain a large number of international journals and conferences. Some examples of relevant scientific conferences and journals that will be targeted by the project for papers and articles are:

- SIGMOD Record (ACM)
- The VLDB Journal (Springer)
- IEEE Transactions on Knowledge and Data Engineering (IEEE)
- Data Mining and Knowledge Discovery (Springer)
- Software and Systems Modeling (Springer)
- Information and Software Technology (Elsevier)
- Journal of Systems and Software (Elsevier)
- Journal on Data Semantics (Springer)
- IEEE Software and IEEE Computer (IEEE)
- Communications of the ACM (ACM)
- IEEE International Conference on Big Data (BigData)
- ACM SIGKDD International Conference on Knowledge discovery and data mining (KDD)
- ACM International Conference on Information and Knowledge Management (CIKM) IEEE International

Conference on Data Mining (ICDM)
- ACM SIGMOD/PODS Conference (SIGMOD)
- The Annual Meeting of the Association for Computational Linguistics (ACL)
- Empirical Methods in Natural Language Processing (EMNLP)
- The International Conference on Computational Linguistics (COLING)

2.2.1.2.3 Knowledge Management and Protection We plan to disseminate the CLOUDMINER results through:

- Platinum (free) open-access journals
- Platinum (free) open-access workshop proceedings such as CEUR
- Gold open access journals/proceedings: These will be used to disseminate mature research results of CLOUDMINER. The decision on which papers will be funded by the consortium for publication under a gold open access scheme will be made by the Project Board to ensure a fair and effective distribution of the available budget. At least one paper per technical work-package will be published in a gold open access journal by the end of the project.
- Green open access journals/proceedings: These will be used to disseminate interim research results of CLOUDMINER. Such outlets typically enforce an embargo period of 6-24 months, during which, authors maintain the right to share copies of the published manuscripts in their personal websites (but not in institutional/project repositories)
- Open platforms such as arXiv and HAL for all preprints resulting from the project.
- CLOUDMINER's open-source project website and through developer portals such in the form of technical articles targeting mainstream software developers. Project partners have a substantial track record of producing such articles (e.g. www.eclipse.org/epsilon/doc/articles/).

2.2.2 Exploitation Plans

The awareness and documentation of successes and benefits provided by the project gained from the industrial evaluations and Use Cases will increase opportunities for exploitation of the CLOUDMINER results. It is important to note that the CLOUDMINER consortium includes partners that reflect the entire value chain from research through industrial use and standardisation, so the exploitation strategy reflects the differing interests and exploitation requirements of each type of organisation. These include making the project results broadly available as open source, embedding results in commercial IT products, and industrial user partners having a choice of suppliers from which to obtain technologies and implementation support. The CLOUDMINER exploitation strategy includes the following elements that will ensure the RTD results are fully exploited:

- **Framework Legal Organisation** - TOG will provide a common access point for all CLOUDMINER technologies and a framework by which further research and expansion of CLOUDMINER technologies can continue in an open and collaborative manner.
- **Business Network** - In addition to the IT supplier partners in the project consortium, a pan-European network of interested solution providers and suppliers in the relevant domains (e.g. data mining and analytics product vendors, system integrators and users) will be established through project dissemination actions.
- **Licensing and Management** - The project partners will make available under open source license the technologies developed in the project. Facilities will be set up to allow easy access to the RTD results, and specific procedures and support systems will be provided to encourage further contributions to the technologies, verification and assurance of the ongoing integrity of the tool set, and to manage the continued evolution of the CLOUDMINER's software framework.

Exploitation plans are based on a set of guiding principles that are believed to be appropriate for the nature and targets of the project; in particular the consortium adopts the principle that the obstacles for the project adoption must be removed by following an open source model.

Pre-existing know-how or required background technologies, if any, will be made available on a royalty-free basis during the project and on favourable conditions after the project. Knowledge of foreground generated during the project will be made available on a royalty-free basis during and after the project. Specific tasks are included in the workprogramme to ensure accurate preparation of achievable exploitation plans, which are of vital importance to the project, are prepared and will ensure that the results of the project will not remain confined in research labs, but will follow a clear and rapid path to market.

The exploitation plan for the project is centred on an open-source distribution strategy Three key elements are essential for achieving broad industry adoption of open-source technologies will be put in place by the partners:

- **Industrial community** – dissemination and collaboration actions will be used to build a community of stakeholders surrounding the open source technologies developed within the CLOUDMINER so that organisations beyond the project partners exploit for their own benefit, while a subset will also contribute to the evolution of the open source technologies.
- **Accessibility** – project partners will ensure technologies are readily accessible to all of the communities of interest that could potentially benefit or exploit the project results for data driven applications and services, and under terms and conditions that are easy to understand and encourage further innovation.
- **Open evolution** – an open, transparent and participatory process will be established for determining the evolution of the open-source project technologies so that all stakeholders have a voice in determining which contributions will be included in future updates to the open-source deep data mining technology base.

The open-source exploitation strategy also facilitates rapid exploitation by each of the project partners as IPR, commercial arrangements, and potential joint ownership issues are avoided. Individual partner exploitation plans will be detailed in the project deliverables, but an overall summary of The exploitation actions planned by the project partners are summarised below by type of partner:
University partners (YORK, UDA, MUN, EHU, IPM)

- Extend the capabilities of the project results with features supporting additional types of data sets and mining capabilities.
- Undertake further research addressing next version features identified in the final stages of the project during the industrial Use Case evaluations and extend the project technologies to address further data domains.

- Research various models, algorithms, meta-information, and AI methods for improving performance, scale, complexity and other mining aspects of the CLOUDMINER technologies and contribute advances to the open-source base.
- Utilise existing technology transfer programmes to make project technologies available to industrial actors within their region.
- Extend data sciene teaching curricula to include methods and tools for designing and developing complex data mining pipelines.

Research partners (ATB, CEA)

- Increase the maturity of the technologies developed within the project to make them available for projects and collaborations with industry partners.
- Provide supporting services to assist organisations to utilise CLOUDMINER technologies for advanced data mining and provide industrial support for open-source adoption and deployments.
- Conduct promotional and dissemination actions to create further awareness of the new project techniques within national and regional communities of stakeholders.
- Undertake further development of CLOUDMINER technology as enhancements and updates to provide improvements and additional features needed by industrial research partnerships.
- Expand the project's deep data mining technologies through further nationally or EU funded projects and contribute to the open-source base.

Industrial data driven partners (EDP/CMM, CONTI, INFT, MLGX)

- Utilise CLOUDMINER technologies for development of new data driven applications and services in their respective industrial domains.
- Evaluate the ability of the CLOUDMINER technologies to support additional types of data sets and mining configurations beyond those addressed within the project.
- Undertake internal awareness actions to propagate the use of CLOUDMINER technologies within other parts of their organisation for developing data-driven applications and services.
- Encourage their suppliers and other data-driven business partners to utilise CLOUDMINER technologies to achieve similar improvements in data mining efficiency, productivity and data driven services.
- Monitor development and business improvements achieved from using CLOUDMINER data mining tools and technologies and share these successes at industry level events.

As an industry-sponsored standards organisation with over 800 members including many of the largest multinationals from energy, manufacturing, finance, retail, healthcare and ICT sectors, as well as many smaller organisations providing applications and services, the exploitation actions for TOG focus on providing the infrastructure and resources to support the open-source distribution of the project results, open evolution process, and to build industry consensus for the proposed new standards based on the project results.

2.2.2.1 Dissemination and Exploitation Towards Additional Domains

The project will verify the new CLOUDMINER technologies through real-life case studies in the traffic information and road maintenance management, weather forecasting, smart cities, and automotive domains. However, the consortium fully expects the technology will be applicable and provide similar benefits for data mining and analytics pipeline development in many other domains including automotive, manufacturing, utilities, transportation, and many more. These industries face very similar challenges as the traffic information and road maintenance management, weather forecasting, smart cities, and automotive domains in terms of their ability to analyse and mine extreme data in a scalable manner. The dissemination activities that will be carried out will include actions towards these other domains. For example, the existing international networks of SMEs and large organisations with which collaboration will be established, will include users from automotive, manufacturing, transportation in addition to the first priority targets of traffic information and road maintenance management, weather forecasting, smart cities, and automotive domains for the project. The partners will also participate and present the project results at Conferences, Fairs, and Congresses where other domains are well represented. TOG, INFT, MLGX, EDP, and CONTI are active in these other domains and have strong capabilities to carry out dissemination and exploitation actions that target organisations in domains beyond traffic information and road maintenance management, weather forecasting, smart cities, and automotive.

The project will set as a first priority dissemination and exploitation towards the traffic information and road maintenance management, weather forecasting, smart cities, and automotive domains, as the case studies are expected to provide compelling first-hand data demonstrating the improvements that are possible from the CLOUDMINER tools and processes. As a second, but also high priority the project will disseminate and position the project results as important improvements that are applicable to data mining and analytics in other domains.

2.2.2.2 IPR Management

The main results of the project are software tools, methodologies, and development processes, which will be developed by the project partners and under the terms and conditions of a Consortium Agreement, established before signing of the EC contract. The Consortium Agreement will include the identification of background modules or technologies and will describe in detail the mechanisms for protecting partners intellectual property while assuring project results will be readily available under an open source licence to industry.

As a governing principle the knowledge or result shall be the property of the partner generating it, and each partner as part of the entering into the Consortium Agreement, agrees to ensure no encumbrances would be introduced that would inhibit the availability of the project results under an open source license.

A joint invention, design or work shall arise if, in the course of carrying out work on the Project, at least two partners contribute to an invention design or work with the result that it is not possible to separate them for the purpose of applying for, obtaining

and/or maintaining in force the protection of the relevant intellectual property right, the partners concerned (the "Contributors") agree that they may jointly apply to obtain and/or maintain the relevant rights and shall strive to set up amongst themselves appropriate agreements in order to do so. Such Contributors shall be entitled to use or license any joint invention, design or work, without owing any financial compensation to or requiring the consent of any other Contributors, provided always that such use or licence does not directly or indirectly compete with the current business of any Contractor, unless the latter expressly consents.

In the case where a partner ("Originator") would decide in its sole discretion that it does not intend to seek adequate and effective protection of certain of its Knowledge from the Project, then, the Originator shall inform in writing the other Contractors, through the Coordinator, and any partner interested in applying to obtain and maintain such protection shall advise the other partners through the Coordinator and in writing within one month of receipt of relevant notice. In case several Contractors are interested in so applying, they shall strive to set up amongst themselves and with the Originator appropriate agreements in order to do so. The consortium will comply with IPR rules under Horizon Europe as specified by the Horizon Europe Model Grant Agreement and in conformance with guidelines provided in the Annotated Model Grant Agreement.

2.2.2.3 Raising Public Participation Awareness

The consortium will raise public participation and awareness by keeping the project web site open to everybody upon registration and by providing discussion forums and opportunities to exchange information and provide guidance to the project. As part of the dissemination and exploitation strategy for the project the partners will establish a community of software developers who focus on data mining and analytics systems. This community will provide inputs and guidance in the early stages of the project and will participate in the evolution and further expansion of the tools towards the latter stages using an open source approach. The open source approach along with the establishment of a broad European community of software developers and data scientists involved in the project technologies will substantially raise the public awareness and participation in the project. The project will also target non-technical people by having articles published in large circulation newspapers.

2.2.3 Communication Activities

In addition to the dissemination materials, events and publications described above, the project will undertake a Public Communications Programme to ensure broad awareness of the project across the data mining and analytics and wider ICT community. The major elements of the public communications plan include the following:

- **Initial international press release** announcing the launch of the CLOUDMINER project supported by the EC and the expected impact it will have on data mining and analytics and supporting tools and technology. This will be released simultaneously to all major technical press and journals in Europe, USA and Asia.
- **Public website** with introductory information about the project and public deliverables as they become available (e.g. technical journal paper, specifications proposed as standards, etc.). The website will be regularly updated and will include an RSS based news feed to alert those tracking the project when new information concerning the project, upcoming events, training, etc. are available.
- **Social media** accounts (e.g. Twitter, Facebook, YouTube) which will enable interactive communication between the partners of the project and big-data practitioners and researchers.
- **Interim international press releases** announcing the completion of key milestones for the project. These are planned for the following:
  - First technology results available with completion of the CLOUDMINER platform and methodology
  - First industrial use of CLOUDMINER technology being validated in four Use Case Demonstrators for traffic information and road maintenance management, weather forecasting, smart cities, and automotive
- **International standardisation press release** announcing actions towards new standards resulting from the project. This may be the announcement of a new specification submission to an existing grouping, formation of a new working group (e.g. under OMG or other standards body), or public availability of a specification.
- **Final international press release** announcing the availability of CLOUDMINER technologies, the evaluation results from the four Use Case Demonstrators, and how to access the project results.

Further communications will be undertaken in association with specific events where CLOUDMINER will participate and in collaboration with specific journals where CLOUDMINER papers and articles will be published.

2.2.4 Contributions to standards

Commitment to open standards is a fundamental principle underlying all of the CLOUDMINER RTD activities. The work plan includes project tasks focused on monitoring and maintaining alignment with existing technology standards used within CLOUDMINER . More importantly, CLOUDMINER intends to establish new standards and extensions to existing ones. The technology-independent API for design-time integration of polyglot data mining pipeline components (WP3), and the real-time data mining pipeline execution monitoring protocol (WP4) have been identified as the potential contributions of CLOUDMINER to standards.

Specific action plans to create awareness, garner industrial support, and build consensus will be established for each of the above standards contributions and undertaken through partners participating in the respective standard bodies. CLOUDMINER includes a standards organisation (TOG) as work package leader for standardisation-related tasks who collaborates with many other standards bodies and fora around the globe. TOG has the ability to build industry consensus and publish industry standards and also has close collaborations with other standards bodies including IEEE, ISO, W3C, and in particular, OMG where the project intends to make multiple proposals for adoption of CLOUDMINER technologies as new OMG industry standards. TOG and OMG share joint membership and TOG operates industry certification programmes on behalf of OMG. TOG publishes standards, provides certification and branding of products as conformant to standards, and conducts public "PlugFest" and similar events for key industry standards.

