Big Data Ecosystem

Reference Architecture

By Orit Levin, Microsoft Corporation.

July 1, 2013

# Introduction

In recent years, the ongoing innovation in distributed computing and cloud infrastructure has become one of the main factors for data collection growth, development of new database technologies and analytics, and their broadening usage. From the enterprise data warehouses to the hundreds of companies that comprise the advertising industry, advancements in a big data ecosystem are quickly incorporated in best practices and newest technologies, becoming more cost-effective and fuelling innovation. Policy makers worldwide, traditionally focused on data collection, are broadening their scope of concern to include new data distribution and data usage practices.

The big data ecosystem can be best described using a data-centric reference architecture (RA) showing end-to-end data collection, transformation, distribution, and usage. Many players in the big data ecosystem would benefit from this analysis, particularly managers and policy makers dealing with the rapid changes in the way data is collected and transformed. The purpose of this RA is to assist the NIST Big Data standardization effort through (1) making the life cycle of big data better understood by industry, policy makers, and users; (2) identifying relevant components and functions in order to define their boundaries, interoperability surfaces, and security implications.

# Overview

The big data ecosystem reference architecture is a high level data-centric diagram that depicts the big data flow and possible data transformations from collection to usage.

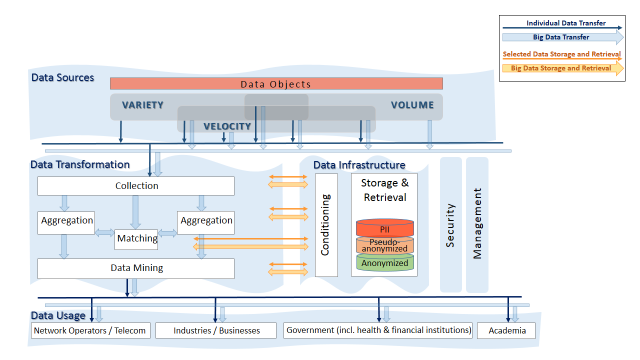
The big data ecosystem is comprised of four main components: Sources, Transformation, Infrastructure and Usage, as shown on Figure 1. Security and Management are shown as examples of additional supporting cross-cutting sub-systems that provide backdrop services and functionality to the rest of the big data ecosystem. 

Figure 1: Big Data Ecosystem Reference Architecture

# Data Sources

Typically, the data behind “big data” is collected for a specific purpose, creating the data objects in a form that supports the known use at the data collection time. Once data is collected, it can be reused for a variety of purposes, some potentially unknown at the collection time.

Data sources can be classified by three characteristics that define big data and are independent of the data content or context: Volume, Velocity, and Variety[[1]](#footnote-1):

Volume: Caused by increased transaction numbers, as well as by new types of data. It can be both a storage issue and a massive analysis issue.

Variety: It is characterized by different formats of information including tabular data (databases), hierarchical data, documents, e-mail, metering data, video, still images, audio, stock ticker data, and financial transactions.

Velocity: This involves streams of data, structured record creation, and availability for access and delivery. Velocity means both how fast data is being produced and how fast the data must be processed to meet demand.

Typically, data that can be characterized by an extreme of at least two out of these characteristics is considered “big data” that would be very difficult to process, use, and manage using traditional databases and data-processing applications.

For each use case, sources of big data can be further classified by various application-specific criteria, such as the presence of information about people (that may have downstream privacy implications), the format, or specific data sources.

# Data Transformation

As data propagates through the ecosystem, it is being processed and transformed in different ways in order to extract the value from the information. For the purpose of defining interoperability surfaces, it is important to identify common transformations that are implemented by independent modules, systems, or deployed as stand-alone services.

*Editor’s Note:* The transformation functional blocks shown in Figure 1 can be performed by separate systems or organizations, with data moving between those entities. (See Use Case I: Advertising.) Similar and additional transformational blocks are being used in enterprise data warehouses, but typically they are closely integrated and rely on a common data base to exchange the information. (See Use Case II: Enterprise Data Warehouse.)

Each transformation function may have its specific pre-processing stage including registration and metadata creation, may use different specialized data infrastructure best fitted for its requirements, and may have its own privacy and other policy considerations.

Common known data transformation include at least:

Collection: Data can be collected in different types and forms. At the initial collection stage, sets of data (e.g., data records) from similar sources and of similar structure are collected (and combined) resulting in uniform security considerations, policies, etc. Initial metadata is created (e.g., subjects with keys are identified) to facilitate subsequent aggregation or lookup method(s).

Aggregation: Sets of existing data collections with easily correlated metadata (e.g., identical keys) are aggregated into a larger collection. As a result, the information about each object is enriched or the number of objects in the collection grows. Security considerations and policies concerning the resultant collection are typically similar to the original collections.

