**NIST Big Data**

**Reference Architecture**

**Version 1.1**

**Reference Architecture Subgroup**

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

**August, 2013**

[Executive Summary 4](#_Toc364669083)

[1 Introduction 4](#_Toc364669084)

[1.1 Objectives 4](#_Toc364669085)

[1.2 How This Report Was Produced 4](#_Toc364669086)

[1.3 Structure of This Report 4](#_Toc364669087)

[2 Big Data Reference Architecture: Use Cases Requirements 4](#_Toc364669088)

[3 Big Data Reference Architecture: Conceptual Model 6](#_Toc364669089)

[4 Big Data Reference Architecture: Main Components 7](#_Toc364669090)

[4.1 Data Sources 7](#_Toc364669091)

[4.2 Transformation 7](#_Toc364669092)

[4.2.1 Collection Services 8](#_Toc364669093)

[4.2.2 Curation Services 8](#_Toc364669094)

[4.2.3 Analytic Services 8](#_Toc364669095)

[4.2.4 Visualization Services 8](#_Toc364669096)

[4.2.5 Access Services 8](#_Toc364669097)

[4.3 Capability Management 8](#_Toc364669098)

[4.3.1 Big Data Framework 8](#_Toc364669099)

[4.3.2 Hardware Management 8](#_Toc364669100)

[4.3.3 Security and Privacy Management 8](#_Toc364669101)

[4.3.4 System Management 8](#_Toc364669102)

[4.3.5 Lifecycle Management 8](#_Toc364669103)

[4.4 Data Usage 8](#_Toc364669104)

[4.4.1 Data Reporting 8](#_Toc364669105)

[4.4.2 Data Retrieval 8](#_Toc364669106)

[4.4.3 Data Rendering 8](#_Toc364669107)

[5 Big Data Reference Architecture: Interfaces 8](#_Toc364669108)

[5.1 Data Service Abstraction 9](#_Toc364669109)

[5.1.1 Overview 9](#_Toc364669110)

[5.1.2 Example: Collection 9](#_Toc364669111)

[5.1.3 Example: Curation 9](#_Toc364669112)

[5.1.4 Example: Analytic 9](#_Toc364669113)

[5.2 Capability Service Abstraction 9](#_Toc364669114)

[5.2.1 Overview 9](#_Toc364669115)

[5.2.2 Example: Streaming 9](#_Toc364669116)

[5.2.3 Example: Aggregation 9](#_Toc364669117)

[5.2.4 Example: Integration 9](#_Toc364669118)

[5.2.5 Example: Transfer 9](#_Toc364669119)

[5.2.6 Example: Search 9](#_Toc364669120)

[5.2.7 Example: Statistics 9](#_Toc364669121)

[5.2.8 Example: RT Analytics 9](#_Toc364669122)

[5.2.9 Example: Batch Analytics 9](#_Toc364669123)

[5.2.10 Example: Interactive Animation 9](#_Toc364669124)

[5.3 Usage Service Abstraction 9](#_Toc364669125)

[5.3.1 Overview 9](#_Toc364669126)

[5.3.2 Example: Retrieval 9](#_Toc364669127)

[5.3.3 Example: Rendering 9](#_Toc364669128)

[5.3.4 Example: Report 9](#_Toc364669129)

[6 Big Data Reference Architecture: Security and Privacy 9](#_Toc364669130)

[7 Big Data Taxonomy 10](#_Toc364669131)

[Appendix A: Big Data Terms and Definitions 10](#_Toc364669132)

[Appendix B: Acronyms 10](#_Toc364669133)

[Appendix C: References 10](#_Toc364669134)

# Executive Summary

# Introduction

## Objectives

## How This Report Was Produced

## Structure of This Report

# Big Data Reference Architecture: Use Cases Requirements

**[This section will be prepared by the NIST BDWG Requirements SG and will contain high level requirements relevant to the design of the Reference Architecture. It is recommended that the section follows the Big Data RA decomposition. For example:**

**Data Source Requirements:**

DSR-1: needs to support reliable real time, streaming, and batch processing to collect data from centralized and distributed data sources, sensors, or instruments.

DSR-2: needs to support slow and high throughput data transmission between data sources and computing clusters.

DSR-3: needs to support diversified data content ranging from text to multimedia to instrumental data.

**Transformation Requirements:**

TR-1: needs to support diversified analytic processing and machines learning techniques

TR-2: needs to support batch and real time analytic processing

TR-3: needs to support processing diversified data content

TR-4: needs to support processing data in motion (streaming, fetching new content, tracking, etc.)

