An Alternative View on Data Processing Pipelines from the DOLAP 2019 Perspective

Oscar Romero [1], Robert Wrembel [2], Il-Yeol Song [3],

[1]*Universitat Politècnica de Catalunya*
[2]*Poznan University of Technology, Institute of Computing Science, Poland*
[3]*Drexel University, United States*

Abstract

Data science requires constructing data processing pipelines (DPPs), which span diverse phases such as data integration, cleaning, pre-processing, and analysis. However, current solutions lack a strong data engineering perspective. As consequence, DPPs are error-prone, inefficient w.r.t. human efforts, and inefficient w.r.t. execution time.We claim that DPP design, development, testing, deployment, and execution should benefit from a standardized DPP architecture and from well-known data engineering solutions. This claim is supported by our experience in real projects and trends in the field, and it opens new paths for research and technology. With this spirit, we outline five research opportunities that represent novel trends towards building DPPs. Finally, we highlight that the best DOLAP 2019 papers selected for the DOLAP 2019 Information Systems Special Issue fall in this category and highlight the relevance of advanced data engineering for data science.

Key words: data integration, ETL/ELT, ETL optimization, data processing pipeline, metadata, data management, data analytics

1 Introduction

This paper includes the introduction to the special section of this issue of the Information Systems journal. The section contains the four best papers submitted to the *21st International Workshop On Design, Optimization, Languages and Analytical Processing of Big Data - DOLAP* 2019 Workshop, held in Lisbon, Portugal on March 26, 2019, in association with the EDBT/ICDT 2019 Joint Conference. The focus of this section is on problems related to

Email addresses: `oromero@essi.upc.edu` (Oscar Romero), `robert.wrembel@put.poznan.pl` (Robert Wrembel), `songiy@drexel.edu` (Il-Yeol Song).

design, development, testing, deployment, and execution of data processing pipelines for data science.

Within the recent years, *data science* and *data analytics* have become one of the most popular research and technological fields. As a consequence, the most wanted IT professions are data scientists and data analysts [8,13], whose main focus is on extracting value from data.

Extracting value from data is a complex task that encompasses multiple disciplines. Specifically, *data science* is the scientific *"interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured"* [9]. Knowledge is typically extracted by means of various *data analytics* algorithms and tools. Such tasks are supported by *data engineering* techniques that *"design and build the data ecosystem that is essential to analytics. Data engineers are responsible for the databases, data pipelines, and data services that are prerequisites to data analysis and data science"* [18]. Finally, *data management* encompasses processes and technologies for managing the whole lifecycle of data, from collecting and integrating data, storing them persistently, providing secure and efficient access, assuring data safety in case of a system crash, to archiving data [12,15,16].

Despite an increasing popularity of data science and data analytics, the fundamental data engineering techniques are typically neglected in the end-to-end data processing conducted by data scientists. This observation is confirmed not only by the experience of the authors of this paper in numerous data analytics projects, but also by trends in the field. Figure 1, drawn with the support of *Google trends*, shows the popularity of 4 topics, namely: data management, data engineering, data analytics, and data science, within the recent 5 years. As we can see, data engineering problems and technologies have received relatively little attention from the Internet users.

Furthermore, current job markets analyses highlight that the industry requires more efficient, automatable, and easy to deploy analytical processes. These characteristics are inherently related to the role of data engineering solutions in such systems. Indeed, the Crunch Report [13], which analyses the current status of the job market in the USA says: "*by 2020 the number of positions for data and analytics talent will increase by 364,000 openings, to 2,720,000. From this, just 61,799 represent advanced analytical roles and the rest require an increasing demand of both data management and analytical skills. Indeed, in 2016, the average salary for a data engineer was of $117,000 compared to $114,000 of data analysts.*" Unfortunately, there is a big gap between what is requested by companies and the job market reality [7]. The International Data Corporation (IDC) estimated that in 2016 the EU had 6 million data workers who collect, store, manage, or analyse data as their primary activity, but around 400,000 jobs could not be filled due to a lack of data workers [8].

