**NIST Big Data Working Group (NBD-WD)**

**NBD-WD-2013/M0057**

|  |  |
| --- | --- |
| **Source:** | Reference Architecture Subgroup |
| **Status:** | Draft |
| **Title:** | Terminology used to describe Reference Architectures as submitted to the Reference Architectures Subgroup |
| **Date:** | July 24, 2013 |
| **Author(s):** | Orit Levin (Microsoft), James Ketner (AT&T), Don Krapohl (Augmented Intelligence) |

Abstract

This list of terminology and definitions is compiled from the submissions (i.e., Input Documents) and the email threads of the NIST Big Data Reference Architecture Subgroup. The purpose of this document is to serve as an input to the Def&Tax Subgroup with the goal to assist its work.

# Source: Bob Marcus, Input Document M0014

* Data Sources
* Processing
  + Stream Processing
  + Data Processing
* Foundational Data Store
* Databases
  + In memory
  + Operational
  + Analytics
* Analytics and Interfaces
  + Real-Time
  + Interactive
  + Batch
* Applications and User Interfaces
* Infrastructure (Cloud?)
* Management
  + System Management
  + Data Resource Management
  + Process Management
* Security
* Design, Develop and Deploy tools

# Source: Orit Levin, Input Document M0015 and M0017

* **Data Sources**[[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.
* **Data Transformation**
  + Collection: 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.
  + 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**: a bundle of servers, storage, and networking (e.g., a datacenter) being used in support of data transformation functions and for storage of data as needed. In order to achieve high efficiencies, data of different volume, variety and velocity would typically be stored and processed using different, often tailored to specific tasks, storage and computing technologies. As a result, often the same data can be processed (either sequentially or in parallel) multiple times using independent data infrastructure solutions.
  + Conditioning (Examples include de-identification, sampling, and fuzzing)
  + Storage & Retrieval (Examples include NoSQL (e.g., Hadoop) and SQL Databases with various specialized types of data load and queries)
* **Data Usage**: The results can be provided in different formats, different granularity and under different security considerations.

# Source: Tim Zimmerlin, “[BigdataReq] Note Concerning Today's NBD-WG Requirements Subgroup Mtg”, 7/16/2013

**Design Concerns**

Description:

* are from agile and factoring methodologies
* verbalized high level drivers for high level requirements
* motivate and steer high level "processing steps"

Examples:

* "data discovery"
* "data analysis" (appears outside the scope of the NBD-WG. It is not generalizable.)
* "data findability"
* "data presentation"

**Processing Steps**

Description:

* maps inputs to outputs
* produce "staged data"
* all steps listed below (except for the “black box”) are controlled vocabulary terms that bridge computer and networks, low level IO and high level overlay networks, automation and human curation, legal and business responsibilities

Examples:

* "data ingestion"
* “non-generalizable” / “black box” processing step (typically unique to one big data system)
* "data extraction" (sometimes has to cough up incredible volumes of data; sometimes has hard time limits)
* “data presentation” (very important to end users and stakeholders, many current big data systems fumble data presentation)

**Data Stages**

Description:

* go back to states of mainframes and databases with transactions plus batch processing systems, including temporarily halting updates for snapshots for batch jobs

Examples:

* externally sourced data
* cleansed data set(s), extracted data set(s), data ready for processing
* "data in updated databases and data stores" initially staged data
* "data in updated databases and data stores" post-processing data
* formatted presentation materials (data accessible in a variety of media like email message, RSS feed, Web page, SQL cursor based client API, FTP file, etc.)

# Source: Gary Mazzaferro, presented on the RefArch call, 7/17/2013

* Visualization Devices
* Data Visualization Applications
* Data Analytics Infrastructure
* Data Processing Infrastructure
* Data Storage Infrastructure
* Hardware and Communications Networking Infrastructure

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)