Big Data Integration

Xin Luna Dong
Google, Inc.
lunadong@google.com

Divesh Srivastava
AT&T Labs–Research
divesh@research.att.com

ABSTRACT

The Big Data era is upon us: data is being generated, collected and analyzed at an unprecedented scale, and data-driven decision making is sweeping through society. Since the value of data explodes when it can be linked and fused with other data, addressing the big data integration (BDI) challenge is critical to realizing the promise of Big Data.

BDI differs from traditional data integration in many dimensions: (i) the number of data sources, even for a single domain, has grown to be in the tens of thousands, (ii) many of the data sources are very dynamic, as a huge amount of newly collected data are continuously made available, (iii) the data sources are extremely heterogeneous in their structure, with considerable variety even for substantially similar entities, and (iv) the data sources are of widely differing qualities, with significant differences in the coverage, accuracy and timeliness of data provided. This tutorial explores the progress that has been made by the data integration community on the topics of schema mapping, record linkage and data fusion in addressing these novel challenges faced by big data integration, and identifies a range of open problems for the community.

1. INTRODUCTION

The Big Data era is the inevitable consequence of our ability to generate and collect digital data at an unprecedented scale, and our concomitant desire to analyze and extract value from this data in making data-driven decisions to alter all aspects of society.

This data is being collected today in a large variety of domains. Examples include Web text and documents, Web logs, large-scale e-commerce, social networks, sensor networks, astronomy, genomics, medical records, surveillance, etc.[1] Since the value of data explodes when it can be linked and fused with other data to create a unified representation, big data integration (BDI) is critical to realizing the promise of Big Data. As one example, recent efforts in mining the Web and building knowledge bases show promise of using integrated big data to improve Web search and Web-scale data analysis.

BDI differs from traditional data integration (which includes virtual integration and materialized warehousing) in many dimensions.

- *Volume:* Not only can each data source contain a huge volume of data, but also the number of data sources has grown to be in the millions. Even for a single domain, this number is now in the tens of thousands. This is much higher than the number of data sources considered in traditional data integration.
- *Velocity:* As a direct consequence of the rate at which data is being collected and continuously made available, many of the data sources are very dynamic, and the number of data sources is also exploding.
- *Variety:* Data sources (even in the same domain) are extremely heterogeneous both at the schema level regarding how they structure their data and at the instance level regarding how they describe the same real-world entity, exhibiting considerable variety even for substantially similar entities.
- *Veracity:* Data sources (even in the same domain) are of widely differing qualities, with significant differences in the coverage, accuracy and timeliness of data provided. This is consistent with the observation that "1 in 3 business leaders do not trust the information they use to make decisions." [2]

This tutorial explores the progress that has been made by the data integration community on the topics of *schema mapping*, *record linkage* and *data fusion* in addressing these novel challenges faced by BDI. We do this using illustrative examples that would be of interest to researchers and practitioners. We also identify a range of open problems for the community in integrating a galaxy of data sources.

2. TUTORIAL OUTLINE

The importance of big data integration has led to a substantial amount of research over the past few years on the topics of schema mapping, record linkage and data fusion to deal with the novel challenges faced by big data integration. Table 1 shows a summary of these techniques. Our tutorial is example driven, and organized as follows.

[1] http://en.wikipedia.org/wiki/Big_data

[2] http://www-01.ibm.com/software/data/bigdata/

Proceedings of the VLDB Endowment, Vol. 6, No. 11

Table 1: Summary of state-of-the-art data integration techniques meeting challenges of big data.

	Schema mapping	Record linkage	Data fusion
Volume	Integrating Deep Web, Web tables/lists	Adaptive blocking, MapReduce-based linkage	Online fusion
Velocity		Incremental linkage	Fusion in a dynamic world
Variety	Dataspace systems	Linking text to structured data	Combining fusion with linkage
Veracity		Value-variety tolerant linkage	Truth discovery

2.1 BDI: Motivation

The tutorial will start with a variety of real-world examples illustrating the importance of big data integration.

