M2 Internship: Developing a "Big Data Model"

Yanlei Diao (École Polytechnique)

Duration: 5-6 months, the starting date is flexible (ideally March 1st, 2016)
Location: École Polytechnique, Palaiseau, France
Keywords: Big Data Analytics, **Databases**

Background: Big Data Analytics

Today data is being generated at an unprecedented rate, so much that 90% of the data in the world has been created in the past two years. Such significant increase of data volume is due to the new ways that we gather data: for example, from software tools that record system and user activities; from sensors and scientific instruments that monitor our built and natural environment; from medical instruments that enable genomic diagnosis of patients; and from human-initiated data sources such as blogs and social networks. However, raw data is of limited use by itself. The value of data is realized only when it is transformed into meaningful and actionable information, such as severe weather events, system anomalies, or user behavior patterns. This area is called "big data" today, which carries the promise to enable major leaps forward in science, business, and society at large in the next few decades.

Challenge: Building a "Big Data Model"

Existing research on big data has focused on BIG VOLUME, which has spawn research and implementation on highly scalable, fault-tolerant data processing, and BIG ANALYTICS, which has led to extensive work on statistical analysis, machine learning, and matrix/graph operations in cluster computing. This project aims to augment big data analytics with the ability to cope with BIG DIVERSITY, that is, the ability to deal with increasingly diverse datasets. A well-known example is integrating a large number of data sources with diverse data types such as structured relational databases and semi-structured RDF datasets. Furthermore, recent applications reveal that even the same dataset may need to be considered under different models/views in complex dataflow programs.

As a simple example, consider a PigLatin [1] program on web page click streams with the following steps: (1) find the top-k visited URLs, (2) find the users that visited at least 20% of the top-k URLs—let us call them the users of "good taste", (3) for only those users of "good taste", find out how their favorite top-k URLs evolved in the past month. The program, when directly implemented on Hadoop [2], requires many passes of data shuffling, with the first pass for grouping data by the URL, the second pass for grouping data by the user ID, the third pass for filtering clicks based on user ID and then grouping the remaining data based on URL again. Such repeated data re-arrangements, a key reason for high network and I/O overheads, stem from different "views" of data in those processing steps regarding how data should be grouped or ordered. Examples of data analytics programs (pipelines) that take different views of the underlying datasets abound, ranging from business analytics to genomic data analysis. Existing research considers different data models and views in isolation. Hence, when they are used together in a single analytics pipeline, tremendous overheads are incurred repeatedly to transform data into the right model or view for the next processing step. New techniques are needed to support such simultaneous views on the underlying datasets and efficient analytics based on these views.

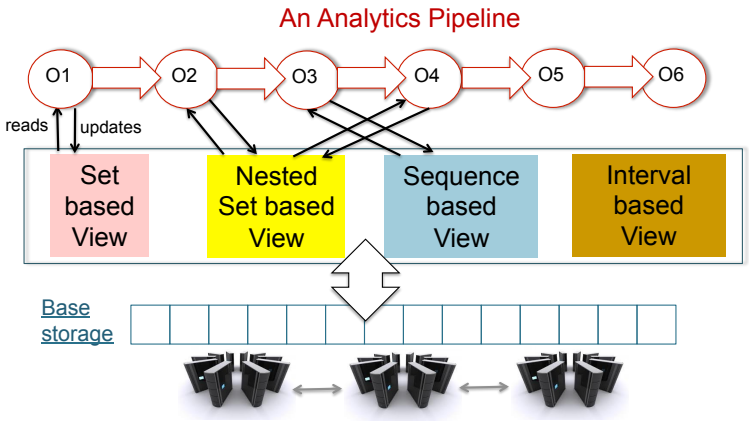

Figure 1: A "Big Data Model" with multiple views for an analytics pipeline.

Internship Goal:

The goal of this internship is to design and implement a new data analytics system that supports a "Big Data Model", which provides different ways to look at data, e.g., as flat tables, nested sets, or sorted sequences, as shown in Fig. 1. The system maintains the base storage of data in a small number of copies, each optionally with a different ordering or grouping property, and multiple concurrent data processing systems on top of it, each of which supports a distinct model/view of the data needed in an analytics pipeline. The fundamental concept explored in this project is a "materialized view" [4, 3], which is a stored computation result from an earlier processing step in a pipeline, such as the computed frequency of each URL, and which can expedite subsequent processing,

Based on our knowledge of common workloads, we aim to design a system that can automatically determine: (1) which form of materialization is needed in each data processing system based on a distinct view, e.g., taking the base storage as is, sorting the base storage by a new criterion, or building a new index on the base storage; (2) as new data is being generated in a recent processing step, how the updates should be propagated back to the base storage, as well as to other data processing systems; (3) what is the most cost-effective way to perform the above tasks.

This project will involve the design of new data structures and algorithms based on sorting, hashing, or indexing, development of new theory on view updates, and implementation using a real storage system. The proposed techniques will be evaluated using real-world analytical workloads such as business analytics and genomic data analysis.

Contact

- Yanlei Diao (yanlei.diao@polytechnique.edu), http://www.lix.polytechnique.fr/~yanlei.diao/

References

[1] Alan Gates, Olga Natkovich, Shubham Chopra, Pradeep Kamath, Shravan Narayanam, Christopher Olston, Benjamin Reed, Santhosh Srinivasan, and Utkarsh Srivastava. Building a highlevel dataflow system on top of mapreduce: The pig experience. *PVLDB*, 2(2):1414–1425, 2009. 1

[2] Hadoop: Open-source implementation of mapreduce. http://hadoop.apache.org. 1

[3] Alon Y. Halevy. Answering queries using views: A survey. *VLDB J.*, 10(4):270–294, 2001. 2

[4] Raghu Ramakrishnan and Johannes Gehrke. *Database management systems (3. ed.)*. McGraw-Hill, 2003. 2