Big Data Architecture Research at UvA

Yuri Demchenko
System and Network Engineering Group, University of Amsterdam

ISO/IEC SGBD Big Data Technologies Workshop
Part of ISO/IEC Big Data Study Group meeting
13-16 May 2014
Outline

• Research on Big Data and Infrastructure technologies at
• Big Data definition
  – From 5 + 1 V’s to 5 parts Big Data Definition
• Paradigm change and new challenges
  – Data centric model and DataBus
• Defining Big Data Architecture Framework (BDAF)
  – From Architecture to Ecosystem to Architecture Framework
• Big Data Infrastructure (BDI) and Big Data Lifecycle Management model
Big Data and Security Research at System and Network Engineering, University of Amsterdam

- Long time research and development on Infrastructure services and facilities
  - High speed optical networking and data intensive applications
  - Semantic description of infrastructure and network services
  - Collaborative systems, Grid, Clouds and currently Big Data
- Focus on Infrastructure definition and services
  - Software Defined Infrastructure based on Cloud/Intercloud technologies
  - Dynamically provisioned security infrastructure and services
- NIST Big Data Working Group
  - Contribution to Reference Architecture, Big Data Definition and Taxonomy, Big Data Security
- Research Data Alliance
  - Interest Group on Education and Skills Development on Data Intensive Science
  - Big Data Analytics Interest Group
- Big Data Interest Group at UvA
  - Non-formal but active, meets two-weekly/monthly
  - Provided input to NIST BD-WG and RDA activities and UvA DSRC
Visionaries and Drivers:
Seminal works, High level reports, Activities

The Fourth Paradigm: Data-Intensive Scientific Discovery.

Riding the wave: How Europe can gain from the rising tide of scientific data.

AAA Study: Study on AAA Platforms For Scientific data/information Resources in Europe,
TERENA, UvA, LIBER, UinvDeb.
(2011-2012)

ISO/IEC SGBD, 13-16 May 2014

NIST Big Data Working Group (NBD-WG)
https://www.rd-alliance.org/
Drivers at SNE/UvA

• Ongoing research on Cyber Infrastructure
• Demand for education on new emerging technologies
• ENVRI EU project
• LifeWatch EU project
• EUBrazil Cloud Connect EU-Brazil project
  – Consortium of 6 Brazilian institutions and 7 European institutions
  – 3 scientific and research use cases
Use Case 1: Leishmaniasis Virtual Laboratory

- Led by ISCIII / FIOCRUZ.
- **Objective:** Improve knowledge on the distribution and susceptibility of epidemiology outburst in Leishmaniasis Disease
- **Technical Challenge:** Easy access to computing and data federation for applications defined as workflows.
- **International Added Value:** Linking data from Brazilian and European leaders and complementary databases and develop a Virtual Research Environment for integrating workflows for epidemiology risk modelling.
Use Case 2: Heart Simulation

- **Led by:** BSC & LNCC.
- **Objective:** Increase the accuracy of blood simulation.
- **Technical Challenge:** Integrate Supercomputing and Cloud computing applications.
- **International Added Value:** Linking boundary conditions of the ADAM Vascular system to the ALYA multilevel heart simulator to achieve beyond the state-of-the-art simulation of the whole Human Vascular System Simulation.
Use Case 3: Biodiversity and Climate Change

- **Led by**: CMCC & UFCG.
- **Objective**: Understand the impact of climate change on terrestrial biodiversity through two workflows based on Earth observation and ground level data.
- **Technical Challenge**: Integrate parallel data analysis with other processing workflows in a geographically distributed environment.
- **International Added Value**: Integration of biodiversity data and modelling with multispectral and remote sensing data for studying the cross-correlation of biodiversity and climate change.

![Climate & Biodiversity Clearing-house](image)
Gartner Technology Hypercycle (October 2013)

Source: http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp
Our/SNE Big Data Technology Research Cycle

Big Data

2011

Mid-End 2013

Cloud Computing

End 2014

Mid 2014

Active research into Big Data domain definition
Building community

Component technologies mastering
Education courses development

Remote BD technology following.
EU Study AAA for Research Data
Main research in Cloud/Intercloud

New style of technology development
Technology consolidation

Source: http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp
• IDC definition of Big Data (conservative and strict approach) :
"A new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis"

• Gartner definition
Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. [http://www.gartner.com/it-glossary/big-data/](http://www.gartner.com/it-glossary/big-data/)
  – Termed as 3 parts definition, not 3V definition

• Big Data: a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques.