2.3 Summary Canvas

KEY ELEMENT OF THE IMPACT SECTION

SPECIFIC NEEDS

- **Ensuring high-quality data** since data under analysis can come from different sources and have different formats.
- **Platform-independent data-mining** as people working with data are less interested in learning new technologies and tools
- **Facilitating advanced data analytics components** such as advanced multidimensional visualizations and deep data mining techniques for distilling meaningful patterns from large, heterogeneous, and dispersed data sources.
- **Providing assistance during data mining** to cope with the complexity of data mining processes and to both flag possible inappropriate choices and provide recommendations during the entire data science process.
- **Optimizing the execution of data pipelines** using distributed environments that support scalable and efficient model training and usage and smart data pipelines autonomously learning from past executions to optimize how new sparse and heterogeneous data have to be retrieved and mined.

EXPECTED RESULTS

- **Data quality management system** to support and optimize pre-processing of high-volume, high-variety, and high-velocity data.
- **Polyglot data pipeline development environment** to enable data scientists to model data pipelines in a platform-independent manner and to facilitate the specification of data mining processes by means of high-level abstractions by providing users with sets of predefined patterns limiting errors, reducing programming time, and facilitating resource exploitation.
- **Advanced analytics techniques for deep data mining** to facilitate, by means of smart data visualization and chatbot facilities, the discovery of meaningful, reliable and useful patterns from large, heterogeneous, and dispersed data sources.
- **Intelligent recommenders for data mining assistance** to provide assistance during the whole data mining process by providing users with recommendations that are relevant for the modelling and execution tasks at hand.
- **Optimised data mining execution environment** to facilitate the deployment and optimized execution of modelled data mining and analytics pipelines by means of a scalable microservice-based architecture.
- **Multiple Industrial Evaluations and Demonstrators** across key data-driven European sectors of Energy, Manufacturing, Smart Cities and weather forecasting for Smart Transport.

D & E & C MEASURES

Key Exploitation elements

- Open source distribution of all project technologies
- Community building around open source amongst a wide range of data science stakeholders
- Sustainable, transparent and open evolution process for the open source project technologies
- Commercial deployments in industrial environments by demonstration partners
- Standardisation of key project technologies

Key Dissemination elements

- Industrial Demo Centres to showcase project technologies in data-driven industries and smart cities contexts
- Network of Industrial and Government Users exploiting the open-source technologies for internal data mining tasks and provisioning of applications and services
- Scientific publications at conferences and in journals with completion of early and full prototype technologies
- Collaboration with Digital Information Hubs, Big Data Value Association, Humane-AI-Network, International Data Spaces Association and other data focused communities for knowledge sharing and support.
- New University courses ensuring new data scientists have required mining tools knowledge and technologies access.

Key Communication elements

- International press releases at major milestones
- Website, brochure, poster, presentation
- Social Media including messaging, podcasts, videos

Figure 6: Key Elements of the Impact section - Part 1

TARGET GROUPS

Stakeholders from multiple industrial sectors, as well as governmental agencies and academic/research communities focused on data sciences and data-driven applications and services including:

- **Data Scientists and Analysts –** create demand for the project technologies as they become aware of the benefits they provide over existing low-code data mining and analytics platforms.
- **Vendors of Data Mining and Analytics Platforms –** an important part of the exploitation strategy because their adoption of the project technologies as updates to their existing data mining platforms, or for use in new platforms, is a key channel for the CLOUDMINER results to reach the data scientists and analysts as commercial offers with the associated supporting structures of training, installation, customisation, and technical assistance.
- **Academia and Researchers communities –** important for two reasons: 1) they currently provide and maintain data mining frameworks that can exploit advances provided by the project technologies; and 2) they can contribute to the evolution of the CLOUDMINER technologies via the sustainable open source process that will be established by the project partners during the lifetime of the project.

OUTCOMES

- CLOUDMINER will natively embrace container-based technologies to produce a **next-generation cloud-native open-source platform for low-code development and scalable execution of polyglot data analytics and mining pipelines**.
- Modular and polyglot architecture of the CLOUDMINER platform will enable **analysis and mining of large, dispersed and heterogeneous data using best-of-breed technologies both locally/in-house and on commodity cloud computing platforms**.
- Project technology adopters will be able to **deploy CLOUDMINER pipelines on infrastructure provided by any cloud provider with minimal configuration effort and no lock-in risk**.
- CLOUDMINER **graphical pipeline development** and **monitoring tools** and the **intelligent recommendation tools** that enable organisations to **involve domain experts who lack a computer programming background in** the design and implementation of **data mining and analytics pipelines**.
- Data-driven organisations able to **use advanced data mining and analysis tools to release precious IT resources and making better use of their domain knowledge of experts** within the organisation.

IMPACTS

Scientific
Methods and tools for systematic and disciplined development of polyglot container-based data analytics pipelines; a first-of-kind microservice-based graphical development and execution environment; novel caching, incrementality, and parallelism algorithms for high-volume and high-velocity data pre-processing; smart data visualization, scalable and privacy-preserving data mining methods, interactive semantic query answering facilities; real-time recommendations for developing and optimising data pipelines.

Technological
A scalable open-source platform that organisations will be able to extend with domain-specific data pre-processing, mining, analytics and visualisation components implemented in any modern technology.

Economic
Lower IT costs through direct participation of domain-experts in the design and implementation of data mining and analytics pipelines; long-lived pipelines that require reduced effort to keep up to date with evolving technologies and frameworks; agility to move deployment to more cost-efficient commodity cloud providers.

Societal
Wider adoption of automated data analytics and mining to support insight extraction, and decision-making processes in the domains of smart cities, automotive, traffic information and road maintenance management, and weather forecasting through the four use-cases of the project.

Figure 7: Key Elements of the Impact section - Part 2

3 Quality and Efficiency of the Implementation

3.1 Work Plan and Resources

This section introduces the work plan structure and discusses the work packages that comprise CLOUDMINER.

3.1.1 Components and Interdependencies

Figure 8 provides a graphical representation of the technical work packages of the project as well as their main products and interactions, which are further discussed in Sections 3.1.4.2-3.1.4.6.

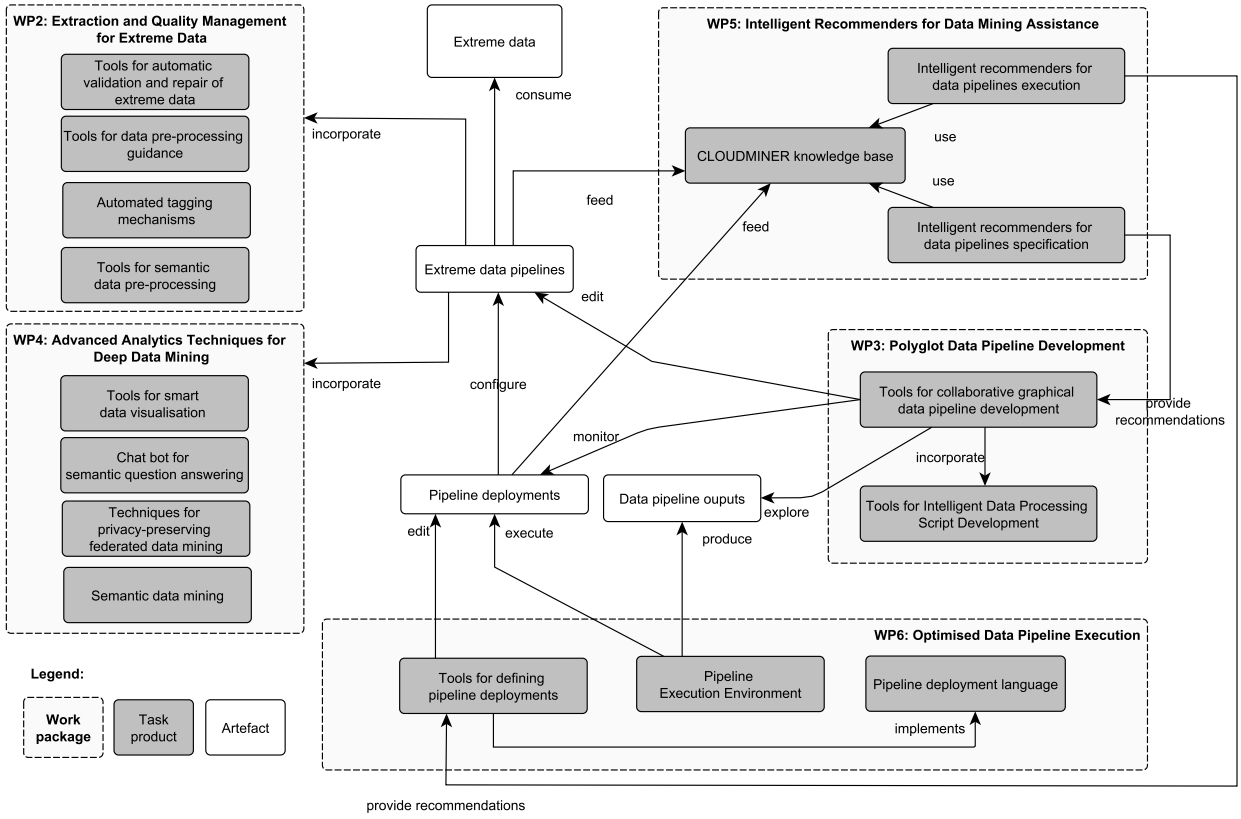

Figure 8: Task Products and Dependencies

3.1.2 Project Planning - Timeline and Effort Distribution

The project duration will be 36 months, and the project will be articulated in the work packages listed in Table 4. Note that the work breakdown into work packages is based on gathering related work, rather than on gathering tasks that occur at around the same time. Therefore, most of the work packages run in parallel throughout the project. The development within the project will be based on frequent iterations and early prototypes/extensions in order to ensure that we are building the right tools the right way.

Table 4: Work Packages

WP#	Title	Lead Participant No	Lead Participant Name	Person months	Start month	End month
WP1	Requirements and Use Cases	1	TOG	50	1	6
WP2	Extraction and Quality Management for Extreme Data	4	CEA	68.5	1	30
WP3	Polyglot Data Pipeline Development	3	YORK	71	1	30
WP4	Advanced Analytics Techniques for Deep Data Mining	5	MUN	48.8	1	30
WP5	Intelligent Recommenders for Data Mining Assistance	2	UDA	31.5	1	30
WP6	Optimised Data Pipeline Execution	7	ATB	42.5	1	30
WP7	Integration, Deployment and Industrial Evaluations	8	CLMS	172	1	36
WP8	Ecosystem, Dissemination, Exploitation and Standardisation	1	TOG	55	1	36
WP9	Project Management	1	TOG	33.8	1	36
	Total			573.1		

The Gantt chart in Table 5 also shows the main deliverables within each of the work packages. Note that much of the work will be carried out continuously: the major deliverables act as checkpoints within the WPs to ensure synchronisation.

Table 5: Project Gantt Chart

	Year 1												Year 2												Year 3											
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
	M1 →						M2 →						M3 →						M4 →						M5 →						M6 →					
WP1						1.1 1.2																														
T1.1																																				
T1.2																																				
T1.3																																				
WP2												2.1						2.2												2.4 2.5						
T2.1																																				
T2.2																																				
T2.3																																				
T2.4																																				
WP3												3.1						3.2 3.3						3.4						3.5						
T3.1																																				
T3.2																																				
T3.3																																				
T3.4																																				
T3.5																																				
WP4												4.1						4.2 4.3												4.4						
T4.1																																				
T4.2																																				
T4.3																																				
T4.4																																				
WP5												5.1						5.2												5.3						
T5.1																																				
T5.2																																				
T5.3																																				
WP6												6.1						6.2 6.3						6.4						6.5						
T6.1																																				
T6.2																																				
T6.3																																				
WP7						7.1												7.2 7.3			7.4 7.5 7.6 7.7									7.8						7.9 7.10 7.11 7.12
T7.1																																				
T7.2																																				
T7.3																																				
T7.4																																				
T7.5																																				
T7.6																																				
T7.7																																				
WP8			8.1			8.2								8.3										8.4												8.5 8.6
T8.1																																				
T8.2																																				
T8.3																																				
T8.4																																				
T8.5																																				
WP9			9.1			9.2			9.3									9.4						9.5												9.6
T9.1																																				
T9.2																																				
T9.3																																				
T9.4																																				
T9.5																																				
	M1 →						M2 →						M3 →						M4 →						M5 →						M6 →					

3.1.3 Deliverables List

Table 6 lists all the project deliverables. Deliverables are numbered Dw.s, where w is the work package number, and s is the deliverable sequence number within the work package. Deliverables of type R are reports, while those of type S consist of software and an accompanying report. Deliverables of type DEC are dissemination and communication artefacts. Note that some deliverables may be revised after their official delivery date based on practical experience in using them to produce other results. In such cases, the effort allocated to the corresponding work package will be used for effecting such changes, and the revised deliverable will be made available on the same terms as the officially delivered version. Further details of the content of the individual deliverables are given in the lists of deliverables in the work package descriptions in the sections that follow.