TR-5: needs to support legacy and advance programming executable and libraries

**Data Infrastructure Requirements (to enable Transformation processing):**

DIR-1: needs to support legacy and advance software packages (subcomponent: SaaS)

DIR-2: needs to support legacy and advance computing platforms (subcomponent: PaaS)

DIR-3: needs to support legacy and advance distributed computing cluster (subcomponent: IaaS)

DIR-4: needs to support elastic data transmission (subcomponent: networking)

DIR-5: needs to support legacy and advance distributed data storage (subcomponent: storage)

**Data Usage Requirements:**

DUR-1: needs to support fast search (~0.1 seconds) from processed data

DUR-2: needs to support diversified output file formats for rendering

DUR-3: needs to support visual layout for results presentation

DUR-4: needs to support rich user interface

DUR-5: needs to support streaming results to clients

**Security & Privacy Requirements:**

SnPR-1: needs to support security and privacy on protected data

SnPR-2: needs to support multi-level access control on protected data

**System Management Requirements:**

SMR-1: needs to support rich user interface from mobile platforms to access processed results

SMR-2: needs to support performance monitoring on analytic processing from mobile platforms

SMR-3: needs to support rich visual content rendering from mobile platforms

**Lifecycle Management Requirements:**

LMR-1: needs to support data quality curation

LMR-2: needs to support dynamic updates on data and user profiles

LMR-3: needs to support data lifecycle policy

LMR-4: needs to support data validation

LMR-5: needs to support human annotation for data validation]

[Editor’s Note: This section may be moved to Appendix.]

# Big Data Reference Architecture: Conceptual Model

[Key points to meet use cases’ requirements:

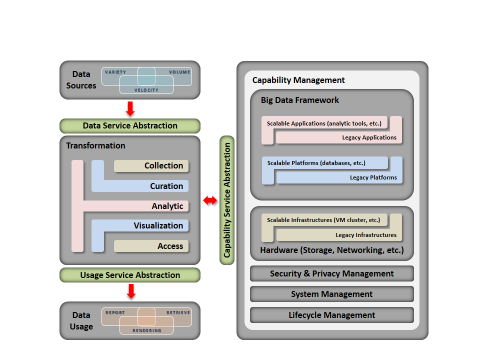
* 1. Be able to support diversified data sources with variety, volume, and velocity of data
  2. Be able to support various transformation functions such as collect, curate, analysis, visual, and access for a given dataset.
  3. Be able to support mixed of legacy and big data framework with the additional resource management of hardware, security and Privacy, systems, and lifecycle management.]

[NBD RA identifies abstract interfaces to meet the use cases’ requirements:

1. ***Data Service Abstraction*** – enables for one or more datasets to be collected, curated, analyzed, visualized, and accessed by the Transformation components.
2. ***Capability Service Abstraction*** – enables Transformation components to use one or more capabilities from the Capability Management Component so that given datasets and algorithms can be securely transferred, stored, and executed by the specific computing fabrics such as data storage, networking, and computing cluster.
3. ***Usage Service Abstraction*** – enables Transformation components to provide their results using different formats, operation and service models such as data retrieval, reporting, and rendering.]

[NBD RA shows the flow of (big) data throughout the system using the red arrows. This data flow is established, managed, and controlled by the signaling programmable interfaces described above.]

[Editor’s Note: review and agree on the direction of the data flow.]



# Big Data Reference Architecture: Main Components

## Data Sources

[This section will discuss different aspects of big data services, such as

1. Data characteristics (variety, volume, velocity, etc.)
2. Sources of data (internal, external, public)]

## Transformation

[This section will introduce the broad inclusive meaning behind the concept of “transformation”.]

[Editor’s note: “Transformation” is a working term that may be replaced in the future.]

[This section will describe a number of specific popular/common examples including “collection”, “curation”, “analytics”, “visualization”, and “access” as suggested below.]

### Collection Services

### Curation Services

### Analytic Services

### Visualization Services

### Access Services

## Capability Management

### Big Data Framework

### Hardware Management

### Security and Privacy Management

### System Management

### Lifecycle Management

## Data Usage

### Data Reporting

### Data Retrieval

### Data Rendering

# Big Data Reference Architecture: Interfaces

[Editor’s Note: Consider how to avoid repetitions and keep the alignment between the interfaces’/services’ examples in this chapter (#5) with the components’ description in the previous chapter (#4).]

## Data Service Abstraction

### Overview

### Example: Collection

### Example: Curation

### Example: Analytic

## Capability Service Abstraction

### Overview

### Example: Streaming

### Example: Aggregation

### Example: Integration

### Example: Transfer

### Example: Search

### Example: Statistics

### Example: RT Analytics

### Example: Batch Analytics

### Example: Interactive Animation

## Usage Service Abstraction

### Overview

### Example: Retrieval

### Example: Rendering

### Example: Report

# Big Data Reference Architecture: Security and Privacy

**[This section will be prepared by the NIST BDWG Security and Privacy SG and will contain high level security and privacy considerations relevant to the design of the Reference Architecture.]**

# Big Data Taxonomy

**[This section will be prepared by the NIST BDWG Def&Tax SG and will contain high level taxonomy relevant to the design of the Reference Architecture.]**

# Appendix A: Big Data Terms and Definitions

**[This section will contain terms and definitions used in this document.]**

# Appendix B: Acronyms

# Appendix C: References