ExpoDB: An Exploratory Data Science Platform

Mohammad Sadoghi

Computer Science Department
Purdue University
msadoghi@cs.purdue.edu

ABSTRACT

The success of relational databases is due in part to the simplicity of the tabular data, the clear separation of the physical and logical view of data, and the simple representation of the logical view as a flat schema (meta-data). But we are now witnessing a paradigm shift owing to the explosion of data volume, variety, and veracity, and as a result, there is a real need to knit together a data model that is naturally heterogeneous, but deeply interconnected. To be useful in this world, we argue that today's tabular data model must evolve into a *unified data model* that views meta-data as a new semantically rich source of data and unifies data and meta-data such that the data becomes *descriptive*. Furthermore, given the dynamicity of data, we argue that fundamental changes are needed in how data is consolidated continuously under uncertainty to make the data model naturally more *adaptive*. We further envision that the entire query model must evolve into a context-aware model in order to automatically refine, discover, and correlate data across many independent sources in real-time within the context of each query. We argue that enriching data with semantics and exploiting the context of the query are the two key prerequisites for realizing our vision of ExpoDB: *an exploratory data science platform*.

1. INTRODUCTION

We observe that today's data is no longer limited to systems of records; we now have a variety of data coming from thousands of sources. Data is being generated at an astonishing rate of 2.5 billion gigabytes daily, and further, 80% of data is unstructured and comes in the variety of forms [1]. These emerging data sources are heterogeneous by nature and are independently produced and maintained, yet the data are inherently related. Leaving data trapped in disconnected islands of information forces analytics-driven decision making to be carried out in isolation and on stale (and possibility irrelevant) data. More importantly, existing database technologies fail to alleviate the data exploration challenges that continue to be a daunting process especially at a time when an army of data scientists are forced to manually and continuously refine their analyses as they sift through these islands of disconnected data sources, a labor-intensive task occupying 50-80% of time spent [3].

We argue that today's database systems need to be fundamentally re-designed to capture data heterogeneity and the semantic relationship among data instances (within and across data sources) as first class-citizens. To address these requirements, we propose a *multi-layer data model* to capture all dimensions of the data[1], so that we can push the burden of semantic enrichment and fusion of the data in a systematic and transparent way into the database engines [6].

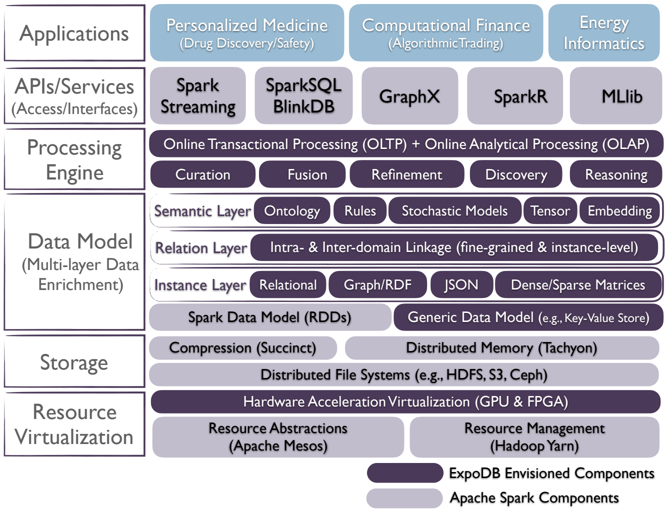
Applications
Personalized Medicine (Drug Discovery/Safety)
Computational Finance (AlgorithmicTrading)
Energy Informatics
APIs/Services (Access/Interfaces)
Spark Streaming
SparkSQL BlinkDB
GraphX
SparkR
MLlib
Processing Engine
Online Transactional Processing (OLTP) + Online Analytical Processing (OLAP)
Curation
Fusion
Refinement
Discovery
Reasoning
Data Model (Multi-layer Data Enrichment)
Semantic Layer
Ontology
Rules
Stochastic Models
Tensor
Embedding
Relation Layer
Intra- & Inter-domain Linkage (fine-grained & instance-level)
Instance Layer
Relational
Graph/RDF
JSON
Dense/Sparse Matrices
Spark Data Model (RDDs)
Generic Data Model (e.g., Key-Value Store)
Storage
Compression (Succinct)
Distributed Memory (Tachyon)
Distributed File Systems (e.g., HDFS, S3, Ceph)
Resource Virtualization
Hardware Acceleration Virtualization (GPU & FPGA)
Resource Abstractions (Apache Mesos)
Resource Management (Hadoop Yarn)
ExpoDB Envisioned Components
Apache Spark Components

Figure 1: ExpoDB Architecture over Apache Spark

We view the data enrichment as a *gradual process* that transforms the raw data into a new unified entity that has *knowledge-like characteristics*. Building upon the enriched data model, we envision that querying and analytics in general will become explorative in nature to provide deeper and quicker insights by proactively refining and raising new queries based on the context of the query submitted by the user. Thus, queries can be answered by an online consolidation (by exploiting the available semantics) of the most up-to-date data from a variety of sources at query time without the need for offline fusion and curation [6].

As a result, we envision the evolution of today's databases in order to meet the continuous exploration and fusion challenges of the information explosion at Web scale. To this end, in ExpoDB, we reimagine the Apache Spark architecture (that is currently knowledge oblivious) to be transformed into a knowledge exploration platform. Other important aspects of ExpoDB is a unified approach to combine both OLTP and OLAP processing [5]; introducing (virtualized) hardware acceleration at every stage of query processing [4] by making the network, storage, and memory active; and finally giving rise to many prominent applications for data scientists such as personalized medicine for improving drug safety [2]. The high-level architecture of ExpoDB is demonstrated in 1.

2. REFERENCES

[1] The IBM strategy. Annual Report'13. http://www.ibm.com/annualreport/2013/, 2013.

[2] A. Fokoue, M. Sadoghi, O. Hassanzadeh, and P. Zhang. Predicting drug-drug interactions through large-scale similarity-based link prediction. In *ESWC'16*.

[3] S. Lohr. For big-data scientists, "janitor work" is key hurdle to insights. The New York Times, 2014.

[4] M. Najafi, M. Sadoghi, and H. Jacobsen. The FQP vision: Flexible query processing on a reconfigurable computing fabric. *SIGMOD Record'15*.

[5] M. Sadoghi, S. Bhattacherjee, B. Bhattacharjee, and M. Canim. L-Store: A real-time OLTP and OLAP system. *arXiv'16*.

[6] M. Sadoghi, K. Srinivas, O. Hassanzadeh, Y. Chang, M. Canim, A. Fokoue, and Y. A. Feldman. Self-curating databases. In *EDBT'16*.

[1] Partially enabled by the recent semi-supervised machine learning techniques (e.g., tensor factorization) to capture semantic relationships within data.