Challenges in Enabling Quality of Analytics in the Cloud

HONG-LINH TRUONG, Distributed Systems Group, TU Wien, Austria
AITOR MURGUZUR, Microsoft, Spain
ERICA YANG, Visual Analytics and Imaging Systems group, STFC Rutherford Appleton Laboratory, UK

CCS Concepts: • **Information systems → Information integration**; • **Computing methodologies → Distributed algorithms**; • **Software and its engineering** → *Development frameworks and environments*;

Additional Key Words and Phrases: Cloud computing, big data analytics, data quality, service management

ACM Reference format:
Hong-Linh Truong, Aitor Murguzur, and Erica Yang. 2018. Challenges in Enabling Quality of Analytics in the Cloud. *J. Data and Information Quality* 9, 2, Article 9 (January 2018), 4 pages.
https://doi.org/10.1145/3138806

9

1 DATA ANALYTICS CONTEXTS AND QUALITY-AWARE DATA ANALYTICS

Currently, domain scientists (DSs) face challenges in managing *quality across multiple data analytics contexts* (DACs). We identify and define quality of analytics (QoA) in dynamic and diverse environments, e.g., based on cloud computing resources for big data sources, as a composition of quality of data (data quality), performance, and cost, to name just the main factors. QoA is a complex matter and not just about quality of data or performance, which are typically considered separately when evaluating existing data analytics frameworks/algorithms. Frequently, the DS needs to utilize multiple frameworks to run different (sub)analytics, and, at the same time, the sub-analytics executed in these frameworks exchange inputs and outputs each other. In these cases, we observe different DACs, where *a DAC refers to a particular situation* in which the DS works with a specific framework to run a sub-analytics carried out by pipeline(s) or tasks in a pipeline. Each DAC has a set of interactions in the following categories:

—*Interactions with data processing frameworks:* Depending on the type of (sub-)analytics within a DAC, the DS could utilize a specific data processing framework. Potential frameworks are for batch processing and data analytics workflows (e.g., Hadoop/MapReduce, well-known scientific workflows (Taylor et al. 2006), Google Dataflow, Oozie, and Airflow), streaming processing (e.g., Storm, Flink, Apex, Spark Streaming, and Azure Stream Analytics), hybrid processing (e.g., Summingbird and Spark), and data operation systems and brokers (e.g., YARN, Mesos, and Kafka) (Sakr et al. 2013; Singh and Reddy 2014). The

Authors' addresses: H.-L. Truong, Distributed Systems Group, Argentinierstrasse 8/184-1, A-1040 Vienna, Austria; email: hong-linh.truong@tuwien.ac.at; A. Murguzur, Microsoft, Paseo del Club Deportivo 1, 28223 Madrid; email: aimurg@microsoft.com; E. Yang, Visual Analytics and Imaging Systems Group, Building R89, Scientific Computing Department, Rutherford Appleton Laboratory, Science and Technology Facilities Council (STFC), Didcot, Oxfordshire OX11 0QX; email: erica.yang@stfc.ac.uk.

https://doi.org/10.1145/3138806

same framework, when instantiated with different configurations, can create different processing offerings. Obviously, different frameworks also provide different offerings for (the same) analytics. These offerings are strongly related to QoA, e.g., data processing granularity (e.g., real streaming or micro-batching), response time, scalability and elasticity (to deal with volume and velocity of data), availability of data quality assessment tools for data types and formats variety, and possibilities of utilizing data cleansing and enrichment in (near)realtime to deal with data veracity. The DS typically selects a framework and controls the framework based on the expected QoA (e.g., expensive, short-running time, and high data quality versus cheap, long-running time, and high data quality).

—*Interactions with different input and output data sources*: Data processing and analytics jobs can deal with different types of input and output data sources due to the variety and veracity of data. Technically, these data sources could be interfaced through different means, such as Database-as-a-Service, Sensor-as-a-Service, distributed file systems, Data-as-a-Service, and data marketplaces. Furthermore, they can have different states, like streaming or static (data at rest), besides other known characteristics, such as volumes and velocity. The DS needs such interactions to dynamically adjust the expected and measured QoA. For example, given a low data quality in the output, the DS might take a new input data to enrich the current analytics to produce a better output, instead of stopping the analytics, which might be expensive due to the amount of resources spent. However, not only the quality of data sources but also the interfaces of data sources and communications during the interactions can strongly influence the QoA.

—*Interactions with different system services for data provisioning, monitoring, and control*: With different frameworks for a data analytics, the DS will interact with several other cloud services for provisioning, monitoring, and controlling computing resources, storage, network functions, and so on. The reasons are as follows: (i) not all underlying data processing frameworks have been equipped with such services for supporting on-demand computing and data resources and (ii) connecting different processing frameworks needs to deal with additional services between these frameworks. Such interactions are needed because, for example, a data quality control and assurance between two sub-analytics to meet the expected QoA will require extra resources and monitoring services to be deployed to assess the quality of data exchanged, which might lead to some problems w.r.t. performance metrics associated with QoA.

The key point in managing these interactions is not just to make sure that the functionality of data processing frameworks is correct (as in the focus of current research) but also to deliver and control the results with the expected QoA, covering *analytics time, cost* and *quality of data*, across these contexts. Quality of data strongly influences analytics time and cost, and vice versa. Such interactions are needed to change QoA, but they also introduce cost and performance overheads. However, current techniques lack (i) capabilities to deal with QoA (mostly, they focus on data quality (Missier et al. 2006), performance (Xue et al. 2016), or costs (Kiran et al. 2015)) and (ii) capabilities to deal with QoA as a whole across multiple frameworks (e.g., how to combine data quality in a framework with processing performance in another framework to create a global view on QoA).