• “Data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn’t fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it.”
  – Ed Dumbill, program chair for the O'Reilly Strata Conference

• Termed as the Fourth Paradigm *)
“The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration.” (Jim Gray, computer scientist)

Improved: 5+1 V’s of Big Data

Generic Big Data Properties
- Volume
- Variety
- Velocity

Acquired Properties (after entering system)
- Value
- Veracity
- Variability

6 Vs of Big Data

Volume:
- Terabytes
- Records/Arch
- Tables, Files
- Distributed

Velocity:
- Batch
- Real/near-time
- Processes
- Streams

Value:
- Correlational
- Statistical
- Events
- Hypothetical

Variability:
- Correlations
- Statistical
- Events
- Hypothetical

Veracity:
- Trustworthiness
- Authenticity
- Origin, Reputation
- Availability
- Accountability

Variety:
- Structured
- Unstructured
- Multi-factor
- Probabilistic
- Linked
- Dynamic

- Changing data
- Changing model
- Linkage

Volume:
- Structured
- Unstructured
- Multi-factor
- Probabilistic
- Linked
- Dynamic

Variety:
- Changing data
- Changing model
- Linkage

Variability:
- Correlations
- Statistical
- Events
- Hypothetical

Veracity:
- Trustworthiness
- Authenticity
- Origin, Reputation
- Availability
- Accountability

Commonly accepted 3V's of Big Data
- Terabytes
- Records
- Transactions
- Tables, Files

- Structured
- Unstructured
- Semi-structured
- All the above
Big Data Definition: From 5+1V to 5 Parts (1)

(1) Big Data Properties: 5V
   – Volume, Variety, Velocity, Value, Veracity
   – Additionally: Data Dynamicity (Variability)

(2) New Data Models
   – Data Lifecycle and Variability
   – Data linking, provenance and referral integrity

(3) New Analytics
   – Real-time/streaming analytics, interactive and machine learning analytics

(4) New Infrastructure and Tools
   – High performance Computing, Storage, Network
   – Heterogeneous multi-provider services integration
   – New Data Centric (multi-stakeholder) service models
   – New Data Centric security models for trusted infrastructure and data processing and storage

(5) Source and Target
   – High velocity/speed data capture from variety of sensors and data sources
   – Data delivery to different visualisation and actionable systems and consumers
   – Full digitised input and output, (ubiquitous) sensor networks, full digital control
Big Data Definition: From 5+1V to 5 Parts (1)

(1) Big Data Properties: 5V
   – Volume, Variety, Velocity, Value, Veracity
   – Additionally: Data Dynamicity (Variability)

(2) New Data Models
   – Data linking, provenance and referral integrity
   – Data Lifecycle and Variability/Evolution

(3) New Analytics
   – Real-time/streaming analytics, interactive and machine learning analytics

(4) New Infrastructure and Tools
   – High performance Computing, Storage, Network
   – Heterogeneous multi-provider services integration
   – New Data Centric (multi-stakeholder) service models
   – New Data Centric security models for trusted infrastructure and data processing and storage

(5) Source and Target
   – High velocity/speed data capture from variety of sensors and data sources
   – Data delivery to different visualisation and actionable systems and consumers
   – Full digitised input and output, (ubiquitous) sensor networks, full digital control
Refining Gartner definition

“Big data is (1) high-volume, high-velocity and high-variety information assets that demand (3) cost-effective, innovative forms of information processing for (5) enhanced insight and decision making”

- **Big Data (Data Intensive) Technologies** are targeting to process (1) high-volume, high-velocity, high-variety data (sets/assets) to extract intended data value and ensure high-veracity of original data and obtained information that demand cost-effective, innovative forms of data and information processing (analytics) for enhanced insight, decision making, and processes control; all of those demand (should be supported by) **new data models** (supporting all data states and stages during the whole data lifecycle) and **new infrastructure services and tools** that allows also obtaining (and processing data) **from a variety of sources** (including sensor networks) and delivering data in a variety of forms to different data and information consumers and devices.