Table 6: Deliverables by chronological order

ID	Title	WP	Partner	Type	Dissem. level	Delivery date
D8.1	Project Website	WP8	TOG	DEC	PU	3
D9.1	Quality Management Plan	WP9	TOG	R	RE	3
D1.1	Project Requirements	WP1	TOG	R	RE	6
D1.2	Evaluation Plan	WP1	TOG	R	RE	6
D7.1	Architectural Guidelines Report	WP7	CLMS	R	PU	6
D8.2	Project Presentation and Brochure	WP8	TOG	DEC	PU	6
D9.2	Data Management Plan	WP9	TOG	R	RE	6
Milestone 1: Requirements and Case Studies Completion (M6)						
D9.3	1st Interim Project Report	WP9	TOG	R	RE	9
D2.1	Automatic validation and repair of extreme data	WP2	CEA	R	PU	12
D3.1	Graphical Data Pipeline Development Tools	WP3	YORK	S	PU	12
D4.1	Smart Data Visualisation Tool	WP4	MUN	S	PU	12
D5.1	The CLOUDMINER Knowledge Base for Intelligent Recommenders	WP5	UDA	R	PU	12
D6.1	Data Pipeline Deployment Language	WP6	ATB	R	PU	12
Milestone 2: CLOUDMINER Languages Definition Completion (M12)						
D8.3	Initial Dissemination and Use Plan	WP8	TOG	R	RE	14
D2.2	Data preprocessing task interoperability and user guidance	WP2	EHU	R	PU	18
D3.2	Intelligent Data Processing Script Development Facilities	WP3	YORK	S	PU	18
D3.3	Natural Language Interaction as a Pipeline Design Aid	WP3	EHU	R	PU	18
D4.2	Privacy-preserving Federated Data Mining Tool	WP4	EHU	S	PU	18
D4.3	Semantic Data Mining Tool	WP4	MUN	S	PU	18
D5.2	The CLOUDMINER Intelligent Recommender Systems for Data Pipelines Specification	WP5	UDA	S	PU	18
D6.2	Data Pipeline Deployment Tools	WP6	ATB	S	PU	18
D6.3	Data Pipeline Execution Environment - Initial Version	WP6	ATB	S	PU	18
D7.2	Integrated Platform - Initial Version	WP7	CLMS	S	PU	18
D7.3	Evaluation Methodology	WP7	TOG	R	PU	18
D9.4	1st Periodic Project Report	WP9	TOG	R	RE	18
Milestone 3: CLOUDMINER Platform and Methodology - Initial Version (M18)						
D7.4	Deep Data Mining for Traffic Statistics and Modal Share Monitoring Use Case Evaluation - Interim Version	WP7	INFT	R	RE	21
D7.5	Deep Data Mining for Improved Meteorological Data and Forecasting Services Use Case Evaluation - Interim Version	WP7	MLGX	R	RE	21
D7.6	Energy Data for Maia City Use Case Evaluation - Interim Version	WP7	EDP,IPM,CMM	R	RE	21
D7.7	Process Data Mining and Analytics to Achieve Early Fail Detection, Reduce Customer Returns and Reduce Process Cycle Time Use Case Evaluation - Interim Version	WP7	CONTI	R	RE	21
D3.4	Graphical Pipeline Monitoring and Debugging Extensions	WP3	YORK	S	PU	24
D6.4	Data Pipeline and Recommender Transformation Tools	WP6	ATB	S	PU	24
D8.4	Press and Media Materials	WP8	TOG	DEC	PU	24
D9.5	2nd Interim Project Report	WP9	TOG	R	RE	24
Milestone 4: CLOUDMINER Platform and Methodology - Interim Version (M24)						
D2.4	Semantic data preprocessing	WP2	EHU	R	PU	30
D2.5	Data preparation and quality	WP2	CEA	S	PU	30
D3.5	Pipeline Differencing, Conflict Detection and Merging Tools	WP3	YORK	S	PU	30
D4.4	Semantic Query Answering Tool	WP4	EHU	S	PU	30
D5.3	The CLOUDMINER Intelligent Recommender Systems for Optimizing Data Pipelines Execution	WP5	UDA	S	PU	30
D6.5	Data Pipeline Execution Environment - Final Version	WP6	ATB	S	PU	30
D7.8	Integrated Platform - Final Version	WP7	CLMS	S	PU	30
Milestone 5: CLOUDMINER Platform Ready for Evaluation (M30)						
D2.3	Automated tagging of exteme data	WP2	CEA	R	PU	31
D7.9	Deep Data Mining for Traffic Statistics and Modal Share Monitoring Use Case Evaluation - Final Version	WP7	INFT	R	RE	36
D7.10	Deep Data Mining for Improved Meteorological Data and Forecasting Services Use Case Evaluation - Final Version	WP7	MLGX	R	RE	36
D7.11	Energy Data for Maia City Use Case Evaluation - Final Version	WP7	EDP,IPM,CMM	R	RE	36
D7.12	Process Data Mining and Analytics to Achieve Early Fail Detection, Reduce Customer Returns and Reduce Process Cycle Time Use Case Evaluation - Final Version	WP7	CONTI	R	RE	36
D8.5	Final Dissemination and Use Plan	WP8	TOG	R	RE	36
D8.6	Standardisation Report	WP8	TOG	R	RE	36
D9.6	Final Project Report	WP9	TOG	R	RE	36
Milestone 6: Project completion (M36)						

3.1.4 Work package description

3.1.4.1 WP1: Requirements and Use Cases

Work package	1	**Start month**	1	**End month**	6
Work package title	Requirements and Use Cases				
Participant name (number)	**TOG (1)**	UDA (2)	YORK (3)	CEA (4)	
Person-months	**5**	4	4	4	
Participant name (number)	MUN (5)	EHU (6)	ATB (7)	CLMS (8)	
Person-months	4	4	4	4	
Participant name (number)	INFT (9)	MLGX (10)	EDP (11)	IPM (12)	
Person-months	3.5	3.5	3	1	
Participant name (number)	CMM (13)	CONTI (14)			
Person-months	2	4			

Objectives: Within this work package the requirements will be defined for the new tools, technologies and processes proposed by CLOUDMINER, driven from both an industry and technology perspective. The use cases provided by the industrial user partners in the project will be analysed and industry driven requirements for the project will be defined and prioritised. The technology partners will further detail technical requirements that address the industrial user needs within the use cases and fulfill the technological breakthroughs targeted by the project. Requirements will be established for each of the technology areas within the development work packages to guide the technology partners during the research and development within the project. The requirements will be later used during industrial evaluations to validate the results of the project and as one of several measures to indicate the degree to which the project delivers the expected impact and benefits for industry.

Description of Work:
Task 1.1: Use Case Analysis (TOG, UDA, YORK, CEA, MUN, EHU, ATB, CLMS, EDP, IPM, CMM, CONTI, INFT, MLGX) The industrial use cases will be specified by each of the industrial partners, analysed to establish a coherent structure and categorisations for the requirements set, followed by the extraction of individual requirements in each category. The requirement set will be prioritised and reviewed to ensure industrial driven requirements are established for each technology area being addressed by the development work packages.
Task 1.2: Technology Analysis (TOG, UDA, YORK, CEA, MUN, EHU, ATB, CLMS, EDP, IPM, CMM, CONTI, INFT, MLGX) In this task the technology partners will define the ultimate objectives for all technology areas and the desired breakthroughs to be achieved in alignment with the industrial user requirements established in the previous task. The result of the technology analysis task will be further detailed technical requirements that must be achieved to fulfill the industrial requirements and provide the expected benefits to European developers of data mining and analytics software systems.
Task 1.3: Evaluation Planning (TOG, UDA, YORK, CEA, MUN, EHU, ATB, CLMS, EDP, IPM, CMM, CONTI, INFT, MLGX) This is the first of two tasks that will establish the detailed evaluation methodology to be carried out during the project operation. This initial task will address in more detail the outlined methodology (described in Section 1.2.6.5) and specifies for each industrial Use Case the evaluation scenarios with regard to the structure, volume, and consistency, availability and partitioning requirements of the data involved, as well as the expected stakeholder participants. Specific success criteria and performance indicators will be established related to the impacts targeted by the project. These will be used later in the project in establishing the detailed measurement methods and metrics that will be defined in work package 6.

Partners Roles:
TOG will lead the work package and tasks taking into account requirements from the industry focused project partners, the research and development partners and from it's over 400 members from the ICT user and supplier community.
ATB, CEA, CLMS, EHU, MUN, YORK, and UDA will contribute to establishing the technical requirements that will drive the research and development in each of the work packages 2–6, which they lead or contribute.
INFT, MLGX, EDP, and CONTI will specify industry focused Use Cases that represent the needs within the specific domain they represent, which will form the basis for the development and prioritisation of industrial user requirements that will drive the project development and form the basis for evaluations later in the project.

Deliverables:
D1.1: Project Requirements (TOG) (M6) Overall requirements specification for the project detailing the needed industrial data mining and analytics capabilities to be addressed within each of the other development work packages. This report will include the specifications of the use cases, the associated requirements for each of the targeted industry domains, specifications of the data mining activities that will be targeted in each domain, and specifications of user requirements for each CLOUDMINER technology component. Requirements will be categorised and prioritised in order to guide the research and development in the other work packages.
D1.2: Evaluation Plan (TOG) (M6) A report describing the targeted demonstrators for each domain, which will be representative of each industry and used to validate the CLOUDMINER technologies. The deliverable will include

an overview of the demonstrator usage and industrial context, the processes addressed by each demonstrator, the CLOUDMINER technologies targeted for validation, the platform to be utilised and the measures that will be monitored to assess the impact of CLOUDMINER technologies in supporting decision making using data mining and analytics pipelines.

3.1.4.2 WP2: Extraction and Quality Management for Extreme Data

Work package	2	**Start month**	1	**End month**	30
Work package title	Extraction and Quality Management for Extreme Data				
Participant name (number)	UDA (2)	**CEA (4)**	MUN (5)	EHU (6)	
Person-months	3	**43**	8	9	
Participant name (number)	CLMS (8)	IPM (12)			
Person-months	3.5	2			

Objectives: This work package aims to facilitate preprocessing of extreme data, particularly the extraction, cleaning, and transformation of such data. The proposed work will focus on two orthogonal dimensions. First, it will propose novel and scalable tools and techniques that can be used to automate data preprocessing activities. Secondly, it will facilitate the integration of heterogeneous datasets and data preprocessing tasks in data mining pipelines using care labels. To facilitate scalability of CLOUDMINER's data preprocessing components, the proposed technologies will support *incrementality* for repeated computations on different versions of the data, *caching* to avoid recomputation of previously computed values, and *parallelism* to enable distribution and acceleration of tasks.

Description of Work:
Task 2.1: Automatic validation and repair of extreme data (CEA, EHU, UDA, MUN, IPM) CLOUDMINER will improve upon existing automatic data quality solutions by providing a novel data validation and repair language to support the specification of complex validation patterns and repair behaviours. Automated data validation constraint inference will be supported. Evolutionary computation will be used to search for possible repairs and recommend them to the user. Incremental evaluation of constraints and their associated metrics will be supported to account for high velocity and volume data. Distribution of data validation tasks will be provided to the user in a transparent manner.
Task 2.2: Data preprocessing task interoperability and user guidance (CEA, EHU, MUN) This task will provide intelligent guidance to users on which preprocessing tools and algorithms to adopt. To this end, the concept of a care label for data preprocessing components will be developed. Care labels would then be used to compare and contrast different alternatives for specific data preprocessing tasks. Moreover, this task will also support the easy integration of pipeline components based on their care labels. Care labels will capture in a machine-readable format sets of pre- and post-conditions and invariants of preprocessing components. In addition, they will also store other relevant requirements and constraints for every component, such as pre-requisites for executing a particular algorithm. This information will then be used to integrate the different preprocessing components following a design-by-contract approach.
Task 2.3: Automated tagging of exteme data (CEA, EHU, UDA, MUN) The aim of this task will be the enhancement of datasets with semantically-rich metadata. A toolset will be developed to facilitate metadata specification and the definition of data patterns corresponding to specific metadata types. Moreover, this task will devise techniques for metadata inference based on previously-seen datasets and semantic analysis of datasets. In addition, automated tagging of datasets will be supported. Finally, this task will devise query and sophisticated metadata management of datasets for structured, semi-structured (e.g., JSON, XML), and unstructured data.
Task 2.4: Semantic data preprocessing (CEA, EHU, CLMS, UDA, MUN) This task will focus on extracting data from heterogeneous sources and tagging them with semantically-rich metadata. A domain-specific language and an architecture will be specified to abstract away from the different types of data sources and the related data extractors. Incremental and parallel data extraction and support for data streams will be provided to accommodate extreme data. Users of the CLOUDMINER technological solution will be able to integrate data from different sources using the provided metadata. The tools will also provide recommendations on data integrations based on the metadata and previously-seen integrations.

Partners Roles:
CEA will lead this work package, and will perform the majority of the technical work related to its tasks. CEA will also coordinate the contributions of other partners and consolidate their inputs to the deliverables of the work package.
UDA will contribute to developing recommender systems that will be devised in this work package and that concern Task 2.1, 2.3, and 2.4.
ATB will ensure the smooth integration of the data management tools with the pipeline execution environment developed in WP6.
EHU and MUN will contribute to the definition of data preprocessing tasks, particularly automatic tagging, semantic annotations, and care label definitions.
IPM will contribute to automatic data validation and repair with a focus on EDP data.

Deliverables:
D2.1: Automatic validation and repair of extreme data (CEA) (M12) This deliverable will present the automated and scalable data validation and repair components as an extension of the CLOUDMINER Studio (see WP3) developed in the context of Task 2.1.
D2.2: Data preprocessing task interoperability and user guidance (EHU) (M18) This deliverable will present the concept of a care label, developed in Task 2.2, and how it can be used for recommendations of data preprocessing components and for integrating such components in pipelines.
D2.3: Automated tagging of exteme data (CEA) (M31) This deliverable will present how data can be automatically enhanced with semantic information in an automated manner, based on the work in Task 2.3.
D2.4: Semantic data preprocessing (EHU) (M30) This deliverable will present how data can be extracted from heterogeneous sources and how they can be integrated using semantic information, based on the work in Task 2.4.
D2.5: Data preparation and quality (CEA) (M30) This deliverable with present the final version of the software tools developed in Tasks 2.1-2.4, as well as an empirical evaluation of these tools.

3.1.4.3 WP3: Polyglot Data Pipeline Development

Work package	3	**Start month**	1	**End month**	30
Work package title	Polyglot Data Pipeline Development				
Participant name (number)	UDA (2)	**YORK (3)**	MUN (5)	EHU (6)	
Person-months	5	**36**	5	15	
Participant name (number)	ATB (7)	CLMS (8)			
Person-months	7	3			

Objectives: This work package aims to facilitate the graphical development, monitoring, and debugging of data mining and analytics pipelines in an intelligent and modular web-based environment. The graphical development environment will support smart editing of embedded data transformation scripts (e.g., in Python, R, Julia) by providing support for existing LSP language servers, and will facilitate live collaborative development, differencing, conflict detection and merging of different versions of pipelines.