Creating value from data is done by deploying *data processing pipelines* (DPPs),

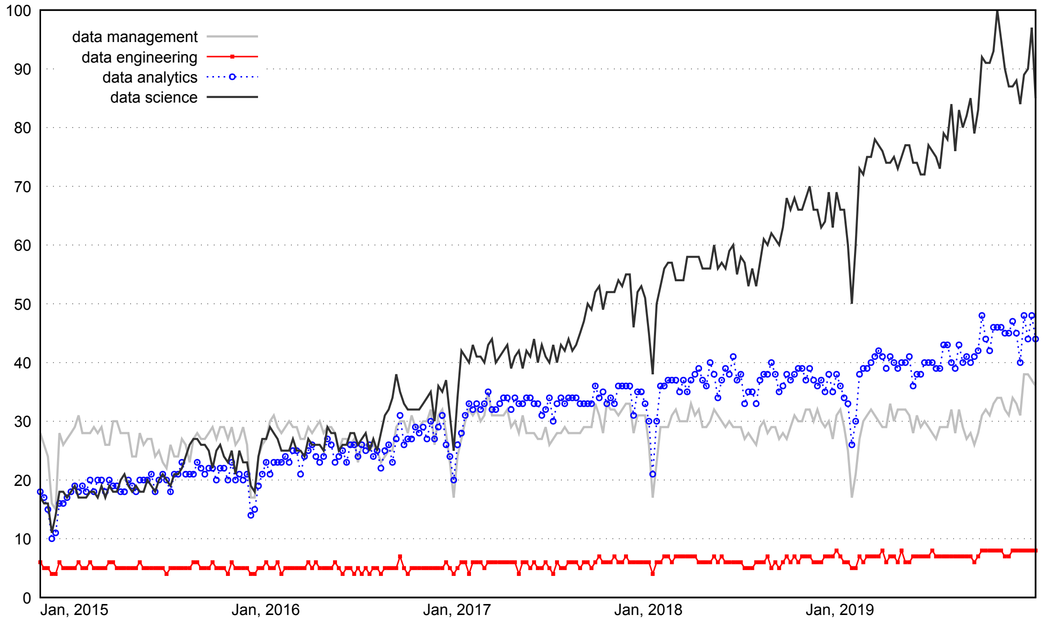

Fig. 1. The popularity of *data management*, *data engineering*, *data analytics*, and *data science* topics within the recent 5 years (by *Google trends*). Numbers represent search interest relative to the highest point on the chart, for the given region and time, i.e., value 100 represents the highest popularity.

which typically include two types of tasks, namely: data preparation and data analysis.The main data preparation tasks include: (1) integrating heterogeneous and distributed data as well as transforming them into a common model and representation, (2) cleaning and standardizing data, (3) eliminating duplicates, and (4) storing cleaned data in a centralized repository.

The main data analysis tasks include: (5) extracting a data view from a centralized repository (and potentially storing it in a data view: i.e., another database, data mart), (6) specific pre-processing for the analytical task (algorithm) at hand, in order to increase for example prediction accuracy, (7) creating test and validation data sets, which also include labeling data for supervised machine learning, and (8) analyzing the data by means of either descriptive statistical analysis (e.g., reporting or OLAP) or by advanced data analysis, including among others predictive models (e.g., machine learning or data mining [10,11]) or graph data analytics [14].

As reported by experts in the field [1–3], up to 80% of time is spent on tasks (1)-(3) and (7) of a DPP. This is due to the fact that DPPs have not been standardized yet and practitioners build DPPs in ad-hoc manners. For this reason, a research and technological challenge lies in developing techniques and architectures for semi-automatic design of DPPs. To this end, standardization is needed [7]. As stressed in [5], new data management techniques are required to adapt to the new analytical settings but building up on the already existing data engineering techniques. Moreover, the Beckman Report [4] also advocates for a more comprehensive and automated DPP.