Matching: Sets of existing data collections with dissimilar metadata (e.g., keys) are aggregated into a larger collection. (For example, in advertising industry matching services correlate HTTP cookies’ values with person’s real name.) As a result, the information about each object is enriched. The security considerations and policies concerning the resultant collection are subject to data exchange interfaces design.

*Editor’s Note:* With the development of Use Cases, more common transformational blocks to be added.

Data Mining: According to DBTA[[2]](#footnote-2), “[d]ata mining can be defined as the process of extracting data, analyzing it from many dimensions or perspectives, then producing a summary of the information in a useful form that identifies relationships within the data. There are two types of data mining: descriptive, which gives information about existing data; and predictive, which makes forecasts based on the data.”

# Data Infrastructure

Big data infrastructure is a bundle of data storage or database software, servers, storage, and networking used in support of the data transformation functions and for storage of data as needed.

*Editor’s Note*: In Figure 1, Data Infrastructure is to the right of the Data Transformation, to emphasize the natural role of Data Infrastructure in support of data transformations. Note that the horizontal data retrieval and storage paths exists between the two, which are different from the vertical data paths between them and Data Sources and Data Usage.

In order to achieve high efficiencies, data of different volume, variety and velocity would typically be stored and processed using computing and storage technologies tailored to those characteristics. The choice of processing and storage technology is also dependent on the transformation itself. As a result, often the same data can be transformed (either sequentially or in parallel) multiple times using independent data infrastructure.

[Under development] Conditioning: Examples include de-identification, sampling, and fuzzing.

[Under development] Storage and Retrieval. Examples include NoSQL and SQL Databases with various specialized types of data load and queries.

# Data Usage

[Under development] The results can be provided in different formats, different granularity and under different security considerations.

# Use Case I: Advertising

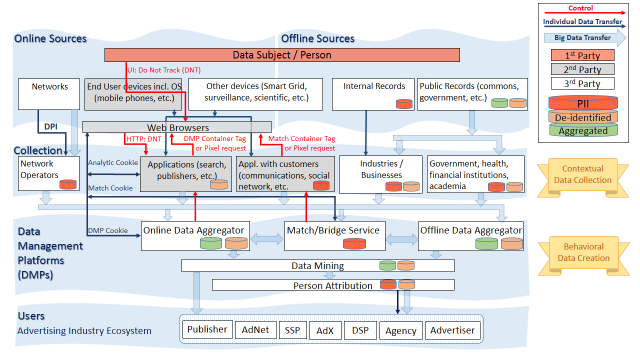
The purpose of this use case is to illustrate the Big Data RA using the well-known advertising industry ecosystem[[3]](#footnote-3) as an example. 

Figure 2: Use Case: Advertising

# Use Case II: Enterprise Data Warehouse

The purpose of this use case is to illustrate the Big Data RA using an abstract enterprise data warehouse decomposition[[4]](#footnote-4).

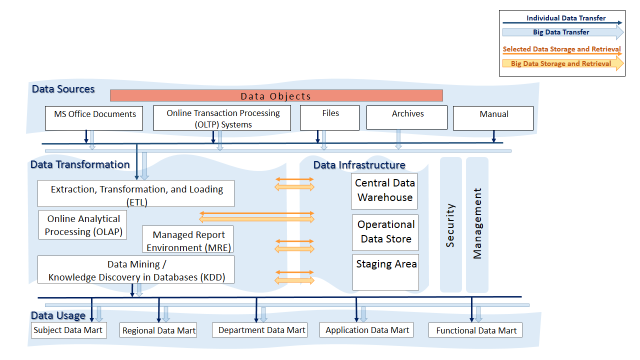


Figure 3: Use Case: Enterprise Data Warehouse

1. Gartner Press Release, “[Gartner Says Solving ‘Big Data’ Challenge Involves More Than Just Managing Volumes of Data](http://www.gartner.com/it/page.jsp?id=1731916)”, June 27, 2011. [↑](#footnote-ref-1)
2. DataBase Trends and Applications, <http://www.dbta.com/Articles/Editorial/Trends-and-Applications/What-is-Data-Analysis-and-Data-Mining-73503.aspx>, Jan 7, 2011 [↑](#footnote-ref-2)
3. [www.iab.net/data/ecosystem.html](http://www.iab.net/data/ecosystem.html) [↑](#footnote-ref-3)
4. The terminology and the breakdown into subsystems is based on <http://www.dbta.com/Articles/Editorial/Trends-and-Applications/What-is-Data-Analysis-and-Data-Mining-73503.aspx>, <http://www.breteler.com/DWH.htm>, and <https://www.ibm.com/developerworks/community/blogs/Sabir/tags/architectures?lang=en>. [↑](#footnote-ref-4)