2.2 BDI: Schema Mapping

Schema mapping in a data integration system refers to (i) creating a mediated (global) schema, and (ii) identifying the mappings between the mediated (global) schema and the local schemas of the data sources to determine which (sets of) attributes contain the same information.

Early efforts in integrating a large *number* of sources involved integrating data from the Deep Web. Two types of solutions were proposed. The first is to build mappings between Web forms (interfaces to query the Deep Web) as a means to answer a Web query over all Deep Web sources. The second is to crawl and index the Deep Web data. More recent efforts include extracting and integrating structured data from Web tables and Web lists.

The number of sources also increases the *variety* of the data. Traditional data integration systems require a significant schema mapping effort before the system can be used, so is obviously infeasible when the heterogeneity is at the BDI scale. The basic idea of *dataspace systems* is to provide best-effort services such as simple keyword search over the available data sources at the beginning, and gradually evolve schema mappings and improve search quality over time.

2.3 BDI: Record Linkage

Record linkage refers to the task of identifying records that refer to the same logical entity across different data sources, especially when they do not share a common identifier across the data sources. It has traditionally focused on linking a static set of structured records that have the same schema. In BDI, (i) data sources tend to be heterogeneous in their structure and many sources (e.g., tweets, blog posts) provide unstructured text data, and (ii) data sources are dynamic and continuously evolving. These characteristics make record linkage particularly challenging in BDI.

When there are a large number of sources and a large volume of data, traditional record linkage approaches become inefficient and ineffective in practice. To address the *volume* dimension, new techniques have been proposed to enable parallel record linkage using MapReduce. These include techniques for adaptive blocking and techniques that balance load among different nodes.

When data sources are dynamic and continuously evolving, applying record linkage from scratch for each update becomes unaffordable. To address the *velocity* aspect, incremental clustering techniques have been proposed.

Record linkage between structured and unstructured data sources arises, e.g., when linking shopping transactions of people with tweets or blog posts about their shopping experience. Highly heterogeneous information spaces (e.g., the Web and dataspaces) also demand new record linkage techniques. To address the *variety* aspect, techniques have been proposed that tag and match free text to structured data.

Finally, in the BDI environment, information is typically more imprecise and noisy. To address this *veracity* aspect, a variety of clustering and linkage techniques that are robust to noise or evolving values have been proposed.

2.4 BDI: Data Fusion

Data fusion refers to resolving conflicts from a collection of sources and finding the truth that reflects the real world. Unlike schema mapping and record linkage, data fusion is a new field that has emerged only recently. Its motivation is exactly the *veracity* of data: the Web has made it easy to publish and spread false information across multiple sources and so it is critical to separate the wheat from the chaff for presenting high quality data.

To address such *veracity* related challenges, techniques have been proposed to find the single or multiple truths from conflicting values. Such techniques have been extended to handle the *volume* of data (online data fusion), *velocity* of data (truth discovery for dynamic data), and *variety* of data (combining record linkage and data fusion).

2.5 BDI: Architecture

After we discuss each component of big data integration, we examine whether the main-stream integration architectures suit the current trends in Big Data. First, recent work discusses the pros and cons for integrating data from *all* available sources and proposes how to select a subset of data sources to balance the gain (obtaining a high quality of integrated data) and cost (data purchase cost and integration resources) of data integration. Second, we compare the offline warehousing and data aggregation architecture, which builds a central repository for structured data from various data sources and provides search on this repository, and the online data integration architecture, which reformulates and sends a user query to a multitude of underlying sources and returns a union of the answers retrieved from those sources. We suggest possible improvements for big data integration.

2.6 BDI: Open Problems

Finally, we discuss cutting-edge open problems for big data integration, such as integrating crowd sourcing data, integrating data from data markets, providing an exploration tool for data sources, and so on.

3. CONCLUSIONS

This tutorial reviews state-of-the-art techniques for data integration in addressing the challenges raised by Big Data: *volume* and number of sources, *velocity, variety,* and *veracity.* We discuss how close we are to meeting these challenges and identify many open problems for future research.