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

**NBD-WD-2013/M0125**

**Source: Requirements Subgroup**

**Status: Draft, v2**

**Title: High-level Requirements Extraction based from Submitted Use Cases (M0096)**

**Author: Wo Chang (NIST)**

1. **Background**

The following high-level requirements were extracted from the submitted use cases of M0096. The requirements are divided based from the four key components (Data Sources, Transformation, Data Infrastructure, and Data Usage) and along with subcomponents (Security and Privacy, Lifecycle Management, and System Management) from within the Data Infrastructure. The numbers associated with each requirement such as (t: i, i+1, n) referred to “t” as the total number use cases referenced to the respective requirement and the “i,i+1,n” represents the specific use case called for such requirement.

1. **How to Extract Requirements from Use Cases**

Table below shows the use case template adopted after discussion within the requirements/ use-case subgroup and within input from other groups so that information from the use cases can generate requirements that feed into architecture discussions. This impact is described in section 3 where feedback from use cases is classified by the seven architecture components listed below where we also give some of use case sections that will drive input to each component.

1. Data Source: from Big Data Characteristics, Data Types
2. Transformation (or filter): from Data Analytics
3. Resource Requirements: from Current Solutions
4. Data Usage: from Goals, Use Case Description and Visualization
5. Security & Privacy: from Security & Privacy Requirements
6. Lifecycle Management: from Veracity, and Data Quality
7. System Management and Other issues

Note that use cases explicitly or implicitly specify requirements and further give details on how the problem is tackled today. Note also user requirements leave many questions unanswered as they will not for example specify the system management directly; they may have important system management effects identified from their current solutions.

**NBD (NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | |  | |
| **Vertical (area)** | |  | |
| **Author/Company/Email** | |  | |
| **Actors/Stakeholders and their roles and responsibilities** | |  | |
| **Goals** | |  | |
| **Use Case Description** | |  | |
| **Current**  **Solutions** | **Compute(System)** | |  |
| **Storage** | |  |
| **Networking** | |  |
| **Software** | |  |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | |  |
| **Volume (size)** | |  |
| **Velocity**  **(e.g. real time)** | |  |
| **Variety**  **(multiple datasets, mashup)** | |  |
| **Variability (rate of change)** | |  |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues, semantics)** | |  |
| **Visualization** | |  |
| **Data Quality (syntax)** | |  |
| **Data Types** | |  |
| **Data Analytics** | |  |
| **Big Data Specific Challenges (Gaps)** | |  | |
| **Big Data Specific Challenges in Mobility** | |  | |
| **Security & Privacy**  **Requirements** | |  | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | |  | |
| **More Information (URLs)** | |  | |
| **Note:** <additional comments> | | | |

**Note: No proprietary or confidential information should be included**

1. **Requirements**

The numbers associated with each requirement such as (t: i, i+1, n) where “t” is the total number use cases reference to the respective requirement and the “I,i+1,n” represent the specific use case called for such requirement.

**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.

(6: M0166, M0167, M0165, M0078, M0090, M0103)

DSR-2: needs to support slow and high throughput data transmission between data sources and computing clusters. (2: M0167, M0078)

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

(5: M0167, M0165, M0078, M0089, M0090)

**Transformation Requirements:**

TR-1: needs to support diversified analytic processing and machines learning techniques (6: M0166, M0164, M0167, M0078, M0089, M0103)

TR-2: needs to support batch and real time analytic processing (4: M0164, M0165, M0090, M0103)

TR-3: needs to support processing diversified data content (2: M0166, M0089)

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

(6: M0166, M0164, M0165, M0078, M0090, M0103)

TR-5: needs to support legacy and advance programming executable and libraries (5: M0166, M0164, M0167, M0078, M0089)

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

DIR-1: needs to support legacy and advance software packages (subcomponent: SaaS) (5: M0166, M0164, M0167, M0078, M0089)

DIR-2: needs to support legacy and advance computing platforms (subcomponent: PaaS) (3: M0164, M0078, M0089)

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

(6: M0166, M0164, M0167, M0078, M0089, M0090)

DIR-4: needs to support elastic data transmission (subcomponent: networking) (3: M0089, M0090, M0103)

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

(6: M0166, M0164, M0167, M0165, M0078, M0089,)

**Data Usage Requirements:**

DUR-1: needs to support fast search (~0.1 seconds) from processed data (1: M0165)

DUR-2: needs to support diversified output file formats for rendering (7: M0166, M0164, M0167, M0165, M0078, M0089, M0090)

DUR-3: needs to support visual layout for results presentation (2: M0167, M0165)

DUR-4: needs to support rich user interface (2: M0167, M0089)

DUR-5: needs to support streaming results to clients (1: M0164)

**Security & Privacy Requirements:**

SnPR-1: needs to support security and privacy on protected data (7: M0166, M0164, M0167, M0165, M0078, M0089, M0103)

SnPR-2: needs to support multi-level access control on protected data (6: M0166, M0167, M0165, M0078, M0089, M0103)