(1) Big Data Properties: 5V
(2) New Data Models
(3) New Analytics
(4) New Infrastructure and Tools
(5) Source and Target
From Big Data to All-Data – Paradigm Change

Breaking paradigm changing factor
- Data storage and processing
- Security
- Identification and provenance

Traditional model
- BIG Storage and BIG Computer with FAT pipe
- Move compute to data vs Move data to compute

New Paradigm
- Continuous data production
- Continuous data processing
- **DataBus as a Data container and Protocol**

Big Data

Network

Big Computer

?

Move or not to move?

Distributed Big Data Storage

Data Abstraction

Data Bus

Infrastructure Abstraction

Distributed Compute and Analytics

DataBus:
1. Data Container
2. Metadata, State
3. Data Transfer Protocol

ISO/IEC SGBD, 13-16 May 2014
• Current IT and communication technologies are host based or host centric
  – Any communication or processing are bound to host/computer that runs software
  – Especially in security: all security models are host/client based

• Big Data requires new data-centric models
  – Data location, search, access
  – Data integrity and identification
  – Data lifecycle and variability
  – Data centric (declarative) programming models
  – Data aware infrastructure to support new data formats and data centric programming models

• Data centric security and access control
Defining Big Data Architecture Framework

**Architecture vs Ecosystem**
- Big Data undergo a number of transformations during their lifecycle
- Big Data fuel the whole transformation chain
  - Data sources and data consumers, target data usage
- Multi-dimensional relations between
  - Data models and data driven processes
  - Infrastructure components and data centric services

**Architecture vs Architecture Framework**
- Separates concerns and factors
  - Control and Management functions, orthogonal factors
- Architecture Framework components are inter-related
Big Data Architecture Framework (BDAF) (1)

(1) Data Models, Structures, Types
- Data formats, non/relational, file systems, etc.

(2) Big Data Management
- Big Data Lifecycle (Management) Model
  - Big Data transformation/staging
- Provenance, Curation, Archiving

(3) Big Data Analytics and Tools
- Big Data Applications
  - Target use, presentation, visualisation

(4) Big Data Infrastructure (BDI)
- Storage, Compute, (High Performance Computing,) Network
- Sensor network, target/actionable devices
- Big Data Operational support

(5) Big Data Security
- Data security in-rest, in-move, trusted processing environments
<table>
<thead>
<tr>
<th>Col: Used By</th>
<th>Data Models Structures</th>
<th>Data Managmnt &amp; Lifecycle</th>
<th>BigData Infrastr &amp; Operations</th>
<th>BigData Analytics &amp; Applicatn</th>
<th>BigData Security</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row: Requires This</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Models &amp; Structures</td>
<td>+</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td></td>
</tr>
<tr>
<td>Data Managmnt &amp; Lifecycle</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td></td>
</tr>
<tr>
<td>BigData Infrastruct &amp; Operations</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>+++</td>
<td></td>
</tr>
<tr>
<td>BigData Analytics &amp; Applications</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td>++</td>
<td></td>
</tr>
<tr>
<td>BigData Security</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+</td>
<td></td>
</tr>
</tbody>
</table>
Big Data Ecosystem: Data, Transformation, Infrastructure

Data Source

Data Collection & Registration → Data Filter/Enrich, Classification → Data Analytics, Modeling, Prediction → Data Delivery, Visualisation → Consumer

Big Data Target/Customer: Actionable/Usable Data
Target users, processes, objects, behavior, etc.

Big Data Source/Origin (sensor, experiment, logdata, behavioral data)

Big Data Analytic/Tools

Storage General Purpose → Compute General Purpose → High Performance Computer Clusters → Storage Specialised Databases Archives (analytics DB, in memory, operational)

Intercloud multi-provider heterogeneous Infrastructure

Data Management

• Security Infrastructure
• Network Infrastructure: Internal

• Infrastructure Management/Monitoring

Data categories: metadata, (un)structured, (non)identifiable
General BDI services and components

- Data management infrastructure and tools
- Registries, search/indexing, ontologies, schemas, namespace
- Collaborative Environment (user/groups managements)
- Heterogeneous multi-provider Inter-cloud infrastructure
  - Compute, Storage, Network (provisioned on-demand dynamically scaling)
  - Federated Access and Delivery Infrastructure (FADI)
- Advanced high performance (programmable) network
- Security infrastructure (access control, Identity and policy management, confidentiality, privacy, trust)
Big Data Infrastructure and Analytic Tools

Big Data Source/Origin (sensor, experiment, logdata, behavioral data)

Big Data Target/Customer: Actionable/Usable Data
Target users, processes, objects, behavior, etc.