Description of Work:
Task 3.1: Graphical Data Pipeline Development (YORK, UDA, ATB) This task will develop a modular web-based graphical (diagram-based) tool (CLOUDMINER Studio) for assembling and configuring data mining and analytics pipelines from existing container-based data processing and transformation components. CLOUDMINER Studio will feature facilities through which users will be able to discover and reuse components, and intelligent connectors and auto-layout features for assembling them into complex pipelines. The configuration user interface of data transformation, learning and visualisation components will be contributed by microservices in their respective containers to facilitate extensibility and loose coupling. CLOUDMINER Studio will also feature real-time collaborative development of pipelines by leveraging the Graphical Language Server Protocol.
Task 3.2: Intelligent Data Processing Script Development (YORK, ATB, MUN) As experience with similar platforms (e.g., Apache NiFi) has demonstrated, glue scripts in languages such as Python, R and Julia are often necessary for complex data mining and analytics pipelines. Unfortunately, existing platforms offer very basic tooling for writing such scripts. In the best case, an embedded editor with syntax-highlighting capabilities is offered, without support for context-aware code completion of error reporting. This task will use existing language servers for popular languages such as Python, R, and Julia to provide intelligent script editing capabilities in CLOUDMINER Studio to improve the developer experience and detect and report errors before a pipeline gets executed.
Task 3.3: Natural Language Interaction as a Pipeline Design Aid (EHU, YORK, MUN) This task will focus on developing an add-on facility for CLOUDMINER Studio to enable user interaction through natural language. Instructions in natural language will be analysed to extract terms and relationships; they will be matched against the available data transformation, learning, and visualisation components of the platform, also taking into account the implicit or explicit schema of the data under analysis, and they will be transformed into pipeline fragments that users will be able to refine and fine-tune.
Task 3.4: Graphical Pipeline Monitoring and Debugging (YORK, ATB, CLMS) This task will extend CLOUDMINER Studio with pipeline monitoring, runtime configuration, and debugging capabilities. While CLOUDMINER Studio will not be necessary for executing pipelines, it will be able to hook into the execution of local or remote pipelines and monitor the execution of individual components (e.g., execution time, incoming/outgoing data, backlog), scale up/down individual components at runtime by instructing the container orchestrator of Work Package 6 to adjust the number of container instances respectively, and place breakpoints and input/output watches in individual components to enable debugging.
Task 3.5: Pipeline Differencing, Conflict Detection and Merging (YORK, ATB, CLMS) Pipelines need to be versioned like any other piece of software. However, as pipeline specifications will be stored in a structured format (e.g., XMI, XML, JSON), performing differencing, conflict detection, and merging on them using standard text-based facilities will be tedious. Therefore, in this task, we will implement custom facilities for identifying changes between different pipeline versions and detecting and resolving conflicts in a semi-automated way (some conflicts will require user intervention to resolve).

Partners Roles:
YORK will lead this work package, and will perform the majority of the technical work related to its tasks. YORK will also coordinate the contributions of other partners and consolidate their inputs to the deliverables of the work package.
EHU will lead Task 3.3 and the corresponding deliverable (D3.3). EHU will be responsible for research and development on natural language processing methods needed to map conversation snippets in free language to data processing/mining concepts and relations among them.
ATB will participate in the development of data pipeline development tools to ensure consistency with the data pipeline deployment tools and data pipeline execution environment.
UDA will ensure consistency of the pipeline development tools with the intelligent recommenders conceived in WP5
MUN will ensure the adoption of the data mining techniques developed in WP4 from the pipeline development tools.
CLMS will ensure the integration of the pipeline development tools in the CLOUDMINER integrated platform.

Deliverables:
D3.1: Graphical Data Pipeline Development Tools (YORK) (M12) This deliverable will present the CLOUDMINER Studio web-based development environment for assembling and configuring data mining and analytics pipelines. CLOUDMINER Studio itself will be made publicly available as open-source software in accordance with the open-source strategy of CLOUDMINER.
D3.2: Intelligent Data Processing Script Development Facilities (YORK) (M18) This deliverable will present the extension of CLOUDMINER Studio with intelligent textual editors for commonly-used scripting languages such as Python, R, and Julia based on existing LSP servers for these languages. A new release of CLOUDMINER Studio with the extended script editing capabilities will also be made in time for the first evaluation cycle of the project in Work Package 7.
D3.3: Natural Language Interaction as a Pipeline Design Aid (EHU) (M18) This deliverable will present the design of the natural language component for CLOUDMINER Studio. It will propose and evaluate methods for recognising technical terms and relations among them in user instructions, matching them against the available CLOUDMINER Studio components, and transforming them into pipeline fragments.
D3.4: Graphical Pipeline Monitoring and Debugging Extensions (YORK) (M24) This deliverable will present the pipeline monitoring, runtime configuration, and debugging extensions of CLOUDMINER Studio developed in Task 3.4. In addition, a new version of CLOUDMINER Studio with these capabilities integrated will also be released.
D3.5: Pipeline Differencing, Conflict Detection and Merging Tools (YORK) (M30) This deliverable will present the differencing, conflict detection, and merging facilities of CLOUDMINER Studio that will be developed in Task 3.5. In addition, a feature-complete version of CLOUDMINER Studio will be released with this deliverable to feed into the second evaluation round of the project in Work Package 7.

3.1.4.4 WP4: Advanced Analytics Techniques for Deep Data Mining

Work package	4	**Start month**	1	**End month**	30
Work package title	Advanced Analytics Techniques for Deep Data Mining				
Participant name (number)	UDA (2)	**MUN (5)**	EHU (6)	CLMS (8)	
Person-months	3	**27**	15.8	3	

Objectives: The aim of this work package is to develop advanced analytics techniques and tools for deep data mining. The techniques developed in this work package aim to discover and distil meaningful, reliable and useful patterns from large, heterogeneous, and dispersed data sources. Users will also be provided with a chatbot to explore date before and after processing using Natural Language (NL) conversations. The techniques will be implemented as extensions to CLOUDMINER Studio.

Description of Work:
Task 4.1: Smart data visualisation (MUN, CLMS) This task will create an extension to CLOUDMINER Studio for interactive visualisations over large and complex datasets. We will devise methods to learn effective multi-dimensional data visualisations arising from combining the semantics of the underlying data and user activity visualisation strategies. The extension will implement FAIR for data visualisations by enabling author-enhancement of automatically generated metadata and enabling direct embedding into research publications.
Task 4.2: Privacy-preserving federated data mining (EHU, CLMS) This task will explore methods for distributed machine learning and deep learning models for predictive data analytics in multi-node settings to improve performance, increase accuracy and allow models to scale to larger data sizes, into the big data and extreme data level. Due to the statistics-based nature of these algorithms, increasing the size of the training data can significantly reduce the learning error. Distributed methods will consider multiple CPUs and processing nodes in a computer cluster, distributed data sources (e.g., edge devices, personal data vaults, and distributed datasets), security, and privacy preservation. Existing and novel federated methods will be available to CLOUDMINER users as alternative components.
Task 4.3: Semantic data mining (MUN, UDA, CLMS) This task will develop new techniques to efficiently uncover patterns in large, multi-dimensional, and heterogeneous datasets. We will explore graph-based methods such as community detection to help reduce the search space and use the semantic mapping technology developed in WP2 to help guide the search path.

Task 4.4: Chatbot for semantic question answering (EHU, CLMS) This task will develop a chatbot to answer questions against the data under investigation, including the results of analyses performed. The chatbot will help users understand and construct well-formulated questions, based on a controlled natural language, with suggested types, relations, and qualifiers from indexed data and computed results. It will make use of NL interaction components developed in Task 3.3.

Partners Roles:
MUN will lead this work package, and will perform the majority of the technical work related to its tasks. MUN will also coordinate the contributions of other partners and consolidate their inputs to the deliverables of the work package.
EHU will lead Tasks 4.2 and 4.4 and the write-up of the corresponding deliverables, i.e., D4.2 and D4.4. EHU will be responsible for research and development on methods for federated Machine Learning and Deep Learning and Semantic Question Answering application of models.
CLMS will ensure the integration of the advanced analytics tools in the CLOUDMINER integrated platform.
UDA will contribute the definition of the graph-based methods in Task 4.3.

Deliverables:
D4.1: Smart Data Visualisation Tool (MUN) (M12) This deliverable will present an intelligent data visualisation tool as an open-source extension of CLOUDMINER Studio.
D4.2: Privacy-preserving Federated Data Mining Tool (EHU) (M18) This deliverable will present research and evaluation of distributed training of Machine Learning and Deep Learning models. Evaluation concerning accuracy and performance will consider datasets provided by CLOUDMINER partners and freely available datasets and benchmarks. The deliverable will also present CLOUDMINER platform components that will *i)* create and deploy secure, privacy-preserving data nodes, and *ii)* enable distributed data mining across a network of selected nodes.
D4.3: Semantic Data Mining Tool (MUN) (M18) This deliverable will present an open-source semantic data mining component for CLOUDMINER Studio.
D4.4: Semantic Query Answering Tool (EHU) (M30) This deliverable will present details about the design and development process of the chatbot of CLOUDMINER Studio for semantic query mining. It will also discuss several use cases that focus on the datasets provided by CLOUDMINER partners.

3.1.4.5 WP5: Intelligent Recommenders for Data Mining Assistance

Work package	5	**Start month**	1	**End month**	30
Work package title	Intelligent Recommenders for Data Mining Assistance				
Participant name (number)	**UDA (2)**	YORK (3)	ATB (7)	CLMS (8)	
Person-months	**18**	7	3	3.5	

Objectives: This WP will develop intelligent recommender systems to support data science processes, employing relevant recommendations for the specification and execution of the data pipelines under development. Machine learning algorithms will be employed to enable the automated identification of relevant data mining components and pipeline patterns that developers can reuse to accomplish the tasks at hand. Like any machine learning approach, data availability is of paramount importance. This WP will define the knowledge base structure that will represent in a homogenous manner data mining components, NLP tasks, data sources, complete pipelines, and their relations and dependencies. The knowledge base will enable innovative features, e.g., automated identification of pipeline fragments that developers can consult for insights and ideas about how to develop the wanted data mining process further, identification of similar pipelines, best practices that serve the required data, and provide developers with automatically classified pipelines, which can be considered for reuse according to quality criteria. The recommendation mechanisms that will be developed in this work package will support both the definition and the optimized execution of data pipelines.

Description of Work:
Task 5.1: CLOUDMINER Knowledge Base Development (UDA, YORK, CLMS) This task will design the CLOUDMINER knowledge base that will underpin the development of automatic and real-time recommendations. The knowledge base will be designed to encode data pipelines and constituting elements in graph-based representations so that ML techniques like collaborative filtering and content-based machine learning models can be employed. Furthermore, the knowledge base will be developed as microservices to facilitate adoption by CLOUDMINER components like the graphical data pipeline development tool defined in Task 3.1, the pipeline monitoring and debugging capabilities developed in Task 3.4, and the optimization facilities for pipeline execution conceived in WP6.
Task 5.2: Development of Intelligent Recommenders for Data Pipelines Specification (UDA, YORK) Existing platforms offer limited support for assisting users while defining data pipelines. Also, the reuse possibilities are limited to the concept of reusable templates and inventories of components that are not organized for the current development context. This

task will use machine learning techniques to provide real-time recommendations throughout the whole development process, from selecting and ranking reusable entities that best fit the current problem to their composition in working data pipelines.
Task 5.3: Development of Intelligent Recommenders for Optimizing Data Pipelines Execution (UDA, ATB) In this task, we will collaborate closely with WP6 to provide users with actionable guidelines to optimize pipelines under development. For example, replacing sequential with parallel flows, applying data quality checks, or executing data filtering operations are possible suggestions that the CLOUDMINER recommender systems will provide to improve data pipelines to increase their execution performance. Furthermore, in this task, we will develop facilities to analyze the data produced by the execution components conceived in WP6. This way, CLOUDMINER will give recommendations that consider past executions of pipelines similar to the one under analysis (e.g., those that share structural characteristics or mine data from the same sources).

Partners Roles:
UDA will lead this work package, and will perform the majority of the technical work related to its tasks. UDA will also coordinate the contributions of other partners and consolidate their inputs to the deliverables of the work package.
ATB will participate in developing intelligent recommenders for optimizing data pipelines execution to ensure consistency with the data pipeline deployment tools and data pipeline execution environment.
YORK will lead the integration of the intelligent recommenders developed in this WP with the pipeline development environment of WP3.
CLMS will ensure the integration of the intelligent recommenders in the CLOUDMINER integrated platform.

Deliverables:
D5.1: The CLOUDMINER Knowledge Base for Intelligent Recommenders (UDA) (M12) This deliverable will present the CLOUDMINER knowledge base designed in Task 5.1.
D5.2: The CLOUDMINER Intelligent Recommender Systems for Data Pipelines Specification (UDA) (M18) This deliverable will present the intelligent recommender systems for data pipelines specification as developed in Task 5.2.
D5.3: The CLOUDMINER Intelligent Recommender Systems for Optimizing Data Pipelines Execution (UDA) (M30) This deliverable will present the intelligent recommender systems for data pipelines execution as developed in Task 5.3.

3.1.4.6 WP6: Optimised Data Pipeline Execution

Work package	6	**Start month**	1	**End month**	30
Work package title	Optimised Data Pipeline Execution				
Participant name (number)	UDA (2)	YORK (3)	**ATB (7)**	CLMS (8)	
Person-months	6	14	**19**	3.5	

Objectives: This WP aims to provide an execution environment for data pipelines and a graphical tool to create deployments for developed data pipelines. The graphical deployment tools will support data engineers in creating container images that are directly deployable to container or container orchestration solutions (e.g., Docker, Kubernetes). The data pipeline execution environment will be based on a state-of-the-art execution environment (e.g., Apache Airflow) and support DataOps based on distributed micro-services (data pipelines). Within this WP, a web-based Deployment Dashboard that allows to visualise the current status of deployed data pipelines and monitor deployed data pipelines will be developed.