**Lifecycle Management Requirements:**

LMR-1: needs to support data quality curation (3: M0166, M0167, M0165)

LMR-2: needs to support dynamic updates on data and user profiles (1: M0164)

LMR-3: needs to support data lifecycle policy (1: M0165)

LMR-4: needs to support data validation (1: M0090)

LMR-5: needs to support human annotation for data validation (1: M0089)

**System Management Requirements:**

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

(1: M0164, M0078)

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

(1: M0167)

SMR-3: needs to support rich visual content rendering from mobile platforms (4: M0166, M0165, M0078, M0089)

1. **Use Cases with Requirements Based from the RA Components**

Note: “C#1” refers to Use Case #1 from the M0096. The requirements in RED are corresponding to Section 2 as for cross referencing.

|  |  |
| --- | --- |
| Data Sources | **M0166**  1. needs to support real time data from Accelerator and Analysis instruments  2. needs to support a-synchronization data collection 3. needs to support calibration of instruments  **M0164** 1. needs to support user profiles and ranking info  **M0167** 1. needs to provide reliable data transmission from aircraft sensors/instruments or removable disks from remote sites 2. needs to support data gathering in real time 3. needs to support varieties of datasets  M0165 1. needs to support distributed data sources 2. needs to support streaming data 3. needs to support multimedia content  **M0078** 1. needs to support high throughput compressed data (300GB/day) from various DNA sequencers 2. needs to support distributed data source (sequencers) 3. needs to support various file formats either structured and unstructured data  **M0089** 1. needs to support high resolution spatial digitized pathology images 2. needs to support various image quality analysis algorithms 3. needs to support various image data formats especially BIGTIFF with structured data for analytical results  4. needs to support image analysis, spatial queries and analytics, feature clustering and classification  **M0090** 1. needs to support real time distributed datasets 2. needs to support various formats, resolution, semantics, and metadata  **M0103** 1. needs to support centralized and real time distributed sites/sensors  **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. (6: M0166,M0167,M0165,M0078,M0090,M0103)  DSR-2: needs to support slow and high throughput data transmission between data sources and computing clusters. (2: M0167,M0078)  DSR-3: needs to support diversified data content ranging from text to multimedia to instrumental data. (5: M0167, M0165, M0078, M0089, M0090) |
| Transformation | **M0166** 1. needs to support experimental data from ALICE, ATLAS, CMS, LHb 2. needs to support histograms, scatter-plots with model fits 3. needs to support Monte-Carlo computations  **M0164**  1. needs to support streaming video contents to multiple clients 2. needs to support analytic processing for matching clients' interest in movie selection 3. needs to support various analytic processing techniques for consumer personalization 4. needs to support robust learning algorithms 5. needs to support continued analytic processing based on the monitoring and performance results  **M0167** 1. needs to support legacy software (Matlab) and language (C/Java) binding for processing 2. needs signal processing and advance image processing to find layers  **M0165** 1. needs to support dynamic fetching content over the network 2. needs to support linking user profiles and social network data  **M0078** 1. needs to support for processing raw data in variant calls 2. needs to support machine learning for complex analysis on systematic errors from sequencing technologies are hard to characterize  **M0089** 1. needs to support high performance image analysis to extract spatial information 2. needs to support spatial queries and analytics, and feature clustering and classification 3. needs to support analytic processing on huge multi-dimensional large dataset and be able to correlate with other data types such as clinical data, -omic data.  **M0090** 1. needs to support MapReduce, SciDB, and other scientific databases 2. needs to support continuously computing for updates 3. needs to support event-specification language for data mining and event searching 4. needs to support semantics interpretation and optimal structuring for 4-dimensional data mining and predictive analysis  **M0103** 1. needs to support tracking items based on the unique identification with its sensor information, GPS coordinates 2. needs to support real time updates on tracking items  **Transformation Requirements:**  TR-1: needs to support diversified analytic processing and machines learning techniques (6: M0166, M0164, M0167, M0078, M0089, M0103)  TR-2: needs to support batch and real time analytic processing (4: M0164, M0165, M0090, M0103)  TR-3: needs to support processing diversified data content (2: M0166, M0089)  TR-4: needs to support processing data in motion (streaming, fetching new content, tracking, etc.) (6: M0166, M0164, M0165, M0078, M0090, M0103)  TR-5: needs to support legacy and advance programming executable and libraries (5: M0166, M0164, M0167, M0078, M0089) |
