**NIST Big Data**

**General Requirements**

**Version 1.1**

**Requirements & Use Cases Subgroup**

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

**August, 2013**

[Executive Summary 4](#_Toc364851168)

[1 Introduction 4](#_Toc364851169)

[1.1 Objectives 4](#_Toc364851170)

[1.2 How This Report Was Produced 4](#_Toc364851171)

[1.3 Structure of This Report 4](#_Toc364851172)

[2 Use Case Scenarios – Vertical Applications 4](#_Toc364851173)

[2.1 Large-scale Deep Learning (Machine Learning/AI) 4](#_Toc364851174)

[2.2 UAVSAR Data Processing, Data Product Delivery, and Data Services (Scientific Research: Earth Science) 4](#_Toc364851175)

[2.3 MERRA Analytic Services MERRA/AS (Scientific Research: Earth Science) 4](#_Toc364851176)

[2.4 IaaS (Infrastructure as a Service) Big Data Business Continuity & Disaster Recovery (BC/DR) Within A Cloud Eco-System (Large Scale Reliable Data Storage) 5](#_Toc364851177)

[2.5 DataNet Federation Consortium DFC (Scientific Research: Collaboration Environments) Reagan Moore, University of North Carolina at Chapel Hill 5](#_Toc364851178)

[2.6 Semantic Graph-search on Scientific Chemical and Text-based Data (Management of Information from Research Articles) 5](#_Toc364851179)

[2.7 Atmospheric Turbulence - Event Discovery and Predictive Analytics (Scientific Research: Earth Science) 5](#_Toc364851180)

[2.8 Pathology Imaging/digital pathology (Healthcare) 5](#_Toc364851181)

[2.9 Genomic Measurements (Healthcare) 6](#_Toc364851182)

[2.10 Cargo Shipping (Industry) 6](#_Toc364851183)

[2.11 Radar Data Analysis for CReSIS (Scientific Research: Polar Science and Remote Sensing of Ice Sheets) 6](#_Toc364851184)

[2.12 Particle Physics: Analysis of LHC Large Hadron Collider Data: Discovery of Higgs particle (Scientific Research: Physics) 6](#_Toc364851185)

[2.13 Netflix Movie Service (Commercial Cloud Consumer Services) 7](#_Toc364851186)

[2.14 Web Search (Commercial Cloud Consumer Services) 7](#_Toc364851187)

[3 Use Case Requirements 7](#_Toc364851188)

[3.1 Use Case Template 7](#_Toc364851189)

[3.2 Data Source Requirements 9](#_Toc364851190)

[3.3 Transformation Requirements 9](#_Toc364851191)

[3.4 Resource Requirements 9](#_Toc364851192)

[3.5 Data Usage Requirements 9](#_Toc364851193)

[3.6 Security & Privacy Requirements 9](#_Toc364851194)

[3.7 Lifecycle Management Requirements 9](#_Toc364851195)

[3.8 System Management and Other Requirements 9](#_Toc364851196)

[4 Service Abstractions (SAs) Requirements 9](#_Toc364851198)

[4.1 What is an SA? 9](#_Toc364851200)

[4.2 Service Abstractions Objectives 9](#_Toc364851201)

[4.3 Service Abstractions Management 10](#_Toc364851202)

[4.4 Types of Service Abstractions 10](#_Toc364851203)

[4.4.1 Data Service Abstraction 10](#_Toc364851204)

[4.4.2 Transport Service Abstraction 10](#_Toc364851205)

[4.4.3 Usage Service Abstraction 10](#_Toc364851206)

[5 Conclusions and Recommendations 10](#_Toc364851207)

[6 Reference 10](#_Toc364851208)

[Appendix A: Raw Use Cases Data 10](#_Toc364851209)

# Executive Summary

# Introduction

## Objectives

## How This Report Was Produced

## Structure of This Report

# Use Case Scenarios – Vertical Applications

## Large-scale Deep Learning (Machine Learning/AI)

## UAVSAR Data Processing, Data Product Delivery, and Data Services (Scientific Research: Earth Science)

**Application Description**

Synthetic Aperture Radar (SAR) can identify landscape changes caused by seismic activity, landslides, deforestation, vegetation changes and flooding. This is for example used to support earthquake science as well as disaster management. This use case supports the storage, application of image processing and visualization of this geo-located data with angular specification. Data from planes and satellites is processed on NASA computers before being stored with substantial data communication; clouds are suitable hosts but not used today in production. The data is made public as soon as processed and requires significant curation due to instrumental glitches. The current ~150TB data size could increase dramatically if Earth Radar Mission launched.