Description of Work:
Task 6.1: Definition of Pipeline Deployment Models (ATB, UDA, YORK) This task will define the syntax and semantics of the data pipeline deployment language. The deployment language will allow to build container images for each data pipeline automatically. Data Pipeline Deployment models will be produced in an automated (but customisable) manner based on the data pipeline specifications given with the environment developed in WP3.
Task 6.2: Development of Pipeline Deployment Tools (ATB, UDA, YORK) This task will design and develop the tool that engineers can use to refine data pipeline deployment specifications automatically obtained from data pipeline specifications defined and developed with the tools of WP3. Each pipeline component will be built as its own container image. The recommendations provided by the recommenders developed in WP5 will also be exploited in this task. Produced container images will be deployable on container and container orchestration solutions, such as Docker or Kubernetes.
Task 6.3: Development of Pipeline Execution Environment (ATB, UDA, YORK, CLMS) In this task, the execution environment will be designed and developed. The execution environment will be based on state-of-the-art execution environments, such as Apache Airflow, and extended to support pipeline executions and DataOps based on distributed micro-services (data pipelines). The CLOUDMINER execution environment will also include DevOps concepts, such as continuous integration and delivery, to automate the testing and deployment of data pipelines. A web-based Deployment Dashboard that visualizes the current status of deployed data pipelines and monitors deployed data pipelines will be

developed within CLOUDMINER and be part of the execution environment. Furthermore, the Pipeline Execution Environment will provide APIs to enable remote monitoring and debugging capabilities from external tools.

Partners Roles:
ATB will lead this work package, and will perform the majority of the technical work related to its tasks. ATB will also coordinate the contributions of other partners and consolidate their inputs to the deliverables of the work package.
YORK will ensure the seamless interoperation of the pipeline execution environment with the graphical pipeline development environment of of WP3 for pipeline debugging and monitoring.
UDA will ensure the adoption of the intelligent recommenders to optimize the execution of specified data pipelines.
CLMS will ensure the integration of the pipeline execution environment in the CLOUDMINER integrated platform.

Deliverables:
D6.1: Data Pipeline Deployment Language (ATB) (M12) This deliverable will present the Data Pipeline Deployment language.
D6.2: Data Pipeline Deployment Tools (ATB) (M18) This deliverable will present the design and implementation of the data pieline deployment tools that will be used to define the data pipeline deployments based on data pipeline specifications.
D6.3: Data Pipeline Execution Environment - Initial Version (ATB) (M18) This deliverable will present the design and initial implementation of the tools for transforming recommendations for data pipeline specifications into corresponding data pipeline deployment models.
D6.4: Data Pipeline and Recommender Transformation Tools (ATB) (M24) This deliverable will present the final implementation of the data pipeline deployment tools, that will be used to define the data pipeline deployments based on data pipeline specifications.
D6.5: Data Pipeline Execution Environment - Final Version (ATB) (M30) This deliverable will present the design and final implementation of the transformation tools for transforming recommendations for data pipeline specifications into the data pipeline deployment models.

3.1.4.7 WP7: Integration, Deployment and Industrial Evaluations

Work package	7	**Start month**	1	**End month**	36
Work package title	Integration, Deployment and Industrial Evaluations				
Participant name (number)	TOG (1)	UDA (2)	YORK (3)	CEA (4)	
Person-months	4.5	5	5	5	
Participant name (number)	MUN (5)	EHU (6)	ATB (7)	**CLMS (8)**	
Person-months	5	5	5	**29.5**	
Participant name (number)	INFT (9)	MLGX (10)	EDP (11)	IPM (12)	
Person-months	30	29	12	10	
Participant name (number)	CMM (13)	CONTI (14)			
Person-months	6	21			

Objectives: The aim of this work package is to guide and synthetize the findings of work packages 2-6 into an integrated workbench for implementing and executing data analytics pipelines that will be readily available to users. To this end, the work package has the following objectives:

- Provide the architectural guidelines that will ensure that the contributions of work packages 2-6 will be interoperable by construction;
- Provide an integrated platform that will consolidate the tools contributed by work packages 2-6. A **release of the integrated platform** will be made **every 6 months** to allow the integration of the different parts all along the project from M12 to M30;
- Evaluate and validate the CLOUDMINER tools within the context of relevant industry driven Use Cases;
- Quantify the industrial user partner experiences in terms of development effort, cost savings, developer productivity, etc., to document for others the benefits achieved from the CLOUDMINER technologies.

Evaluation will be carried out **in two iterations**. In the first iteration (M19-21), use case partners will carry out a **preliminary evaluation** on an interim version of the technical contributions of the project on redacted versions of their use cases and will compile reports that will present their findings and provide recommendations that will guide the remaining technical work. In the second iteration (M30-36), use case partners will carry out a **full-scale evaluation** of the facilities of the platform on their complete use cases. Feedback obtained during the second iteration will be incorporated in the final version of the platform which is due on M36.

Description of Work:
Task 7.1: Architectural Guidelines Establishment (CLMS, ATB, YORK, UDA, MUN, EHU, CEA) In order to minimise the risk that technical work packages 2-6 will produce results that are challenging to integrate later on in the context of a common platform, this task will capture core design decisions and establish architectural guidelines that work packages 2-6 will then be required to adhere to.
Task 7.2: Platform Integration and Evaluation Support (CLMS, ATB, YORK, UDA, MUN, EHU, CEA) This task will integrate the technical contributions of work packages 2-6 into a unified development studio and deployment platform, and will provide support to the demonstrator use-cases. We envision that the CLOUDMINER development tools will be implemented on top of a cloud-based IDE.
Task 7.3: Evaluation Methodology (TOG, UDA, YORK, CEA, MUN, EHU, ATB, CLMS, EDP, IPM, CMM, CONTI, INFT, MLGX) This task will specify the detailed measurement methods and metrics, along with the evaluation procedures (e.g. side-by side-comparisons, performance measurements, etc.) that will be used within the four Use Cases to evaluate the extent to which the project has achieved the success criteria defined early in the project (see work package 1). This task focuses on methods of evaluating the intrinsic performance of the CLOUDMINER tools based on KPIs (e.g. response time, use of computing resources, data complexity, data volumes, etc). Testbeds, probes and test cases will be specified and the resulting deliverable will ensure the generation of quantitative, consistent, comparable and industry relevant evaluation results from each Use Case.
Task 7.4: Deep Data Mining for Traffic Statistics and Modal Share Monitoring Use Case (INFT) This task will evaluate the CLOUDMINER technologies in the context of the traffic information and road maintenance management use case discussed in Section 1.2.6.1, and assess the level of achievement with respect to the applicable target measures established for the project.
Task 7.5: Deep Data Mining for Improved Meteorological Data and Forecasting Services Use Case (MLGX) This task will evaluate the CLOUDMINER technologies in the context of the weather forecasting use case discussed in Section 1.2.6.2, and assess the level of achievement with respect to the applicable target measures established for the project.
Task 7.6: Energy Data for Maia City Use Case (EDP, IPM, CMM) This task will evaluate the CLOUDMINER technologies in the context of the smart cities use case discussed in Section 1.2.6.3, and assess the level of achievement with respect to the applicable target measures established for the project.
Task 7.7: Process Data Mining and Analytics to Achieve Early Fail Detection, Reduce Customer Returns and Reduce Process Cycle Time Use Case (CONTI) This task will evaluate the CLOUDMINER technologies in the context of the automotive use case discussed in Section 1.2.6.4, and assess the level of achievement with respect to the applicable target measures established for the project.

Partners Roles:
CEA, YORK, MUN, UDA, ATB will contribute to Tasks 7.1 - 7.2 to ensure integration with tools developed in the work packages they lead, i.e., WP2, WP3, WP4, WP5, and WP6, respectively.
EHU will contribute to Tasks 7.1 - 7.2 to ensure integration with tools developed in WP2-WP6.
TOG will contribute to Tasks 7.1 - 7.3 to ensure the satisfaction of the requirements elicited in WP1 and to support the evaluation procedures.
INFT will demonstrate the CLOUDMINER tool chain for the traffic information and road maintenance management use case.
MLGX will demonstrate the CLOUDMINER tool chain for the weather forecasting use case.
EDP, IPM and CMM will demonstrate the CLOUDMINER tool chain for the smart cities use case.
CONTI will demonstrate the CLOUDMINER tool chain for the automotive use case.

Deliverables:
D7.1: Architectural Guidelines Report (CLMS) (M6) This deliverable will capture and present the core design decisions and architectural guidelines that will ensure that the contributions of work packages 2-6 will be interoperable by construction. The report will outline the architecture of the platform and the extension points that tool developers can leverage to integrate additional technologies with the platform.
D7.2: Integrated Platform - Initial Version (CLMS) (M18) This deliverable will comprise a software prototype and a report. synthesise relevant technical contributions from work packages 2-6 into an integrated platform that will enable data pipeline specification, deployment, and execution based on a first distributed architecture. The platform will be appropriately tested and packaged and will be ready for the industrial users to install and use in order to perform their case studies.
D7.3: Evaluation Methodology (TOG) (M18) This deliverable builds upon the initial Evaluation Plan from work package 1, to provide a detailed specification of the evaluation methodologies and measurements that will be carried out for each of the four Use Cases. It will include the scope of validation activities, the list of metrics (KPIs) to be evaluated comprising also productivity/efficiency/economical benefits and the testbeds, test cases and methods to gather and evaluate these measures.
D7.4: Deep Data Mining for Traffic Statistics and Modal Share Monitoring Use Case Evaluation - Interim Version (INFT) (M21) This deliverable will report on the findings of the interim evaluation of the components of the CLOUDMINER platform that will be ready to use by M18 and will provide recommendations for the direction of the technical work in the remaining of the project from the point of view of the traffic information and road maintenance management use case.
D7.5: Deep Data Mining for Improved Meteorological Data and Forecasting Services Use Case Evaluation - Interim Version (MLGX) (M21) This deliverable will report on the findings of the interim evaluation of the components of the CLOUDMINER platform that will be ready to use by M18 and will provide recommendations for the direction of the technical work in the remaining of the project from the point of view of the weather forecasting use case.

D7.6: Energy Data for Maia City Use Case Evaluation - Interim Version (EDP,IPM,CMM) (M21) This deliverable will report on the findings of the interim evaluation of the components of the CLOUDMINER platform that will be ready to use by M18 and will provide recommendations for the direction of the technical work in the remaining of the project from the point of view of the smart cities use case.
D7.7: Process Data Mining and Analytics to Achieve Early Fail Detection, Reduce Customer Returns and Reduce Process Cycle Time Use Case Evaluation - Interim Version (CONTI) (M21) This deliverable will report on the findings of the interim evaluation of the components of the CLOUDMINER platform that will be ready to use by M18 and will provide recommendations for the direction of the technical work in the remaining of the project from the point of view of the automotive use case.
D7.8: Integrated Platform - Final Version (CLMS) (M30) This deliverable will provide a software prototype and a report. The prototype will provide the final version of the integrated platform that will now include the final versions of all the tools developed in CLOUDMINER . The prototype will synthesise relevant technical contributions from work packages 2-6 into an integrated platform that will enable the development and deployment of cloud-native, polyglot and scalable data mining and analytics pipelines. The platform will be appropriately tested and packaged and will be ready for the industrial users to install and use in order to perform their case studies.
D7.9: Deep Data Mining for Traffic Statistics and Modal Share Monitoring Use Case Evaluation - Final Version (INFT) (M36) An evaluation report based on applying the CLOUDMINER tools for the development of the traffic information and road maintenance management use case. This report will present the development tasks undertaken, the tools that were utilised for construction and implementation of a domain-specific demonstrator, the experiences gained, and the improvements achieved from applying CLOUDMINER technologies. An assessment of the level of achievement with respect to the applicable target measures established for the project will be provided.
D7.10: Deep Data Mining for Improved Meteorological Data and Forecasting Services Use Case Evaluation - Final Version (MLGX) (M36) An evaluation report based on applying the CLOUDMINER tools for the development of the weather forecasting use case. This report will present the development tasks undertaken, the tools that were utilised for construction and implementation of a domain-specific demonstrator, the experiences gained, and the improvements achieved from applying CLOUDMINER technologies. An assessment of the level of achievement with respect to the applicable target measures established for the project will be provided.
D7.11: Energy Data for Maia City Use Case Evaluation - Final Version (EDP,IPM,CMM) (M36) An evaluation report based on applying the CLOUDMINER tools for the development of the smart cities use case. This report will present the development tasks undertaken, the tools that were utilised for construction and implementation of a domain-specific demonstrator, the experiences gained, and the improvements achieved from applying CLOUDMINER technologies. An assessment of the level of achievement with respect to the applicable target measures established for the project will be provided.
D7.12: Process Data Mining and Analytics to Achieve Early Fail Detection, Reduce Customer Returns and Reduce Process Cycle Time Use Case Evaluation - Final Version (CONTI) (M36) An evaluation report based on applying the CLOUDMINER tools for the development of the automotive use case. This report will present the development tasks undertaken, the tools that were utilised for construction and implementation of a domain-specific demonstrator, the experiences gained, and the improvements achieved from applying CLOUDMINER technologies. An assessment of the level of achievement with respect to the applicable target measures established for the project will be provided.

3.1.4.8 WP8: Ecosystem, Dissemination, Exploitation and Standardisation

Work package	8	**Start month**	1	**End month**	36
Work package title	Ecosystem, Dissemination, Exploitation and Standardisation				
Participant name (number)	**TOG (1)**	UDA (2)	YORK (3)		CEA (4)
Person-months	**8**	4	4		4
Participant name (number)	MUN (5)	EHU (6)	ATB (7)		CLMS (8)
Person-months	4	7.5	4		4
Participant name (number)	INFT (9)	MLGX (10)	EDP (11)		IPM (12)
Person-months	4	3.5	2		2
Participant name (number)	CMM (13)	CONTI (14)			
Person-months	1	3			

Objectives: The aim of this work package is to execute supporting activities including dissemination, exploitation, standardization and documentation that ensures the broad take-up of CLOUDMINER results by industry, and to encourage further research and on-going development of CLOUDMINER technologies. The key objectives of this work package are to

- act as an observatory to identify market trends, opportunities, industrial needs, etc.;
- lead a European-wide dissemination activity to maximize the impact of the results from the CLOUDMINER project;
- lead an exploitation strategy to facilitate the successful exploitation and take-up of CLOUDMINER results; and
- lead a strategy for standardization of CLOUDMINER project results.

