Quality-aware query processing for big datasets

Project Summary

Data quality is essential for a large variety of applications in many domains. Data quality is commonly defined as the “fitness for use” [1]. The disconnect between data source and data use is one of the prime reasons behind the data quality issues. Researchers classify data quality issues into: (i) identifying duplicate data, (ii) resolving data discrepancies, (ii) imputation of missing data, (iv) link and integrate data coming from different data sources[2].

In Big Data, these issues are particularly challenging due to the characteristics of the sources, known as 5 V's. While the volume, velocity and variety are relatively well-defined and can be measured, value is not well-defined and hard to measure and veracity is a complex theoretical construct [3]. Most of the existing real-life data are dirty, and when it comes to Big Data the quality problem becomes a more urgent issue. Usually researcher have been focused on how to deal with the big volume of data, the quantitative aspect, letting apart the qualitative aspect.

This project aims to propose a framework dealing with quality-aware query processing for large scale datasets. The main challenge is to make trade-offs between the query cost and the answer quality, in the context of Big Data. Big Data require storage and analysis capabilities that can only be addressed by distributed computing systems. The challenge related to big data is then to adapt the quality-driven query processing to a distributed environment, namely: defining distributed query execution plans while dealing with the quality assessment for distributed and scalable data. We will then ensure to adapt the query costs, that involves reducing the communication cost of query processing in distributed environments, to users’ quality requirements.
In this work, we particularly focus on developing a model-driven framework, where queries are answered by models of data using machine learning techniques. The main goal of the proposed framework is to develop an approximate querying engine that is more efficient and more accurate using intelligent techniques, as traditional approaches cannot address the Big Data challenges.

The main research directions to achieve this goal are: (i) apply learning approaches to learn and identify quality rules/metrics and (ii) adapt query processing techniques with the quality constraints.

References

[1] Wang RY. A product perspective on total data quality management. Communications of the ACM. 1998; 41(2 (Feb)):58–65.
[2] Ganti, V., and Sarma, A. D. Data cleaning: A practical perspective. Synthesis Lectures on Data Management, vol. 5, no. 3 (2013), pp. 1–85.
[3] Abiteboul, S., Dong, X.L., Etzioni, O., Srivastava, D., Weikum, G., Stoyanovich, J., Suchanek, F.M.: The elephant in the room: getting value from big data. In: Proceedings of the 18th International Workshop