Batch processing innovative technologies for rapidly increasing historical data

Dr. Cesare Cugnasco - July 22, 2019

Many of you may have heard saying that "[Big] Data is the oil of the 21st century" when explaining the importance of data in modern society, and how it is changing our everyday life. After all, it is not by chances that many of the most valued companies in the world base their business in the harvesting of data[1]. However, this analogy can also give an idea of the engineering challenges and the complexity of the whole process that goes from the collection of raw data to the extraction of useful information.

Figure 1. This is not a data pipeline.

An oil company must invest money and resources to find a suitable petroleum reservoir, build a well, extract and transport the crude oil to a refinery where the oil is transformed into multiple useful products that are later commercially distributed. And of course, every part of this vast infrastructure must be tested and validated to avoid leakages and explosions.

Similarly, any data-driven company must find its "data reservoir", collect and safely store the data, build a data warehouse where to refine the raw information to produce useful knowledge and finally use this final product to interact in the world, whether to better serve customers, improve R&D or reduce costs.

[1] Amazon beats Apple and Google to become the world's most valuable brand

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 780787.

The goal of the I-BiDaaS platform is to simplify the overall process of harvesting information, providing a simple to use interface for both the expert and less "tech-savvy" users. In the context of the overall project, the WP3 activities focus on the refinery of data, transforming the raw information into a valuable asset.

However, the analogy stops here, because unlike any other industrial process, Big Data pipelines are cumulative: the information produced yesterday is used to improve the one generated today. As a result, any system we design must be engineered in such a way it can work with an always increasing mole of work, being able to add more computers and resources at will. And that is precisely the goal of WP3: the implementation and development of batch processing innovative technologies for the rapidly-increasing historical data.

How does WP3 tackle such a challenge? With three main approaches: better algorithms, better data and resource management, and better testing.

The University of Novi Sad (UNSPMF) leads the effort toward the development of innovative distributed Big Data machine learning algorithms (like ADMM, LASSO, Random Forest and many more) that naturally adapt to the architecture of modern data centers, making parallelism and asynchronous communication part of the core design.

The Barcelona Supercomputing Center (BSC) focuses on providing programming frameworks and tools that abstract the complexity of managing and distributing data (Hecuba and Qbeast) and of allocating work between different computing resources (COMPSs) so that what we build for the data of today, will also work tomorrow when the size decuples.

Last but not least, a Big Data platform must be tested, validated, and tuned to support the quantity of data that the system is going to manage now and in the future.

If we use real data, we can ensure the system will work today, but we cannot predict how it will behave in the future. At the same time, using real data often poses serious privacy concerns and thus is unsuitable.

A better way is to use synthetic data: artificially generated data that has the same structure and the same statistical properties of the real one, without exposing any sensible information. To do so, IBM and BSC are working together to generate automatic descriptive statistics of a large dataset and to integrate them with the IBM's Test Data Fabrication platform.

For a more detailed description of the achievements of Work Package 3 and of the current status of work, do not miss the report D3.2, available in the section Deliverables of the I-BiDaaS' website.

Find & Follow us

Website | **Twitter** | **LinkedIn**

Zenodo | **OpenAIRE** | **GitHub**

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 780787.