Fulfillment of these objectives will result in a wide range of European data scientists and technology providers benefiting from the new data mining and analytics tools and technologies that will be developed in the project.

Description of Work:
Task 8.1: Dissemination and Promotion (TOG, UDA, YORK, CEA, MUN, EHU, ATB, CLMS, EDP, IPM, CMM, CONTI, INFT, MLGX) This task will carry out dissemination activities in order to ensure a substantial impact both at the EU level and international level. The actions planned within this task include the following.

- **Dissemination material**—Logo, brochures, templates, and other actions that create identity, consistency and awareness of the CLOUDMINER project.
- **Events Participation**—CLOUDMINER will be presented in multiple related European Events, such as conferences, fairs, and congresses,—a minimum of at least six events will be targeted—plus at least 2 International Events (outside of Europe).
- **Publications**—articles will be produced for industrial conferences, technical journal papers, and publicity (e.g. White Paper) by both research and industrial partners within the project.

All partners will contribute to this task.
Task 8.2: Exploitation and IPR (TOG, UDA, YORK, CEA, MUN, EHU, ATB, CLMS, EDP, IPM, CMM, CONTI, INFT, MLGX) This task will define a plan to support the members of CLOUDMINER consortium to make an effective exploitation of project results. The specific exploitation actions of the project will be coordinated through the establishment of an Exploitation Board consisting of the research partners, the ICT partners, and industrial user partners. The exploitation actions include the following:

- **Business Network**—In addition to the IT supplier partners in the project consortium, a pan-European network of solution providers/ suppliers involved with extreme data mining will be established to license the CLOUDMINER results for commercial solutions and services.
- **Licensing and Management**—The project partners will make available under open source license the foreground technologies developed in the project. Online facilities will be set-up to allow easy access to the project results, and specific procedures and facilities will be provided to encourage contributions to the open source base, and to manage the continued evolution and versioning of the open source results from the project.

Specific actions for each exploitable result will be identified in the Plan for Using and Disseminating Knowledge deliverable prepared early in the project and periodically updated. All partners will contribute to this task.
Task 8.3: Communications (TOG, UDA, YORK, CEA, MUN, EHU, ATB, CLMS, EDP, IPM, CMM, CONTI, INFT, MLGX) The project will undertake a Public Communications Programme to ensure broad awareness of the project across the wider IT supplier and industrial extreme data mining community. The major elements of the public communications plan include initial international press releases announcing the launch of the CLOUDMINER project and progress at key milestones released simultaneously to all major technical press and journals in Europe, USA and Asia, regular updates to the public project website to alert those tracking the project when new information concerning the project, upcoming events, etc. are available. Further communications will be undertaken in association with specific conferences and events where CLOUDMINER will participate, and in collaboration with specific journals where CLOUDMINER papers and articles will be published. All partners will contribute to this task.
Task 8.4: Market Watch (TOG, UDA, YORK, CEA, MUN, EHU, ATB, CLMS, EDP, IPM, CMM, CONTI, INFT, MLGX) The Market Watch observatory will monitor external developments and opportunities relevant to the CLOUDMINER project. The key activities of the observatory are to ensure that CLOUDMINER is positioned as well as possible both technically and commercially by:

- regularly monitoring advances in state-of-the-art, new technology breakthroughs useful for the project, the development of other international projects, dissemination and exploitation opportunities;
- conducting socio economic analysis of the project results and to deliver cost benefit analysis (CBA) that will assist in the exploitation strategy;
- tracking changes to standards in domains germane to open source tools and platforms.

The Observatory will also be responsible for coordinating actions for sharing and gathering information with selected industrial and research networks.
Task 8.5: Contribution to Standards (TOG, UDA, YORK, CEA, MUN, EHU, ATB, CLMS, EDP, IPM, CMM, CONTI, INFT, MLGX) For each of the project results that are strategic to the CLOUDMINER toolset and are intended as standards, a Standardization Leader selected from one of the consortium partners will be designated. The Open Group will work with each designated Standards Leader to identify the appropriate standards bodies or industry grouping (e.g. OMG, Eclipse/Apache Foundations, etc.), to produce the specifications or technologies in formats appropriate for standards submission, and to establish a plan of action for achieving consensus, including collaboration with members of the grouping, events or initiatives.

Partners Roles:
TOG will lead the dissemination, exploitation, and standardisation activities with contributions from all partners to ensure the broadest awareness and take-up of the project technologies. TOG will host the project website, manage the project's social media accounts, ensure dissemination materials are professionally prepared and will lead the preparation of the exploitation plans and strategies involving all project partners.
IPM, CMM, INFT, MLGX, EDP, and CONTI will undertake industry-focused dissemination actions in their respective domains as well as planning for the exploitation of the project results within their own organisations. They will also provide motivational presentations, testimonials, evaluation results and other materials at appropriate conferences and encourage external organisations to exploit the project results to benefit the wider data science community. They will also support

the standardisation activities reinforcing industry support for the newly proposed standards arising from the project. YORK, UDA, CEA, MUN, CLMS, ATB, and EHU will carry out specific dissemination actions of presentations at conferences, preparation of journal articles, and planning for further development and exploitation of the project results. They will also participate in the standardisation actions by assisting in the preparation of proposed standards specifications and working to achieve consensus within the targeted standards bodies.

Deliverables:
D8.1: Project Website (TOG) (M3) A public website describing the project objectives, approach, partners involved and expected results in the initial version. Later updates will include announcements of workshops and conferences, downloadable papers and articles and the public deliverables from the project. The website will be updated periodically.
D8.2: Project Presentation and Brochure (TOG) (M6) Materials to present the project to interested parties including details on the technical challenges and the approaches being developed within the project to address the challenges, the expected impact from both a technical and societal standpoint, and where to obtain further information and details concerning the project research and development work.
D8.3: Initial Dissemination and Use Plan (TOG) (M14) Report that provides a first planning of the dissemination and exploitation activities the partners plan to undertake including dissemination goals and channels, target groups, and exploitation approach for each project result. It also contains a draft of the long-term business and research goals of the consortium partners.
D8.4: Press and Media Materials (TOG) (M24) Summary of Press Releases and other supporting materials and activities used during the operation of the project to promote and create awareness of the project and expected impact for industry. The materials also includes a summary of articles, interviews and other write-ups and articles that reference the CLOUDMINER project generated from the communications activities carried out within the project.
D8.5: Final Dissemination and Use Plan (TOG) (M36) Report that provides a further detailed planning of the dissemination and exploitation activities the partners plan to undertake including further refinements of the dissemination goals and channels, target groups, and exploitation approach for each project results, along with detailed individual plans for exploitation for each partner. It also contains market positioning and opportunity analysis along with key messages for promoting the exploitation of the project results within the targeted communities of tool vendors, software developers and researchers, and data scientists.
D8.6: Standardisation Report (TOG) (M36) Report describing the activities carried out within the project towards industry standardization of the project results including preparation of submissions to standards grouping, consensus building within the standardization communities, and addressing any technical challenges or inclusion of alternative approaches. The report will describe the progress the project has achieved through the standardization process within the appropriate standards bodies.

3.1.4.9 WP9: Project Management

Work package	9	**Start month**	1	**End month**	36
Work package title	Project Management				
Participant name (number)	**TOG (1)**	UDA (2)	YORK (3)	CEA (4)	
Person-months	**14**	5	1.4	1.3	
Participant name (number)	MUN (5)	EHU (6)	ATB (7)	CLMS (8)	
Person-months	1.3	1.3	1.3	1.5	
Participant name (number)	INFT (9)	MLGX (10)	EDP (11)	IPM (12)	
Person-months	1.3	1.3	1.3	1.3	
Participant name (number)	CONTI (14)				
Person-months	1.5				

Objectives: The work package ensures the overall management of the project including and includes the following objectives: (1) Overall project management and coordination of the project in accordance with the contractual obligations, especially ensuring the timely delivery of all project outputs and reports; (2) Effective management of project internal and external relations and communications, including effective communication with the EC; (3) Efficient management of time and resource allocation at a consortium level, facilities, representatives at meetings and general administrative duties; (4) Execute quality control and actively address risks with effective contingency measures. Further project management details and procedures are described in more detail in Section 3.2.

Description of Work:
Task 9.1: Consortium start up (TOG, UDA, YORK, CEA, MUN, EHU, ATB, CLMS, EDP, IPM, INFT, CONTI, MLGX) In accordance to the detailed project plan and consortium agreement, all the project start up activities will be established and supported in this task. Responsibilities, collaborations of partners, review and conflict resolution procedures, as well as reporting and detailed delivery description will be refined and assigned to project participants. Project communications facilities, infrastructure for managing project artefacts (documents, code, etc), will be established.
Task 9.2: Project progress tracking and coordination (TOG) Quarterly coordination, planning and quality assurance

meetings (Management Board) will be held. The responsible leaders of the consortia members and one to two stake-holders of member companies or organisation will meet to provide advice and steering inputs for next steps and resolve problems that come up.
Task 9.3: Quality Management (TOG) This task will include the procedures to ensure the quality of the project results meets the expectations of the partners, the Commission and the targeted industries as important research technologies capable of being exploited. Procedures to ensure consistency and quality content will be undertaken. Review tasks will be assigned to specific partners to provide comments and recommendations on deliverables, as well as use of external trusted experts to provide additional feedback.
Task 9.4: Project reporting to EU (TOG, UDA, YORK, CEA, MUN, EHU, ATB, CLMS, EDP, IPM, INFT, CONTI, MLGX) Semi-annual reporting of achievements and status for each phase. The project reports are used for reporting the progress of work carried out from the beginning of the project or the previous project report. This also makes deviations and major risks and counter measures explicit.
Task 9.5: Project reviews and contract revisions (TOG, UDA, YORK, CEA, MUN, EHU, ATB, CLMS, EDP, IPM, INFT, CONTI, MLGX) Regular project reviews for the project will be planned and carried out in collaboration with the Commission to assess progress, discuss strategic direction and gain further guidance from eternal experts. Activities to address any revisions to the workplan that may be required either as formal contract amendment or informal clarification will be undertaken.

Partners Roles:
TOG will lead the coordination and project management activities with contributions from all partners to ensure the project progresses according to plan and fulfills all obligations towards the EC. TOG will also establish quality management procedures and ensure project results meet consistent quality standards.
UDA, YORK, CEA, MUN, EHU, ATB, CLMS, IPM, INFT, MLGX, EDP, and CONTI will each contribute by providing required reporting, participating in EC project reviews, participating in risk assessment and mitigation, and providing reviews of deliverables according to quality management guidelines.

Deliverables:
D9.1: Quality Management Plan (TOG) (M3) To ensure the high quality of the project results, this task will deliver and implement a quality and risk-monitoring plan that will establish required procedures and provide for regular monitoring of the operational performance of the project and progress towards its objectives.
D9.2: Data Management Plan (TOG) (M6) Report documenting the consortium's data management plan in compliance with the guidelines provided in Annex 2 of the EC "Guidelines on Data Management in Horizon 2020" document. More specifically, the deliverable will outline how research data will be collected, handled and made available to the community during and after the lifespan of CLOUDMINER. This will be a live document that will evolve beyond its delivery and capture the current status of each data set used in the project including its name and description, any standards that it conforms to, a description of how it is shared with the community (or why it cannot be shared), and a plan for long term archiving and preservation.
D9.3: 1st Interim Project Report (TOG) (M9) Report on technical progress, dissemination and communication actions, any challenges or issues that arose during the first half of the first contractual reporting period, and how they were addressed within the project.
D9.4: 1st Periodic Project Report (TOG) (M18) Report on technical progress, resource utilised, formal cost reports from each partner, assessment of risks and mitigation actions, dissemination and communication actions, and any challenges or issues that arose during the annual reporting period and how they were addressed within the project.
D9.5: 2nd Interim Project Report (TOG) (M24) Report consolidating the technical progress, dissemination and communication actions, any challenges or issues that arose during the first half of the second contractual reporting period, and how they were addressed within the project.
D9.6: Final Project Report (TOG) (M36) Report consolidating the technical progress, resource utilised, formal cost reports from each partner, dissemination and communication actions, and any challenges or issues that arose during the final reporting period and how they were addressed within the project.

3.1.5 Milestones List

There are six major milestones in this project. These work as checkpoints in order to verify that research and development is progressing in the right direction.

Table 16: List of project-wide milestones

No.	Milestone Name	WP(s) involved	Month	Means of verification
1	Requirements and Case Studies Completion	WP1, WP7, WP8, WP9	6	The project requirements have been established, the architectural guidelines for the CLOUDMINER platform have been defined, the project website, brochure and the evaluation and data management plans have been delivered and the 1st interim project report has been produced.