Big Data Analytic/Tools

Analytics: Refinery, Linking, Fusion
Analytics: Realtime, Interactive, Batch, Streaming

Analytics Applications: Link Analysis, Cluster Analysis, Entity Resolution, Complex Analysis

Storage General Purpose
Compute General Purpose
High Performance Computer Clusters
Storage Specialised Databases, Archives

Data Management

Data categories: metadata, (un)structured, (non)identifiable

Intercloud multi-provider heterogeneous Infrastructure

Security Infrastructure
Network Infrastructure Internal
Infrastructure Management/Monitoring

Federated Access and Delivery Infrastructure (FADI)
Big Data Analytics Infrastructure

- High Performance Computer Clusters (HPCC)
- Specialised Storage, Distributed/Replicated, Archives, Databases, SQL/NoSQL
- Big Data Analytics Tools/Applications
- Analytics/processing: Real-time, Interactive, Batch, Streaming
- Link Analysis, Graph analysis
- Cluster Analysis
- Entity Resolution
- Complex Analysis
Data Transformation/Lifecycle Model

- Does Data Model changes along lifecycle or data evolution?
- Identifying and linking data
  - Persistent identifier

Common Data Model?
- Data Variety and Variability
- Semantic Interoperability

Data Storage

Data (inter)linking?
- Persistent ID
- Identification
- Privacy, Opacity

Data Variety and Variability
- Semantic Interoperability

Data Model (1) → Data Model (1) → Data Model (3) → Data Model (4)
Scientific Data Lifecycle Management (SDLM) Model

Data Lifecycle Model in e-Science

- Data discovery
- Data collection and filtering
- Data analysis
- Data sharing/Data publishing
- Data archiving
- Data linkage to papers
- Data cleanup and retirement
- Data re-purpose

Data Linkage Issues:
- Persistent Identifiers (PID)
- ORCID (Open Researcher and Contributor ID)
- Lined Data

Data Clean Up and Retirement:
- Ownership and authority
- Data Detainment

Open Public Use

Data Links

Metadata & Mgmt
Further Research

• Data centric models
• DataBus concept and related data centric mechanisms
• Data centric security and NoSQL security
• Big Data curriculum development and coordination
Foreseen Big Data Innovations in 2013+

- **Cloud-Based Big Data Solutions**
  - Amazon’s Elastic Map Reduce (EMR) is a market leader
  - Expected new innovative Big Data and Cloud solutions
- **Real-Time Hadoop**
  - Google’s Dremel-like solutions that will allow real-time queries on Big Data and be open source
- **Distributed Machine Learning**
  - Mahout iterative scalable distributed back propagation machine learning and data mining algorithm
  - New algorithms Jubatus, HogWild
- **Big Data Appliances (also for home)**
  - Raspberry Pi and home-made GPU clusters
  - Hardware vendors (Dell, HP, etc.) pack mobile ARM processors into server boxes
  - Adepteva’s Parallella will put a 16-core supercomputer into range of $99
- **Easier Big Data Tools**
  - Open Source and easy to use drag-and-drop tools for Big Data Analytics to facilitate the BD adoption
  - Commercial examples: Radoop = RapidMiner + Mahout, Tableau, Datameer, etc.
  - *LexisNexis*: from ECL (Enterprise Control Language) to KEL (Knowledge Engineering Language)

Source: Big Data in 2013 by Mike Guattieri, Forrester
Evolutional/Hierarchical Data Model

• Common Data Model?
• Data interlinking?
• Fits to Graph data type?
• Metadata

• Referrals
• Control information
• Policy
• Data patterns

Usable Data

Actionable Data
Papers/Reports
Archival Data

Processed Data (for target use)

Processed Data (for target use)

Classified/Structured Data

Classified/Structured Data

Classified/Structured Data

Raw Data

Processed Data (for target use)
Security issues

- CIA and Access control
- Referral integrity
- Traceability
- Opacity