2 RESEARCH CHALLENGES

Challenge 1 – Uniform quality-aware data analytics view: The first challenge is the conceptual model defining QoA for such data analytics involved in multiple data processing frameworks. Several works have discussed data quality in workflows (Hazen et al. 2014; Missier et al. 2006;

Reiter et al. 2011), but they focus on single-workflow frameworks. Our DACs involve different analytics, each associated with a workflow/pipeline executed by different platforms. First, we need novel concepts to represent a uniform view on multi-scale, multi-type data analytics that consist of different sub-analytics and their corresponding data analysis algorithms based on different data processing types, such as batch, streaming, and hybrid processing. The view will also characterize QoA metrics that we must support to ensure the analytics (and its sub-analytics) to meet the expected QoA for the analytics results (defined through performance, data quality, cost, forms of data output, etc.). To this end, we need to leverage runtime performance, data quality, and cost metrics associated with analytics structures, meta-data about data sources (e.g., true-positive coverage, false-positive coverage, and interpretability) and algorithms, and underlying data processing frameworks to define these models. Furthermore, we need to enable the modeling of level of QoA based on characteristics of data volume, velocity, variety, and veracity.

Challenge 2 – Quality of analytics across contexts: When moving from a DAC to another one, we could generate and execute different QoA management processes to carry out suitable actions for resource management, data quality monitoring, and data enrichment. First, we have to develop primitives for assessing and monitoring quality aspects in a (sub-)data analysis. While we could leverage several works for understanding quality associated with computing resources and data (Ousterhout et al. 2015; Batini et al. 2009), a systematically way to develop primitives for data assessment and adjustment goes beyond the capabilities of domain-specific data expert. Together with primitives for resource assessment and control, we could establish primitive action models (PAMs) for data analytics (Nguyen et al. 2015); an action is used to assess and adjust data to meet QoA by instantiating primitives with suitable parameters. Given PAMs and the expected QoA, we need to create suitable data management processes that can be injected and executed along with DACs. We have to research novel techniques to create data assessment and adjustment processes from the quality of input data sources, underlying data integration capabilities, and domain know-how as well as performance information about data processing frameworks and resources and costs. Such processes will include different actions to improve data quality (e.g., by filtering bad data and enriching data by adding better data sources during the data fusion) and to reduce performance overhead (e.g., by leveraging suitable cloud resource control algorithms and optimizing data movement). To support tradeoffs in QoA (the right data output, the performance, the quality of data, etc.) across DACs, first we need to invoke quality assessment processes to obtain quality within a specific DAC and then to invoke suitable adjustment processes that strongly depend on the current computational processing capabilities and data processing algorithms and types of analytics. Second, we need to address challenges in coordinating quality assessment and adjustment between DACs, introducing quality-control feedback loops among sub-analytics to support tradeoffs in QoA.

ACKNOWLEDGMENTS

We thank our colleagues in the SAVVY proposal for fruitful ideas.

REFERENCES

Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.* 41, 3 (2009), 1–52.

Benjamin T. Hazen, Christopher A. Boone, Jeremy D. Ezell, and L. Allison Jones-Farmer. 2014. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics* 154 (2014), 72–80.

Mariam Kiran, Peter Murphy, Inder Monga, Jon Dugan, and Sartaj Singh Baveja. 2015. Lambda architecture for cost-effective batch and speed big data processing. In *Proceedings of the 2015 IEEE International Conference on Big Data (BIG DATA'15)*. IEEE Computer Society, 2785–2792.

Paolo Missier, Suzanne M. Embury, R. Mark Greenwood, Alun D. Preece, and Binling Jin. 2006. Quality views: Capturing and exploiting the user perspective on data quality. In *Proceedings of the 32nd International Conference on Very Large Data Bases*. 977–988.

Tien-Dung Nguyen, Hong Linh Truong, Georgiana Copil, Duc-Hung Le, Daniel Moldovan, and Schahram Dustdar. 2015. On developing and operating of data elasticity management process. In *Proceedings of the 13th International Conference on Service-Oriented Computing (ICSOC'15)*.

Kay Ousterhout, Ryan Rasti, Sylvia Ratnasamy, Scott Shenker, and Byung-Gon Chun. 2015. Making sense of performance in data analytics frameworks. In *Proceedings of the 12th USENIX Symposium on Networked Systems Design and Implementation (NSDI'15)*. 293–307.

Michael Reiter, Uwe Breitenbücher, Schahram Dustdar, Dimka Karastoyanova, Frank Leymann, and Hong Linh Truong. 2011. A novel framework for monitoring and analyzing quality of data in simulation workflows. In *Proceedings of the IEEE 7th International Conference on E-Science (e-Science'11)*. IEEE Computer Society, 105–112.

Sherif Sakr, Anna Liu, and Ayman G. Fayoumi. 2013. The family of mapreduce and large-scale data processing systems. *ACM Comput. Surv.* 46, 1, Article 11 (July 2013), 44 pages.

Dilpreet Singh and ChandanK Reddy. 2014. A survey on platforms for big data analytics. *J. Big Data* 2, 1 (2014). DOI: http://dx.doi.org/10.1186/s40537-014-0008-6

Ian J. Taylor, Ewa Deelman, Dennis B. Gannon, and Matthew Shields. 2006. *Workflows for e-Science: Scientific Workflows for Grids*. Springer-Verlag, Secaucus, NJ.

Ji Xue, Feng Yan, Alma Riska, and Evgenia Smirni. 2016. Scheduling data analytics work with performance guarantees: Queuing and machine learning models in synergy. *Cluster Comput* 19, 2 (June 2016), 849–864. DOI: http://dx.doi.org/10.1007/s10586-016-0563-z

Received March 2016; revised August 2017; accepted August 2017