Table 16: List of project-wide milestones

No.	Milestone Name	WP(s) involved	Month	Means of verification
2	CLOUDMINER Languages Definition Completion	WP2, WP3, WP4, WP5, WP6, WP9	12	The tools for the graphical development of data pipelines have been developed, the CLOUDMINER knowledge base has been delivered, and the language for data pipeline deployment has been produced. The 1st annual project report has been produced.
3	CLOUDMINER Platform and Methodology - Initial Version	WP2, WP3, WP4, WP5, WP6, WP7, WP8, WP9	18	The initial dissemination plan has been delivered. The automated data tagging mechanisms have been developed. The natural language support and intelligent recommenders for pipeline development have been delivered. The smart data visualization and federated mining tools have been produced. The data pipeline deployment and execution environment have been developed and integrated in the initial version of the platform. The evaluation methodology and the 2nd interim project report have been also delivered.
4	CLOUDMINER Platform and Methodology - Interim Version	WP3, WP6, WP7, WP8, WP9	24	The interim evaluation has been conducted by use-case providers. The data preprocessing supporting tools have been delivered. The graphical tools for data pipeline monitoring and debugging have been produced. Advanced tools for semantic data mining have been developed. Press and media materials have been released, and the 2nd annual project report has been delivered.
5	CLOUDMINER Platform Ready for Evaluation	WP2, WP3, WP4, WP5, WP6, WP7	30	All the data preprocessing facilities have been delivered. The tools supporting collaborative development of data pipelines and the semantic query answering tool has been produced. The intelligent recommender supporting the optimized execution of data pipelines have been delivered. The dashboards providing insights about data pipeline executions have been released. The platform has been tested and delivered to use-case providers.
6	Project completion	WP2, WP7, WP8, WP9	36	The case studies have been completed, feedback has been incorporated back to the CLOUDMINER platform and its components, the final dissemination and use plan has been produced, and the standardisation and final project reports have been delivered.

3.1.6 Project Risks

An initial list of identified risks is presented in Table 17; more detailed assessment of risks will be carried out regularly during the execution of the project. As a general prevention measure, we will continuously monitor the progress of the project in order to detect and react early to any problems that may occur.

Table 17: Identified project risks

Description of risk	Level of likelihood	WP(s) involved	Proposed risk-mitigation measures
Technical Risks			
Recommendations of data preprocessing tasks and algorithms are not relevant.	Medium	WP2	The proposed recommendation algorithms will be developed in an incremental manner by continuously validating them in tight collaboration with the use case partners and by considering real scenarios. Interactive recommendations will be considered if needed.
Data preprocessing components underperform in terms of accuracy and running time.	Low	WP2	We will adopt an incremental development process with iterative evaluations to ensure early detection and rectification of any deficiencies. If necessary, additional resources will be allocated to reinforce development and testing of these components.
The graphical tool is not expressive enough for covering types of pipelines which diverge substantially from the use-cases of the project.	Medium	WP3	The graphical tool will be developed in an incremental manner, taking into account diverse types of data pipelines from relevant literature in addition to the project's use-cases.
The ability to merge conflicting pipelines becomes too complex to use effectively.	Low	WP3	Merging capabilities will leverage knowledge and lessons learnt from current state-of-the-art text and model versioning systems to ensure that it remains usable and offers sufficient semi-automated capabilities, with user interaction being guided appropriately.
The developed NL interfaces have low accuracy for recognizing NL text.	High	WP3, WP4	This risk would be more acute when using NL to create data processing pipelines concerning arbitrary data domains. However, we will base our work on existing NL parsers, libraries and frameworks. To simplify our assumptions, we will focus on the English language, but will create a generic design where other languages can be plugged in. We will perform frequent evaluations with both native and non-native English speakers.

Table 17: Identified project risks

Description of risk	Level of likelihood	WP(s) involved	Proposed risk-mitigation measures
Data visualizations are not performant with large data.	Low	WP4	Data visualization capabilities will make use of best in class visualization libraries, which feature asynchronous data loading as well as loading on demand. Real time data summarization techniques will be developed to provide high-level and actionable visualizations.
No ontology and/or ontology mapping is provided to the semantic data mining tool, or these are of poor quality.	Medium	WP4	The proposed data mining tool will operate with or without background knowledge. We will provide brief video tutorials that demonstrate how to find relevant ontologies on the web, or otherwise construct simple ontologies that enable users to experiment with this feature.
The recommendations provided by the CLOUDMINER recommender systems are not relevant.	Medium	WP5	The recommendation systems will be developed in an iterative manner by continuously validating it in tight collaboration with the use case partners and by considering real scenarios.
The knowledge base does not contain enough data to train the machine learning algorithms	Medium	WP5	Close collaboration with the research and especially industrial partners to identify significant data sources to be mined.
Deployment models may express instantiation requirements that are not satisfiable by any of the evaluated or developed tools.	Low	WP6	WP6 will work closely with WP3 and WP5 to ensure the feasibility of all aspects relating to the deployment language, ensuring that transformations from pipeline specifications produce deployment models that express deployable configurations
Expressivness of Pipeline Execution Dashboard not good enough to visualise DevOPs and DataOps concepts.	Medium	WP6	The dashboard will developed in an agile manner allowing a continuous evaluation.
Heterogeneity of independently developed tools makes them impossible to integrate.The tools will be poorly integrated or not integrated at all.	Low	WP7	The common architectural and designed guidelines will be established early on in the project, and their observance will be monitored by integrating interim releases of the produced tools throughout the duration of the project, and reporting misalignments back to developers.
Independently developed tools are late. The integrated platform is not adequately tested and feature-complete.	Low	WP7	Interim versions of the tools will be integrated frequently so that even if the final version of a tool is not ready by the planned deadline, a tested version can still be included in the platform.
Management Risks			
Loss of a partner means that some results cannot be achieved.	Low	WP2-7	Several partners will be involved in high-risk tasks to avoid single points of failure.
Communication problems between partners or work packages can cause delays in collaborative software development and in the delivery of results.	Low	WP2-7	Kick-off meeting will be held to establish personal contacts; Project Handbook for the day-to-day management of the project will be set up.
Lack of resources and/or personnel changes forced upon the project by one or more partners can impact the quality of the outcomes of the project.	Low	WP2-7	Raise the issue urgently with management in partner organisations. In consultation with the commission, consider whether a replacement partner can be sought.

3.1.7 Resources to be Committed

3.1.7.1 Summary of Effort and Costs

The following table provides a summary of the overall effort. The table indicates the number of person months over the duration of the planned work, for each work package and for each participant.

Table 18: Summary of efforts per partner and work package

No	Name	WP1	WP2	WP3	WP4	WP5	WP6	WP7	WP8	WP9	Total
1	TOG	5	0	0	0	0	0	4.5	8	14	31.5
2	UDA	4	3	5	3	18	6	5	4	5	53
3	YORK	4	0	36	0	7	14	5	4	1.4	71.4
4	CEA	4	43	0	0	0	0	5	4	1.3	57.3
5	MUN	4	8	5	27	0	0	5	4	1.3	54.3
6	EHU	4	9	15	15.8	0	0	5	7.5	1.3	57.6
7	ATB	4	0	7	0	3	19	5	4	1.3	43.3
8	CLMS	4	3.5	3	3	3.5	3.5	29.5	4	1.5	55.5
9	INFT	3.5	0	0	0	0	0	30	4	1.3	38.8
10	MLGX	3.5	0	0	0	0	0	29	3.5	1.3	37.3
11	EDP	3	0	0	0	0	0	12	2	1.3	18.3
12	IPM	1	2	0	0	0	0	10	2	1.3	16.3
13	CMM	2	0	0	0	0	0	6	1	0	9
14	CONTI	4	0	0	0	0	0	21	3	1.5	29.5
Total		50	68.5	71	48.8	31.5	42.5	172	55	33.8	573.1

3.1.7.2 Purchase Costs, Subcontracting and Associated partner

Project partner MLGX from Switzerland participates as an Associate Partner with financial support from government agency SBFI support the following costs in Euros: Personnel: 235,027, Travel: 21,391, Equipment: 11,700, Indirect: 67,028, Total: 335,146. Partners TOG, EDP and IPM have Purchase costs that exceed 15% of personnel costs itemised in the following tables.

1/TOG	Cost (€)	Justification
Travel	20,219	Travel expenses to participate in project meetings, dissemination actions, and EC reviews.
Open Access Fees	29,310	Open access fees incurred by research and development partners. Amounts will be transferred to respective partners as journal articles/papers are accepted for publication.
Total	**49,529**	

11/EDP	Cost (€)	Justification
Travel	22,481	Travel expenses to participate in project meetings, dissemination actions, and EC reviews.
Total	**22,481**	

12/IPM	Cost (€)	Justification
Travel	22,191	Travel expenses to participate in project meetings, dissemination actions, and EC reviews.
Total	**22,191**	

Subcontracting: None of the consortium partners intend to use subcontracting to carry out tasks within the project.

3.2 Capacity of Participants and Consortium as a Whole

The consortium as a whole provides the project with the expertise and experience in all the key areas that are required to achieve the project's objectives. Refer to Section 1.1.2 for an overview of the project's key areas and how they are matched by the contributing core competencies of project partners. The partners' core competencies are complementary to each other. The partners are suited and strongly committed to their tasks and their competencies match the tasks they have been assigned to in the project. Moreover, each partner has a strong interest in successfully completing the project, to consolidate their positions and competence in the rapidly growing area of extreme-data mining environments. The consortium consists of four enterprises active in the domains of traffic information and road maintenance management, weather forecasting, smart cities, and automotive, world-leading universities and research centers, and a global consortium that manages open IT standards. This variety ensures that the project is very well grounded in the needs and the know-how that industry has in data mining and analytics. The fourteen partners are from nine countries. All partners have participated in (and several have coordinated) EU projects in the past. Most of the partners have already successfully collaborated with each other in previous projects too. These relationships strengthen the consortium and facilitate good communication and collaboration within and between work packages.

Project partners, namely TOG and YORK, are actively involved in standardization organizations, such as the OMG. OMG is a member of TOG and TOG manages the certification and branding programmes for OMG. YORK has previously contributed to the OMG UML standard and QVT standard. Currently it contributes to the OMG Software Assurance RFP which is producing software assurance metamodels. This work is part of the Software Systems Engineering Initiative (SSEI) based at York. YORK has also contributed to the development of the AADL standard, and is an active member of the Eclipse Foundation, which develops de-facto industry standards and implements de-jure (e.g. OMG) standards in the field of software modelling and Model Driven Engineering. Project partners have ongoing or past cooperation activities with several European Commission or national governments funded research projects in related topics such as domain modelling and domain specific language development, scalable cloud-based infrastructures, data migration, text mining and natural language processing.

3.2.1 Project's Key Areas and Partners' Contributing Expertise

- *Extraction and Quality Management for Extreme Data*

  CEA: Expertise in data engineering and data analysis pipelines. Expertise in distributed data analysis platforms.

  UDA: Expertise in mining software repositories.

  MUN: Expertise in FAIR data, data mining and predictive analytics.

  CLMS: Expertise in development of data management applications.

  IPM: Expertise in development of complex data-driven applications.
- *Polyglot Data Pipeline Development*

  YORK: Expertise in development of graphical, textual and hybrid model-based software engineering workbenches.

  UDA: Expertise in development of modeling environments and model management tools.

  CLMS: Expertise in development of graphical and textual model-based tools in a production environment.

  ATB: Expertise in development of cloud-based software-engineering tools.

  MUN: Expertise in machine and deep learning.

  EHU: Expertise in natural language processing tools.
- *Advanced Analytics Techniques for Deep Data Mining*

  MUN: Expertise in FAIR data, formal knowledge representation and reasoning, and machine learning and deep learning for data mining and predictive analytics.

  EHU: Expertise in natural language processing, machine learning and deep learning for natural language and other data types.

  UDA: Expertise in mining software repository, supervised and unsupervised learning.

  CLMS: Expertise in applying machine learning tools to mine big data.
- *Intelligent Recommenders for Data Mining Assistance*

  UDA: Expertise in development of recommendation systems for supporting the development of complex software systems.

  ATB, **CLMS**, and **YORK:** Expertise in container-based microservice architectures.
- *Optimised Data Pipeline Execution*

ATB and **CLMS**: Expertise in microservice architectures and standards for defining service architecture, protocols and QoS. Expertise in development of graphical tools to create container deployments.

YORK: Expertise in distributed data processing (e.g. messaging middleware, efficient data exchange formats) and in container-based microservice architectures.

UDA: Expertise in recommender systems for supporting different software engineering tasks.

3.2.2 Partner Roles in Project

TOG will provide **overall coordination of the project**, and as an industry organisation with membership that includes most of the largest IT technology vendors and many purchasers of IT, will lead the dissemination and exploitation work package. TOG will also provide additional requirements for low-code data mining and analytics from organisations outside of the consortium ensuring technologies can be taken-up by a range of industries for many types of applications.
UDA will be the **technical coordinator of the project**, and ensure that technical work performed in CLOUDMINER is of high quality and well-aligned with the requirements of industrial partners as they have been specified in WP1. UDA will lead the technical work on CLOUDMINER intelligent model recommenders developed in WP4, and contribute to the platform integration effort. UDA will also build on their experience with leading and contributing research projects. UDA will also collaborate in WP2 based on their experience in the construction of modeling environments and in WP6 to ensure the adoption of recommenders for optimizing the execution of data pipelines.
YORK will lead the technical work on the polyglot data pipeline development environment produced in WP3 and they will develop facilities for debugging and monitoring pipelines executed using the facilities produced by WP5. YORK will also build on their experience with leading high-visibility open-source projects (e.g. www.eclipse.org/epsilon) to establish the open-source project through which the technical results of CLOUDMINER will be made available to the community.
CEA will lead the technical work on CLOUDMINER data preprocessing components (including tasks for data extraction, validation, cleaning, and transformation) developed in the context of WP2. CEA will also contribute the development and maintenance of data mining pipelines in WP3. CEA, will build on their experience with leading and contributing to open source tools and projects, e.g. Papyrus (https://www.eclipse.org/papyrus/), that are used extensively by industry and academia. Finally, CEA will capitalise on its long experience on leading and participating in large collaborative national and international research projects including Confiance.ai (https://www.confiance.ai) and DIH4AI (https://www.dih4ai.eu).
MUN will lead WP4 and lead the technical work on semantic data mining and their semantic data visualization. MUN will build on their experience in platform integration (e.g., NIH/NCATS Biomedical Data Translator https://ncats.nih.gov/translator) and in the development of FAIR data systems (e.g., EOSC-Life https://www.eosc-life.eu/).
EHU will leverage the long experience of its members in natural language processing, data analytics, machine learning and deep learning to lead the design and development of a natural interaction assistant for designing data processing pipelines in Task 2.3, a chatbot for data and processing result exploration using natural language in Tasks 4.4 and pipeline components for federated machine learning in Task 4.2. EHU will be involved in WP2, WP5 and WP6 to support technical work leaders concerning natural language and machine learning. EHU will build on previous experience in participating in several national and EC projects including CROSSMINER, TYPHON, OSSMETER, CHARM, TechUP and CyberGatE.
ATB will lead the technical work on the optimised data pipeline execution in WP6. ATB will collaborate in WP3 on the pipeline development toosl and in WP5 to include recommenders in deployments for optimizing the execution of data pipelines. ATB will also contribute to the platform integration in WP7.
CLMS will lead the technical work on Platform Integration and Evaluation.
INFT will provide a use case demonstrator from the traffic information and road maintenance management domain. It will also contribute throughout the project with its strong business experience in handling traffic data and developing relevant services.
MLGX will provide a use case demonstrator from the weather forecasting domain. It will also contribute throughout the project with its strong business experience in handling meteorological data and developing relevant services.
EDP will provide a use case demonstrator from the smart cities domain. It will also contribute throughout the project with its strong business experience in handling energy data and developing relevant services.
IPM will provide engineering support to EDP.
CMM will provide smart city services in the context of the EDP use case.
CONTI will provide a use case demonstrator from the manufacturing domain for verifying the project results. CONTI will also contribute throughout the project with its strong business experience in handling manufacturing data and developing internal services around this data.

