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

**NBD-WD-2013/M0198**

|  |  |
| --- | --- |
| **Source:** | Reference Architecture Subgroup |
| **Status:** | A proposal for discussion |
| **Title:** | Content towards “NIST Big Data Reference Architecture” |
| **Date:** | August 28, 2013 |
| **Author(s):** | Orit Levin (Microsoft) |

***Abstract***

*This document contains an updated Reference Architecture diagram and a new text for discussion by the Reference Architecture subgroup with the purpose of including them into the subgroup deliverable.*

# Introduction

The “big data” ecosystem is an evolution and a superset of a “traditional data” system exhibiting any or all of the following characteristics or requirements:

* Data sources are diverse in their security and privacy considerations and their business relationships with the data system integrators
* Data imported into the system vary in its structure, exhibit large volume, velocity, variety, and other complex properties
* The nature and the order of data transformations varies between vertical systems; it is not prearranged, and is evolving for a given system
* Storage technologies and databases are tailored to specific transformation needs and their scalability properties scale horizontally, vertically, or both
* Innovative analytic functions continuously emerge; proven technologies get enhanced and abstracted, resulting in frequent updates and outsourcing practices
* Data usage vary in structure and format; new use cases can be easily introduced to the system

The resultant generic “big data” system is represented by the “Big Data Ecosystem Reference Architecture”, which

* Is comprised of functional components that are loosely coupled, can be implemented by different stakeholders, and communicate using interoperability surfaces
* Demonstrates the data security and privacy considerations that are essential to the big data system design from data collection to data usage.

# Conceptual Model

The big data ecosystem reference architecture (RA) is a technology agnostic diagram comprised of five main components: “Sources”, “Transformation”, “Capabilities”, “Usage”, and “Vertical Orchestration”, as shown on Figure 1. The two additional “Security & Privacy” and “System and Lifecycle Management” components are shown as cross-cutting sub-systems that provide services and functionality to the ecosystem components in the areas specific to “big data”.

The “data” arrows show the flow of data between the system’s main components. The data flows between the components either physically (a.k.a., by value) or by providing its location and the means to access it (a.k.a., by reference). The “control” arrows show requests directed to components to either perform operations or to comply with specified policies.

The RA is applicable to big data stakeholders regardless the actual business relationships between the entities implementing its components or subcomponents. The RA is organized around two axes representing the two “big data” views corresponding to its value chains: the information flow (along the vertical axis) and the IT integration (along the horizontal axis).

Along the information flow axis, the value is created by data collection, integration, analysis, and applying the results down the value chain. Along the IT axis, the value is created by providing IT services for hosting, operating and managing of big data for implementing the data transformations that are the building blocks for a generic vertical solution. Note that the transformation component is at the intersection of both axes indicating that data analytics and its implementation are of special value to big data stakeholders in both value chains.



Figure 1: Big Data Reference Architecture

# Main Components

## Data Sources

Data sources are internal and public records, images/audio and videos, sensor data, Web log, HTTP cookies, etc.

Data sources are produced by humans, machines, sensors, Internet technologies, etc.

Different data sources have different security and privacy considerations.

One of the important characteristics of a big data system is the ability to import and use data from a variety of resources.

## Transformation

Represents a generic “vertical system” data life cycle including data collection from various sources, multiple data transformations being implemented using both traditional and new technologies, and diverse data usage.

As the data propagates through the ecosystem, it is being processed and transformed in different ways in order to extract the value from the information. Transformation sub-components can be implemented by independent stakeholders and deployed as stand-alone services.

Each transformation function can use different specialized data infrastructure best fitted for its requirements, and can have its own privacy and other policy considerations.

## Capabilities

Capabilities are abstracted functionalities that exist in order to support big data transformation functions. Capabilities include infrastructures (e.g., VM clusters), platforms (e.g., databases), and applications (e.g., analytic tools).

## Data Usage

Includes data retrieval, reporting, rendering, etc. The exported data can be provided in different formats, different granularity and under different security considerations.

## Vertical Orchestration

Big data systems vary from a single-organization tightly-coupled solutions (integrated by standard or proprietary interfaces) to a loosely-coupled industry-wide ecosystem maintained by a variety of stakeholders bounded by agreements and standard or standard-de-facto interfaces.

The vertical orchestration component abstracts these two extreme cases and represents:

* the ability to integrate different capabilities using technological components and services into a robust scalable system
* the ability to work with an extended supply chain from a variety of data resources and business partners for extracting value from the state-of-the-art analytics.

# Interfaces

## Data Import

**Control-I:** Used by the data owners to specify policies for collection and usage of their data.

**Data-I**: Represents different big data characteristics (Volume, Velocity, Variety, etc.) and formats (records, streams, packets, etc.) of the data being imported into the system. Supports multiple sources of data.

## Capabilities

**Control-C:** Used by transformation functions to request specific functions or services from the supporting infrastructures, platforms, or applications.

**Data-C**: Represents data transfer to, from, and in-between cloud infrastructures, databases, or services often physically and geographically distinct as the data propagates through the transformation stages.

## Vertical Orchestration

**Control-O:** Represents a set of vertical-specific operations or a description of stakeholders’ practices and interfaces in a loosely-coupled vertical industry.

## Data Export

**Control-E:** Used by consumers to tailor the exported data format, granularity, etc. subject to system’s policies.

**Data-E**: Represents different exported data formats for data retrieval, reporting, and rendering. Supports multiple users of exported data.