ExpoDB: An Exploratory Data Science Platform

Mohammad Sadoghi
Computer Science Department
Purdue University
msadoghi@purdue.edu

ABSTRACT

We are entering a new data era: on the one hand, we are witnessing an unprecedented explosion of data volume and variety; and on the other hand, the data is now becoming increasingly interconnected yet disconnected. To derive insights from data, there is a pressing need to knit together a data model that is naturally heterogeneous while deeply interconnected. To construct a *unified view of data*, we must overcome the fundamental mismatch in the way data and its meta-data are expressed in each source by making data model inherently *descriptive*—a step towards developing a unified data representation to serve as a common ground for fusing and exploring data. We argue that we must fundamentally revisit how we cope with the dynamicity of data in order to offer a continuous consolidation of data under uncertainty to make the data model inherently *adaptive*. We envision evolution of today's query model into a context-aware paradigm such that within the context of every query, we automatically refine, discover, and correlate data across many independently maintained data sources in real-time. We argue that developing *unified data model* along with *context-aware query model* are two major steps in realizing our vision of ExpoDB: *an exploratory data science platform* to ultimately achieve a real-time exploration and fusion of enriched data at Web scale.

1. INTRODUCTION

Today's data has surpassed the traditional systems of records found in the enterprise. Data is being generated from increasing number of sources at an astonishing rate of 2.5 billion gigabytes daily [8]. Yet data is trapped in islands of disconnected and heterogeneous sources despite being inherently interconnected. This disconnectedness limits the insight-driven decision making to stale and possibility irrelevant data. Moreover, database systems fail to alleviate these data exploration challenges that is forcing an army of data scientists to manually and continuously refine their analyses as they sift through these islands of disconnected sources that accounts for 50-80% of the overall time spent [2].

Our long-term goal is to redesign today's database systems to systematically push the burden of semantic enrichment, fusion, exploration of the data into the database engine [8]. To capture data heterogeneity and the semantic relationship of data as first class-citizens (within and across sources), we propose a *multi-layered data model*. Essentially, viewing the data enrichment as a *gradual process*, where the raw data is transformed into a new unified representation having *knowledge-like characteristics* [8]. The unified and enriched data model is the prerequisite to make analytics explorative, so that queries can be answered by an online consolidation of the most up-to-date data from a variety of sources at query time without the need for offline fusion and curation [8].

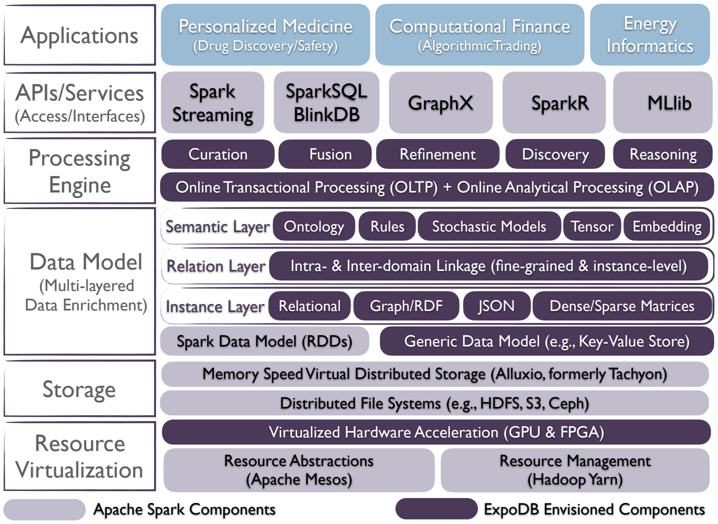

Figure 1: ExpoDB Architecture over Apache Spark

As a first step towards realizing our vision, in ExpoDB, we reimagine the Apache Spark architecture (that is currently knowledge oblivious) to evolve into a knowledge exploration platform (cf. Figure 1). Other notable aspects of ExpoDB is a unified approach to combine both OLTP and OLAP processing [7, 6, 9, 3]; introducing (virtualized) hardware acceleration at every stage of query processing [4, 5] 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 [1].

2. REFERENCES

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

[2] S. Lohr. For big-data scientists, "janitor work" is key hurdle to insights. New York Times'14.

[3] M. Hemmatpour, B. Montrucchio, M. Rebaudengo, and M. Sadoghi. Kanzi: A distributed, in-memory key-value store. In *Middleware'16*.

[4] M. Najafi, M. Sadoghi, and H. Jacobsen. Flexible query processor on FPGAs. *PVLDB*, 6(12):1310–1313, 2013.

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

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

[7] M. Sadoghi, M. Canim, B. Bhattacharjee, F. Nagel, and K. A. Ross. Reducing database locking contention through multi-version concurrency. *PVLDB'14*.

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

[9] K. Zhang, M. Sadoghi, and H. Jacobsen. DL-Store: A distributed hybrid OLTP and OLAP data processing engine. In *ICDCS'16*.

8th Biennial Conference on Innovative Data Systems Research (CIDR '17)
January 8-11, 2017 , Chaminade, California, USA.