References

[1] T. De Bie, L. De Raedt, J. Hernández-Orallo, H. H. Hoos, P. Smyth, and C. K. I. Williams. "Automating Data Science". In: *Communications of the ACM* 65.3 (Mar. 2022), pp. 76–87.

[2] D. Petcu et al. "On Processing Extreme Data". In: *Scalable Computing: Practice and Experience* 16.4 (Jan. 2016), pp. 467–490.

[3] C. Li. "Preprocessing Methods and Pipelines of Data Mining: An Overview". In: *arXiv:1906.08510 [cs, stat]* (June 2019). arXiv: 1906.08510 [cs, stat].

[4] A. R. Munappy, J. Bosch, and H. H. Olsson. "Data Pipeline Management in Practice: Challenges and Opportunities". In: *Product-Focused Software Process Improvement*. Ed. by M. Morisio, M. Torchiano, and A. Jedlitschka. Vol. 12562. Cham: Springer International Publishing, 2020, pp. 168–184.

[5] D. Dennison and T. Harvey. "Data Processing Pipelines". In: *Site Reliability Engineering: How Google Runs Production Systems*. 2016.

[6] F. Nargesian, E. Zhu, R. J. Miller, K. Q. Pu, and P. C. Arocena. "Data lake management: challenges and opportunities". In: *Proceedings of the VLDB Endowment* 12.12 (2019), pp. 1986–1989.

[7] T. Furche, G. Gottlob, L. Libkin, G. Orsi, and N. Paton. "Data wrangling for big data: Challenges and opportunities". In: *Advances in Database Technology—EDBT 2016: Proceedings of the 19th International Conference on Extending Database Technology*. 2016, pp. 473–478.

[8] B. Golshan, A. Halevy, G. Mihaila, and W.-C. Tan. "Data integration: After the teenage years". In: *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI symposium on principles of database systems*. 2017, pp. 101–106.

[9] S. Grafberger, J. Stoyanovich, and S. Schelter. "Lightweight inspection of data preprocessing in native machine learning pipelines". In: *Conference on Innovative Data Systems Research (CIDR)*. 2021.

[10] E. Caveness, P. S. GC, Z. Peng, N. Polyzotis, S. Roy, and M. Zinkevich. "Tensorflow data validation: Data analysis and validation in continuous ml pipelines". In: *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 2020, pp. 2793–2796.

[11] S. Schelter, F. Biessmann, D. Lange, T. Rukat, P. Schmidt, S. Seufert,

P. Brunelle, and A. Taptunov. “Unit testing data with deequ”. In: *Proceedings of the 2019 International Conference on Management of Data*. 2019, pp. 1993–1996.

[12] K. Morik, H. Kotthaus, L. Heppe, D. Heinrich, R. Fischer, A. Pauly, and N. Piatkowski. “The care label concept: a certification suite for trustworthy and resource-aware machine learning”. In: *arXiv preprint arXiv:2106.00512* (2021).

[13] R. Cordingly, H. Yu, V. Hoang, D. Perez, D. Foster, Z. Sadeghi, R. Hatchett, and W. J. Lloyd. “Implications of Programming Language Selection for Serverless Data Processing Pipelines”. In: *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*. 2020, pp. 704–711.

[14] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. “From Data Mining to Knowledge Discovery: An Overview”. In: *Advances in Knowledge Discovery and Data Mining*. USA: American Association for Artificial Intelligence, 1996, 1–34.

[15] N. W. Grady. “KDD meets big data”. In: *2016 IEEE International Conference on Big Data (Big Data)*. IEEE. 2016, pp. 1603–1608.

[16] C. Molnar, G. Casalicchio, and B. Bischl. “Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges”. In: *ECML PKDD 2020 Workshops*. Springer International Publishing, 2020, pp. 417–431.

[17] M. Atzmueller. “Declarative Aspects in Explicative Data Mining for Computational Sensemaking”. In: *Proc. International Conference on Declarative Programming (DECLARE)*. Heidelberg, Germany: Springer, 2018, pp. 97–114.

[18] D. Dou, H. Wang, and H. Liu. “Semantic Data Mining: A Survey of Ontology-based Approaches”. In: Feb. 2015, pp. 244–251.

[19] N. Lavrač and A. Vavpetič. “Relational and Semantic Data Mining”. In: *Logic Programming and Nonmonotonic Reasoning*. Ed. by F. Calimeri, G. Ianni, and M. Truszczynski. Cham: Springer International Publishing, 2015, pp. 20–31.

[20] H. Liu. “Towards semantic data mining”. In: *In Proc. of the 9th International Semantic Web Conference (ISWC2010*. 2010.

[21] P. Ristoski and H. Paulheim. “Semantic Web in data mining and knowledge discovery: A comprehensive survey”. In: *Journal of Web Semantics* 36 (2016), pp. 1–22.

[22] J. Kralj, A. Vavpetic, M. Dumontier, and N. Lavrac. “Network Ranking Assisted Semantic Data Mining”. In: *Bioinformatics and Biomedical Engineering - 4th International Conference, IWBBIO 2016, Granada, Spain, April 20-22, 2016, Proceedings*. Ed. by F. M. O. Guzman and I. Rojas. Vol. 9656. Lecture Notes in Computer Science. Springer, 2016, pp. 752–764.

[23] J. Kralj, M. Robnik-Sikonja, and N. Lavrac. “NetSDM: Semantic Data Mining with Network Analysis”. In: *J. Mach. Learn. Res.* 20 (2019), 32:1–32:50.

[24] C. Sirichanya and K. Kesorn. “Semantic data mining in the information age: A systematic review”. In: *International Journal of Intelligent Systems* 36 (2021), pp. 3880 –3916.

[25] A. K. C. Wong, P.-Y. Zhou, and Z. A. Butt. “Pattern discovery and disentanglement on relational datasets”. In: *Scientific Reports* 11.1 (2021), p. 5688.

[26] K. Zhu, H. Wang, H. Bai, J. Li, Z. Qiu, H. Cui, and E. Chang. “Parallelizing Support Vector Machines on Distributed Computers”. In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt, D. Koller, Y. Singer, and S. Roweis. Vol. 20. Curran Associates, Inc., 2007.

[27] M. Guillame-Bert and O. Teytaud. “Exact Distributed Training: Random Forest with Billions of Examples”. In: *CoRR* abs/1804.06755 (2018).

[28] X. Wang, Y. Han, V. Leung, D. Niyato, X. Yan, and X. Chen. *Edge AI: Convergence of Edge Computing and Artificial Intelligence*. Jan. 2020.

[29] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, and Y. Zhou. “A Hybrid Approach to Privacy-Preserving Federated Learning”. In: *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*. AISec’19. London, United Kingdom: Association for Computing Machinery, 2019, 1–11.

[30] K. Bonawitz, P. Kairouz, B. Mcmahan, and D. Ramage. “Federated Learning and Privacy”. In: *Commun. ACM* 65.4 (2022), 90–97.

[31] L. Erlenhov, F. G. de Oliveira Neto, R. Scandariato, and P. Leitner. “Current and Future Bots in Software Development”. In: *Proceedings of the 1st International Workshop on Bots in Software Engineering*. BotSE ’19. Montreal, Quebec, Canada: IEEE Press, 2019, 7–11.

[32] C. Lebeuf, M.-A. Storey, and A. Zagalsky. “Software Bots”. In: *IEEE Software* 35.1 (2018), pp. 18–23.

[33] S. Pérez-Soler, E. Guerra, and J. de Lara. “Collaborative Modeling and Group Decision Making Using Chatbots in Social Networks”. In: *IEEE Software* 35.6 (2018), pp. 48–54.

[34] R. Paul, J. Arkin, N. Roy, and T. M. Howard. “Grounding Abstract Spatial Concepts for Language Interaction with Robots”. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 2017, pp. 4929–4933.

[35] R. Liu and X. Zhang. “A review of methodologies for natural-language-facilitated human–robot cooperation”. In: *International Journal of Advanced Robotic Systems* 16.3 (2019), p. 1729881419851402. eprint: https://doi.org/10.1177/1729881419851402.

[36] J. Thomason, S. Zhang, R. Mooney, and P. Stone. “Learning to Interpret Natural Language Commands through Human-Robot Dialog”. In: *Proceedings of the 24th International Conference on Artificial Intelligence*. IJCAI’15. Buenos Aires, Argentina: AAAI Press, 2015, 1923–1929.

[37] H. Kaji. “Controlled languages for machine translation: state of the art”. In: *Proceedings of Machine Translation Summit VII*. Singapore, Singapore, 1999, pp. 37–39.

[38] R. Schwitter. “Controlled Natural Languages for Knowledge Representation”. In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. COLING ’10. Beijing, China: Association for Computational Linguistics, 2010, 1113–1121.

[39] *PyTerrier*. Last seen March 2022. 2022.

[40] P. Yang, H. Fang, and J. Lin. “Anserini: Reproducible Ranking Baselines Using Lucene”. In: *J. Data and Information Quality* 10.4 (2018).

[41] *Terrier IR Platform*. Last seen March 2022. 2022.

[42] *PyCarret - low-code machine learning*. Last seen March 2022. 2022.

[43] M. P. Robillard, W. Maalej, R. J. Walker, and T. Zimmermann, eds. *Recommendation Systems in Software Engineering*. en. DOI: 10.1007/978-3-642-45135-5. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[44] B. Dagenais, H. Ossher, R. K. E. Bellamy, M. P. Robillard, and J. P. de Vries. “Moving into a New Software Project Landscape”. In: *Proceedings of the 32Nd ACM/IEEE International Conference on Software Engineering - Volume 1*. ICSE ’10. Cape Town, South Africa: ACM, 2010, pp. 275–284.

[45] M. Mazaheri, G. Kiar, and T. Glatard. “A Recommender System for Scientific Datasets and Analysis Pipelines”. In: *arXiv:2108.09275 [cs]* (Aug. 2021). arXiv: 2108.09275 [cs].

[46] A. Kumar, H. Rasche, B. Grüning, and R. Backofen. “Tool Recommender System in Galaxy Using Deep Learning”. In: *GigaScience* 10.1 (Jan. 2021), giaa152.

[47] S. Naujokat, A.-L. Lamprecht, and B. Steffen. “Loose Programming with PROPHETS”. In: *Proceedings of the 15th International Conference on Fundamental Approaches to Software Engineering*. FASE’12. Tallinn, Estonia: Springer-Verlag, 2012, 94–98.

[48] Y. Gil, V. Ratnakar, J. Kim, P. Gonzalez-Calero, P. Groth, J. Moody, and E. Deelman. “Wings: Intelligent Workflow-Based Design of Computational Experiments”. In: *IEEE Intelligent Systems* 26.1 (Jan. 2011), pp. 62–72.

[49] D. McDermid. *Oracle machine learning for Python user’s guide, release 1.0.* Last seen March 2022.

[50] W. La Cava, H. Williams, W. Fu, S. Vitale, D. Srivatsan, and J. H. Moore. “Evaluating Recommender Systems for AI-driven Biomedical Informatics”. In: *Bioinformatics* 37.2 (Apr. 2021). Ed. by J. Wren, pp. 250–256.

[51] A. Tornede, M. Wever, and E. Hüllermeier. “Extreme Algorithm Selection with Dyadic Feature Representation”. In: *Discovery Science*. Ed. by A. Appice, G. Tsoumakas, Y. Manolopoulos, and S. Matwin. Cham: Springer International Publishing, 2020, pp. 309–324.

[52] N. Marz and J. Warren. *Big data: principles and best practices of scalable real-time data systems*. Manning, 2015.

[53] A. Krettek. *The data processing evolution: A potted history*. https://www.itproportal.com/features/the-data-processing-evolution-a-potted-history//.

[54] J. C. S. Anjos, K. J. Matteussi, P. R. R. De Souza, A. da Silva Veith, G. Fedak, J. L. V. Barbosa, and C. R. Geyer. “Enabling Strategies for Big Data Analytics in Hybrid Infrastructures”. In: *2018 International Conference on High Performance Computing Simulation (HPCS)*. 2018, pp. 869–876.

[55] D. Wu, L. Zhu, X. Xu, S. Sakr, D. Sun, and Q. Lu. “Building Pipelines for Heterogeneous Execution Environments for Big Data Processing”. In: *IEEE Software* 33.2 (2016), pp. 60–67.

[56] H. R. Hartson, T. S. Andre, and R. C. Williges. “Criteria For Evaluating Usability Evaluation Methods”. In: *International Journal of Human-Computer Interaction* 15.1 (2003), pp. 145–181.

[57] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean. “Carbon Emissions and Large Neural Network Training”. In: *arXiv:2104.10350 [cs]* (Apr. 2021). arXiv: 2104.10350 [cs].