**Current Approach**

**Future Approach**

## MERRA Analytic Services MERRA/AS (Scientific Research: Earth Science)

**Application Description**

This MapReduce application produces global temporally and spatially consistent synthesis of 26 key climate variables by combining numerical simulations with observational data. Three-dimensional results are produced every 6-hours extending from 1979-2012. This supports important applications like the intergovernmental Panel on Climate Change (IPCC) research and the NASA/Department of Interior RECOVER wild land fire decision support system; these applications typically involve integration of MERRA with other datasets. There is a current total of 480TB of data growing by one TB a month. The current system is hosted on a 36 node Infiniband cluster but clouds are being investigated.

**Current Approach**

**Future Approach**

## IaaS (Infrastructure as a Service) Big Data Business Continuity & Disaster Recovery (BC/DR) Within A Cloud Eco-System (Large Scale Reliable Data Storage)

## DataNet Federation Consortium DFC (Scientific Research: Collaboration Environments) Reagan Moore, University of North Carolina at Chapel Hill

## Semantic Graph-search on Scientific Chemical and Text-based Data (Management of Information from Research Articles)

## Atmospheric Turbulence - Event Discovery and Predictive Analytics (Scientific Research: Earth Science)

**Application Description**

This builds data mining on top of reanalysis products including the North American Regional Reanalysis (NARR) and the Modern-Era Retrospective-Analysis for Research (MERRA) from NASA where latter described earlier. The analytics correlate aircraft reports of turbulence (either from pilot reports or from automated aircraft measurements of eddy dissipation rates) with recently completed atmospheric re-analyses. This is of value to aviation industry and to weather forecasters. There are no standards for re-analysis products complicating system where MapReduce is being investigated. The reanalysis data is hundreds of terabytes and slowly updated whereas turbulence is smaller in size and implemented as a streaming service. The initial turbulence case can be extended to other ocean/atmosphere phenomena but the analytics would vary in each case.

**Current Approach**

**Future Approach**

## Pathology Imaging/digital pathology (Healthcare)

**Application Description**

**Current Approach**

**Future Approach**

## Genomic Measurements (Healthcare)

**Application Description**

**Current Approach**

**Future Approach**

## Cargo Shipping (Industry)

**Application Description**

**Current Approach**

**Future Approach**

## Radar Data Analysis for CReSIS (Scientific Research: Polar Science and Remote Sensing of Ice Sheets)

**Application Description**

This data feeds into intergovernmental Panel on Climate Change (IPCC) and uses custom radars to measures ice sheet bed depths and (annual) snow layers at the North and South poles and mountainous regions. The initial analysis is currently Matlab signal processing that produces a set of radar images. These cannot be transported from field over Internet and are typically copied to removable few TB disks in the field and flown “home” for detailed analysis. Image understanding tools with some human oversight find the image features (layers) that are stored in a database front-ended by a Geographical Information System. The ice sheet bed depths are used in simulations of glacier flow. The data is taken in “field trips” that each currently gather 50-100 TB of data but an order of magnitude more is projected with improved instrumentation. Demands of processing increasing field data in an environment with more data but still constrained power budget, suggests low power/performance architectures such as GPU systems.

**Current Approach**

**Future Approach**

## Particle Physics: Analysis of LHC Large Hadron Collider Data: Discovery of Higgs particle (Scientific Research: Physics)

**Application Description**

**Current Approach**

**Future Approach**

## Netflix Movie Service (Commercial Cloud Consumer Services)

**Application Description**

**Current Approach**

**Future Approach**

## Web Search (Commercial Cloud Consumer Services)

**Application Description**

**Current Approach**

**Future Approach**

# Use Case Requirements

## Use Case Template

Figure 1 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:

**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**

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.

## Data Source Requirements

## Transformation Requirements

## Resource Requirements

## Data Usage Requirements

## Security & Privacy Requirements

## Lifecycle Management Requirements

## System Management and Other Requirements



# Service Abstractions (SAs) Requirements



## What is an SA?

## Service Abstractions Objectives

## Service Abstractions Management

## Types of Service Abstractions

### Data Service Abstraction

### Transport Service Abstraction

### Usage Service Abstraction

# Conclusions and Recommendations

# Reference

# Appendix A: Raw Use Cases Data