| Data Infrastructure | **M0166** 1. needs to support legacy computing infrastructure (computing nodes) 2. needs to support distributed cached files (storage) 3. needs to support object databases (sw pkg)  **M0164** 1. needs to support Hadoop (platform) 2. needs to support Pig (language) 3. needs to support Cassandra and Hive 4. needs to support huge subscribers, ratings, and searching per day (DB) 5. needs to support huge storage (2 PB) 6. needs to support I/O intensive processing  **M0167** 1. needs to support ~0.5 Petabytes/year of raw data 2. needs to support transfer content from removable disk to computing cluster for parallel processing 3. needs to support MapReduce or MPI plus language binding for C/Java  **M0165** 1. needs to support petabytes of text and rich media (storage)  **M0078** 1. needs to support legacy computing cluster and other PaaS and IaaS (computing cluster) 2. needs to support huge data storage in PB range (storage) 3. needs to support Unix-based legacy sequencing bioinformatics software (sw pkg)  **M0089** 1. needs to support legacy system and cloud (computing cluster) 2. needs to support huge legacy and new storage such as SAN or HDFS (storage) 3. needs to support high throughput network link (networking) 4. needs to support MPI image analysis, MapReduce, Hive with spatial extension (sw pkgs)  **M0090** 1. needs to support other legacy computing systems (e.g. supercomputer) 2. needs to support high throughput data transmission over the network **M0103** 1. needs to support Internet connectivity  **Data Infrastructure Requirements (to enable Transformation processing):**  DIR-1: needs to support legacy and advance software packages (subcomponent: SaaS) (5: M0166, M0164, M0167, M0078, M0089)  DIR-2: needs to support legacy and advance computing platforms (subcomponent: PaaS) (3: M0164, M0078, M0089)  DIR-3: needs to support legacy and advance distributed computing cluster (subcomponent: IaaS) (6: M0166, M0164, M0167, M0078, M0089, M0090)  DIR-4: needs to support elastic data transmission (subcomponent: networking) (3: M0089, M0090, M0103)  DIR-5: needs to support legacy and advance distributed data storage (subcomponent: storage) (6: M0166, M0164, M0167 M0165, M0078, M0089) |
| Data Usage | **M0166** 1. needs to support histograms and model fits (visual)  **M0164**  1. needs to support streaming and rendering media??  **M0167** 1. needs to support GIS user interface 2. needs to support rich user interface for simulations  **M0165** 1. needs to support search time in ~0.1 seconds 2. needs to support top 10 ranked results 3. needs to support page layout (visual)  **M0078** 1. needs to support data format for "Genome browsers"  **M0089** 1. needs to support visualization for validation and training  **M0090** 1. needs to support visualization to interpret results  **Data Usage Requirements:**  DUR-1: needs to support fast search (~0.1 seconds) from processed data (1: M0165)  DUR-2: needs to support diversified output file formats for rendering (7: M0166, M0164, M0167 M0165, M0078, M0089, M0090)  DUR-3: needs to support visual layout for results presentation (2: M0167, M0165)  DUR-4: needs to support rich user interface (2: M0167, M0089)  DUR-5: needs to support streaming results to clients (1: M0164) |
| Security & Privacy | **M0166** 1. needs to support data protection  **M0164** 1. needs to support preservation of users' privacy and digital rights for media  **M0167** 1. needs to support security and privacy on political sensitive issues 2. needs to support dynamic security and privacy policy mechanisms  **M0165** 1. needs to support access control  2. needs to protect sensitive content  **M0078** 1. needs to support security and privacy protection on health records and clinical research databases  **M0089** 1. needs to support security and privacy protection for protected health information  **M0103** 1. needs to support security policy  **Security & Privacy Requirements:**  SnPR-1: needs to support security and privacy on protected data (7: M0166, M0164, M0167 M0165, M0078, M0089, M0103)  SnPR-2: needs to support multi-level access control on protected data (6: M0166, M0167 M0165, M0078, M0089, M0103) |
| Lifecycle Management | **M0166** 1. needs to support data quality on complex apparatus  **M0164** 1. needs to support continued ranking and updating based on user profile and analytic results  **M0167** 1. needs to support data quality assurance  **M0165** 1. needs to support purge data after certain time interval (few months) 2. needs to support data cleaning  **M0089** 1. needs to support human annotations for validation  **M0090** 1. needs to support validation for output products (correlations)  **Lifecycle Management Requirements:**  LMR-1: needs to support data quality curation (3: M0166, M0167, M0165)  LMR-2: needs to support dynamic updates on data and user profiles (1: M0164)  LMR-3: needs to support data lifecycle policy (1: M0165)  LMR-4: needs to support data validation(1: M0090)  LMR-5: needs to support human annotation for data validation (1: M0089) |
| System Management | **M0164** 1. needs to support smart interface accessing movie content on mobile platforms  **M0167** 1. needs to support monitoring data collection instruments/sensors  **M0165** 1. needs to support mobile search and rendering  **M0078** 1. needs to support mobile platforms for physicians accessing genomic data (mobile device)  **M0089** 1. needs to support 3D visualization and rendering on mobile platforms  **System Management Requirements:**  SMR-1: needs to support rich user interface from mobile platforms to access processed results (1: M0164, M0078)  SMR-2: needs to support performance monitoring on analytic processing from mobile platforms (1: M0167)  SMR-3: needs to support rich visual content rendering from mobile platforms (4: M0166, M0165, M0078, M0089) |