A successful story on applying standardization is that of Business Intelligence (BI). For over two decades, a reference architecture for data integration and analytics is being applied by companies. In this architecture, called a *data warehouse architecture* (DWA) [6,17], heterogeneous and distributed data sources are integrated by means of a layer called *extract-transform-load* (ETL) or its *extract-load-transform* (ELT) variant. This layer is composed of processes (workflows) that are responsible for preparing data for further analysis.

We can mirror the current situation of data science to that of BI. For example, the multidimensional model was developed to standardize OLAP analysis and thanks. Therefore, we can nowadays benefit from OLAP engines (e.g., MicroStrategy, Mondrian, Hyperion) that fully automate the generation of dynamic reports by changing the data granularity or slicing them. Other aspects of BI, including the creation and maintenance of ETL processes, are too complex to standardize, even by the most advanced tools (e.g., IBM Data Stage, Informatica PowerCenter, Ab Initio). However, all these ETL tools provide integrated design, test, and run-time environments, equipped with pre-defined building blocks (standardized to some extent) and standardized core functionalities.

Nowadays, companies deploy multiple DPPs to tackle each of the aforementioned tasks, and each of these sub-DPPs is typically run in a separate system. Data management tools, e.g., Apache NiFi, HDFS, HBase, Spark, are used for tasks (1)-(4), while data are typically stored in OS files. Analytical tools, e.g., SAS, R, Statistica, IBM SPSS, or Python scripts, are used for tasks (5)-(8).

To sum up, in the current state of research and technologies supporting data science, the data analytics tasks and data engineering tasks are handled separately and for this reason:

- data transformations along a DPP (from a data source to a destination) cannot be tracked and recorded, i.e., it is not possible to implement data lineage;
- multiple technologies used in a DPP require multiple skills from a designer, which decreases his/her productivity;
- multiple technologies cause that data need to be transferred between various tools (execution environments), which decreases performance due to multiple (and unnecessary) I/O and type conversions;
- multiple tasks in a DPP need to be orchestrated manually;
- there is no mean for optimizing the tasks as they are distributed among different tools and technologies.

As a consequence, designing a DPP is error-prone, inefficient w.r.t. human efforts, and inefficient w.r.t. execution time of the whole DPP. With this respect, a DPP design, development, testing, deployment, and execution could profit from a standardized DPP architecture and from data engineering solutions. This claim opens new paths for research and technology, as outlined in Section 2.

2 Open research issues

The efficiency of designing, maintaining, and executing a DPP can benefit from a **a standardized layered architecture**, including a conceptual, logical, and physical view of the DPP. The *conceptual layer* should be designed by means of a unified processing language that would standardize the common data engineering and data analysis tasks. This layer abstracts from an implementation. The *logical layer* would represent concepts from the previous layer in appropriate technologies, including: a given type of DBMS (e.g., relational, key-value, graph), and an appropriate implementation language for each DPP task. At this level, the workflow is optimized by techniques such as task reordering and parallelization. Finally, the *physical layer* defines physical components, and their parameters. It includes among others: (1) memory size of workstations, (2) the number of CPUs and threads, (3) physical data layout (e.g., column store or row store), (4) physical data structures in a database, (5) if applicable, the size of a cluster of workstations.

