Middleware for Online Exploration of Big Data

Francesca Bugiotti Benoît Groz Jean-Daniel Fekete

LISN, Université Paris-Saclay

December 2024

1 Context

The web has become the largest repository of datasets, serving many large of them for free but with limited support and uncontrollable latency. For example, the city of New York publishes all its taxi trips online, 3 billion trips and growing. This data can be useful for many use cases, and in particular, to make sense of the habits of the New York population. However, to explore and analyze this dataset and similar ones, they need to be downloaded locally before any visualization, exploration, or analysis can start. Downloading data upfront implies that it takes some time before analysts can see any part of the data. Furthermore, even if the data schema is sometimes described somewhere, understanding the data requires more than the schema. Typically, it requires computing the distribution of quantitative attributes, counting unique values, the minimum and maximum values, and so on. We call this information *metadata*.

With this internship, we want to study the services or API needed to be able to explore and analyze data **progressively** to provide important information quickly with approximations if needed and improve the quality of the information over time until the whole dataset is downloaded. Alternatively, the analyst may decide that she received enough information to answer her question before everything is downloaded and abort the remaining downloading and metadata computation, saving power and CO2 emission. Many progressive explorations and analyses can be performed accurately with only a reasonable sample of data. How much is enough can rarely be guessed beforehand.

Progressive Data Analysis (PDA) is a novel data exploration paradigm meant to start exploration and analyses while data is loaded and processed [2]. It allows dealing with the latency of systems caused by the network and algorithm execution. Latency cannot be reduced on the Internet, and complex analysis algorithms take time to run and produce latency. PDA acknowledges these latency issues and is meant to overcome them by providing results step by step, iteratively, starting with approximate results and improving them until the whole data has been processed or the approximate results already computed are accurate enough for the analyst to answer her question.

The possible uses of this novel service are very broad, including data exploration as well as online machine learning [3] with interactive monitoring.

2 Program

The internship will study the services required to perform progressive data analysis as accurately as possible, providing at first an accurate report about the state of the art in the field. After the first phase of the study, the student will propose a strategy for implementing Progressive Data Analysis (PDA) as a proof-of-concept prototype as a middleware to test some of the services on a few use-case applications. The middleware will be used by a progressive data analysis system, the *ProgressiVis toolkit*.

The target use case will consist of a program that wants to download a dataset; it specifies one or several URLs (the Gaia 3 astronomical dataset describing 2 billion stars is distributed with about 3000 compressed CSV files). The most popular format remains CSV, more or less compressed, but newer formats exist, such as *parquet* and *arrow*, with possible pros and cons.

To visualize or analyze a dataset, the middleware should download data locally and compute its metadata progressively. Without external information, the metadata will only be computed approximately and can be provided to a progressive analysis system connected to the middleware.

In addition to computing metadata, the middleware will also need to *shuffle* the data to improve the quality of visualizations or analyses downstream. Perfect shuffling is only possible when the whole dataset is loaded, but progressive shuffling is possible and enhances the data analysis downstream [1]. Few strategies have been explored for this online shuffling. The internship should explore the alternatives and evaluate their performance.

In addition, another important service that will benefit from the approach is *caching* the downloaded data until the analyst decides that the dataset is not useful anymore or using timeout mechanisms. This caching is probably important to improve the shuffling and keep it parallel to the data and the metadata.

The last deliverable of the internship will be a report written in English that will describe the methodology, detail the experiments, and analyze the results. The report will also include the related work written early on and constantly updated and integrated during the development phase.

3 Practical aspects of the internship

The internship will take place at the LISN laboratory of Université Paris-Saclay, building 650 and 660. Co-advised by Benoît Groz and Francesca Bugiotti from the LahDAK team and by Jean-Daniel Fekete from the Aviz team.

The middleware will be programmed in Python. It can use a high-performance analytical database such as DuckDB [4].

References

[1] Yu Cheng, Weijie Zhao, and Florin Rusu. Bi-level online aggregation on raw data. In *International Conference on Scientific and Statistical Database Management*, SSDBM '17. ACM, 2017.

[2] Jean-Daniel Fekete, Danyel Fisher, and Michael Sedlmair. *Progressive Data Analysis: Roadmap and Research Agenda*. Eurographics, November 2024.

[3] Steven C.H. Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

[4] Mark Raasveldt and Hannes Mühleisen. Duckdb: an embeddable analytical database. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, page 1981–1984, New York, NY, USA, 2019. Association for Computing Machinery.