Such an architecture opens multiple research directions, including:

- *Unified metadata storage and management techniques* in order to support data profiling, provenance, and execution optimization. To this end, a standardized metadata representation is needed with method assuring metadata gathering and consistency maintenance.
- *Optimization of a DPP* in order to guarantee acceptable execution time. To this end, the whole DPP needs to be optimized using similar techniques to cost-based query optimization. The problem of DPP optimization is computationally complex and it resembles ETL optimization, which still is an open research and technological problem.
- *Automatic transformation* from the conceptual to logical and from the logical to physical layer. The transitions are guided by requirements (in the spirit of service level agreement), e.g., execution time, monetary cost, or quality of analytical results (quality of models). Developing cost models for such an architecture is challenging and it is an unexplored area.
- *Collaborative exploration of data* in order to increase the performance of analyst. To this end, the architecture must provide means for sharing and reusing results of analyses done so far. Moreover, the results must be easy to annotate and provide provenance. Some of the existing data engineering technologies, like databases and materialized views may be applicable to this problem. Moreover, the system should guide an analyst in the process of data exploration, similarly as in recommender systems. Even though, such technologies exist, they still need to be integrated.
- *Meaningful and innovative visualization and exploitation scenarios* in order to provide meaningful insights into analytical results. Existing visualization techniques like numerous charts or spatial objects may help in this problem, but with a constantly increasing volume of analyzed data, new techniques may be needed.

3 Special issue content

The papers presented in this special section cover problems related to a DPP: from developing a DPP and assuring data quality, through DPP task orchestration, to the last step - data analysis. Thus, they address some of the open issues mentioned in Section 2. These papers are the extensions of the original top 4 papers accepted at DOLAP 2019.

Data preparation

The first paper entitled *Feedback Driven Improvement of Data Preparation Pipelines*, by N. Konstantinou, N.W. Paton, addresses the problem of designing a DPP with the final goal to obtain data of high quality. The problem is solved with the following steps. First, data being produced by a given DPP are labeled either as a true positive or true negative. Labeling is done explicitly by a user. Second, some hypotheses are formed on why some data are incorrect. Third, the significance of these hypotheses is tested by statistical methods. Finally, based on the statistically significance hypotheses, a DPP is refined or reorganized in order to produce data of higher quality. The DPP is reorganized by two algorithms proposed in this paper.

Data pre-processing optimization

The second paper entitled *Two-stage Optimization for Machine Learning Workflow*, by A. Quemy, addresses the problem of an automatic construction of a learned DPP by using machine learning. A DPP is composed of data pre-processing tasks and model building tasks. Thus, a search space of DPP tasks reordering and parametrization is divided into the following tasks: the data pipeline construction and configuration, and a machine learning algorithm selection and configuration. In this approach a DPP is defined as a directed acyclic graph, where nodes represent tasks or algorithms. Then, the optimization problem extends the CASH problem and adds time constraints. The author proposed four policies to allocate time between the data pre-processing and model building tasks. The approach was extensively evaluated by experiments that showed promising results.

The importance of metadata and collaboration exploitation

The third paper entitled *Detecting coherent explorations in SQL workloads*, by V. Peralta, P. Marcel, W. Verdeaux, A.S. Diakhaby, proposes to analyze SQL workloads by extracting features that characterize SQL queries, with the goal to identify sequences of meaningful exploratory queries that represent the same analytical goal (such queries are called explorations). To this end, first the authors proposed a procedure for extraction of query features. Second, based on the extracted features, they investigate three different techniques for segmenting queries within a given user session, namely: (1) an unsupervised learning - based only on similarity between contiguous queries, (2) a supervised learning - based on transfer learning to reuse a model trained over a dataset where ground truth is available, and (3) weak supervision - based on labeling a training set.

Novel means of data analysis

The fourth paper entitled *A-BI+: A Framework for Augmented Business Intelligence*, by M. Francia, M. Golfarelli, S. Rizzi, applies data analysis (in this case, descriptive analytics based on OLAP) in innovative scenarios such as augmented reality. In particular, the authors propose an architecture for the so-called situated analytics, in which users equipped with virtual reality technologies (e.g., glasses) are provided in real-time with analyses of data data relevant to a location a user is and a context (concrete objects) currently perceived by her. The analyses are provided in the form of reports obtained by running OLAP queries on the most relevant enterprise multidimensional cubes. To make it happen, in this paper the authors: (1) proposed an architecture of a system, (2) explained how a-priori expert knowledge can be modeled by mapping context objects to relevant multidimensional elements, (3) proposed an algorithm to generate queries to build contextual analyses, (4) advocate for applying collaborative filtering approach to learn possible analyses from user feedback.

Acknowledgements. The guest editors of this special issue express their gratitude for the Information Systems journal Editors-in-Chief: Felix Naumann, Dennis Shasha, and Gottfried Vossen for their continuous support in the process of publishing the best DOLAP papers in this journal.

We would like to thank DOLAP 2019 Program Committee for their timely, thorough and committed work towards DOLAP. The high quality reviews generated are a key success factor of DOLAP.

The work of Robert Wrembel is supported from the grant of the Polish National Agency for Academic Exchange, within the Bekker programme.

References

[1] Data Science Report. Technical report, CrowdFlower, 2016.

[2] Data Warehouse Trends Report. Technical report, Panoply, 2018.

[3] Data Engineering, Preparation, and Labeling for AI 2019. Technical report, Cognilytica Research, 2019.

[4] D. Abadi, R. Agrawal, A. Ailamaki, M. Balazinska, P. A. Bernstein, M. J. Carey, S. Chaudhuri, J. Dean, A. Doan, M. J. Franklin, J. Gehrke, L. M. Haas, A. Y. Halevy, J. M. Hellerstein, Y. E. Ioannidis, H. V. Jagadish, D. Kossmann, S. Madden, S. Mehrotra, T. Milo, J. F. Naughton, R. Ramakrishnan, V. Markl, C. Olston, B. C. Ooi, C. Ré, D. Suciu, M. Stonebraker, T. Walter, and J. Widom.

The beckman report on database research. *Communications of the ACM*, 59(2):92–99, 2016.

[5] S. Abiteboul, I. Manolescu, P. Rigaux, M. Rousset, and P. Senellart. *Web Data Management*. Cambridge University Press, 2011.

[6] J. Bolton. *Data Warehousing Essentials*. Larsen and Keller Education, 2019.

[7] F. Consulting. Digital Businesses Demand Agile Integration, 2019.

[8] R. Davies. European Parliament Briefing. Big data and data analytics. The potential for innovation and growth. http://www.europarl.europa.eu/RegData/etudes/BRIE/2016/589801/EPRS_BRI(2016)589801_EN.pdf, 2016.

[9] V. Dhar. Data science and prediction. *Comm. of the ACM*, 56(12):64–73, 2013.

[10] D. Frazzetto, T. D. Nielsen, T. B. Pedersen, and L. Siksnys. Prescriptive analytics: a survey of emerging trends and technologies. *VLDB Journal*, 28(4):575–595, 2019.

[11] IBM. Descriptive, predictive, prescriptive: Transforming asset and facilities management with analytics. www.ibm.com/downloads/cas/3V9AA9Y5.

[12] Informatica. What is data management? https://www.informatica.com/services-and-training/glossary-of-terms/data-management-definition.html#fbid=wJQ9ej17rBf.

[13] W. Markow, S. Braganza, B. Taska, S. M. Miller, and D. Hughes. The Quant Crunch: How the demand for data science skills is disrupting the job market. https://www.ibm.com/downloads/cas/3RL3VXGA, 2017.

[14] M. Needham and A. E. Hodler. *Graph Algorithms: Practical Examples in Apache Spark and Neo4j*. O'Reilly, 2019.

[15] Oracle. What is data management? https://www.oracle.com/database/what-is-data-management/.

[16] SAS. Data management. https://www.sas.com/en_us/insights/data-management/data-management.html.

[17] A. A. Vaisman and E. Zimányi. *Data Warehouse Systems - Design and Implementation*. Data-Centric Systems and Applications. Springer, 2014.

[18] D. Wells. Data engineering coming of age. https://www.eckerson.com/articles/data-engineering-coming-of-age, 2018.