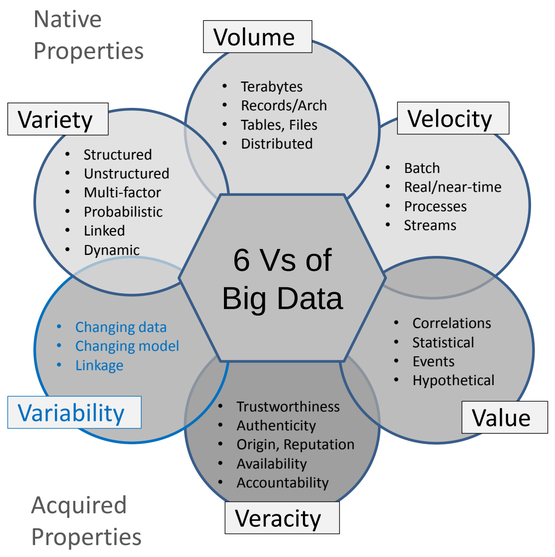
Towards Defining Big Data Architecture Framework

Yuri Demchenko, Marcel Worring, Wouter Los, Cees de Laat

Big Data Definition: From 5+1 Vs to 5 Parts Definition



(1) Big Data Properties: 5+1 V

- Native properties: Volume, Variety, Velocity
- Acquired properties: Value, Veracity, Variability (Dynamicity)

(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) 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

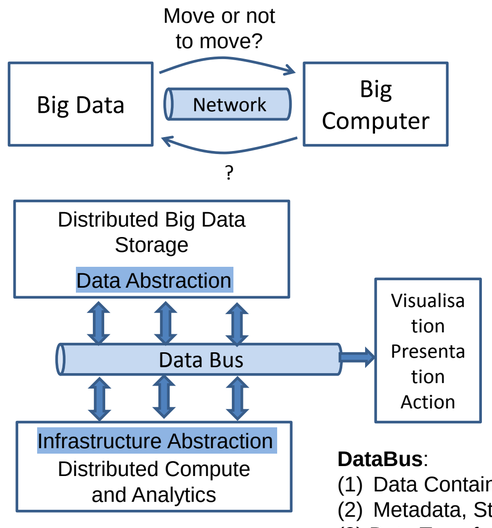
(5) New Infrastructure and Tools

- 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

Refining Gartner definition (http://www.gartner.com/it-glossary/big-data)

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 (3) cost-effective, innovative forms of data and information processing (analytics) for (4) enhanced insight, decision making, and processes control; all of those demand (should be supported by) (2) new data models (supporting all data states and stages during the whole data lifecycle) and (5) new infrastructure services and tools that allows also obtaining (and processing data) from (4) a variety of sources (including sensor networks) and delivering data in a variety of forms to different data and information consumers and devices.

Big Data Paradigm Change: Moving to Data-Centric Models



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

Current IT and communication technologies are host based or host centric (service/message centric)

- Any communication or processing are bound to host/computer that runs software
- For security: all security models are host/client based

Big Data requires new data-centric models

- Data location, replication, search, access
- Data lifecycle, transformation, variability
- Data integrity, identification, ownership
- Data centric security and access control

Paradigm changing factors

- **Big Data properties: 5+1 V's**
- **Data aggregation:** multi-domain, multi-format, variability, linkage, referral integrity
- **Policy granularity:** variety and complex structure, for their access control processing
- **Virtualization:** Can improve security of data processing environment but cannot solve data security "in rest"
- **Mobility** of the different components of the typical data infrastructure: data, sensors or data source, data consumer

Big Data Architecture Framework (BDAF) Components

(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 Analytics Applications
- Target use, presentation, visualisation

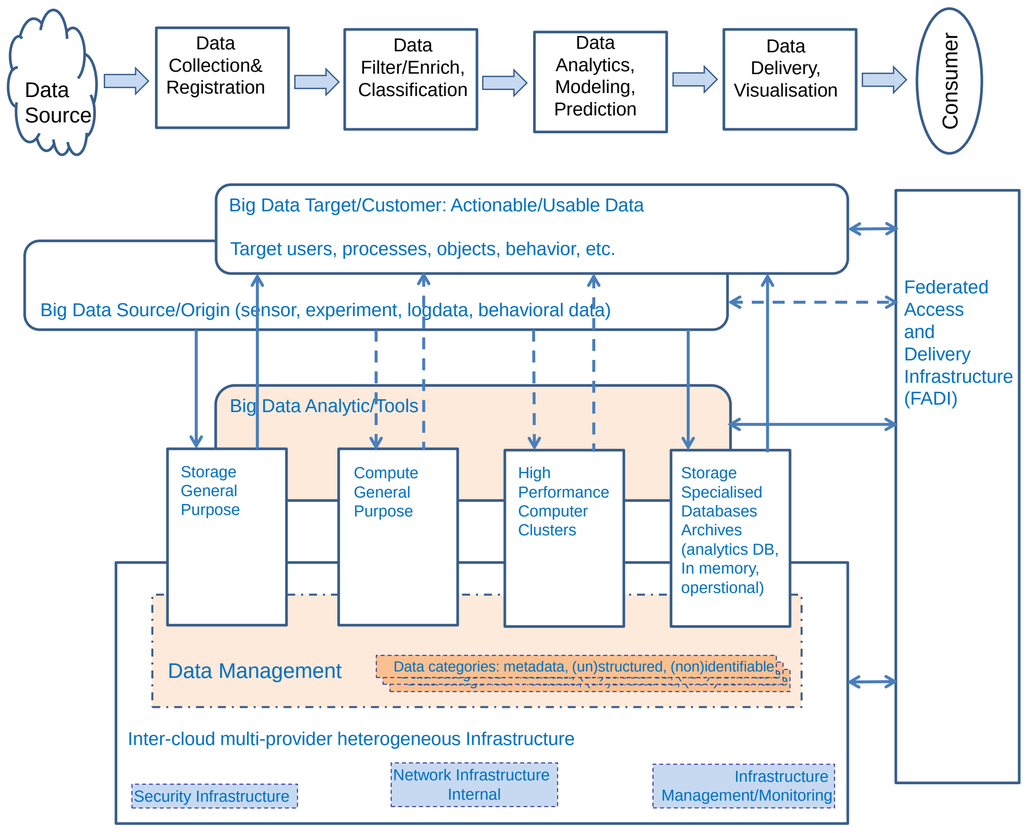
(4) Big Data Infrastructure (BDI)

- 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

Big Data Infrastructure (BDI)



Big Data Infrastructure (BDI) Components

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)
- Security infrastructure (access control, Identity and policy management, confidentiality, privacy, trust)
- Federated Access and Delivery Infrastructure (FADI)

Big Data Analytics Infrastructure

- High Performance Computer Clusters (HPCC)
- Specialised Storage, Distributed/Replicated, Archives, Databases, SQL/NoSQL

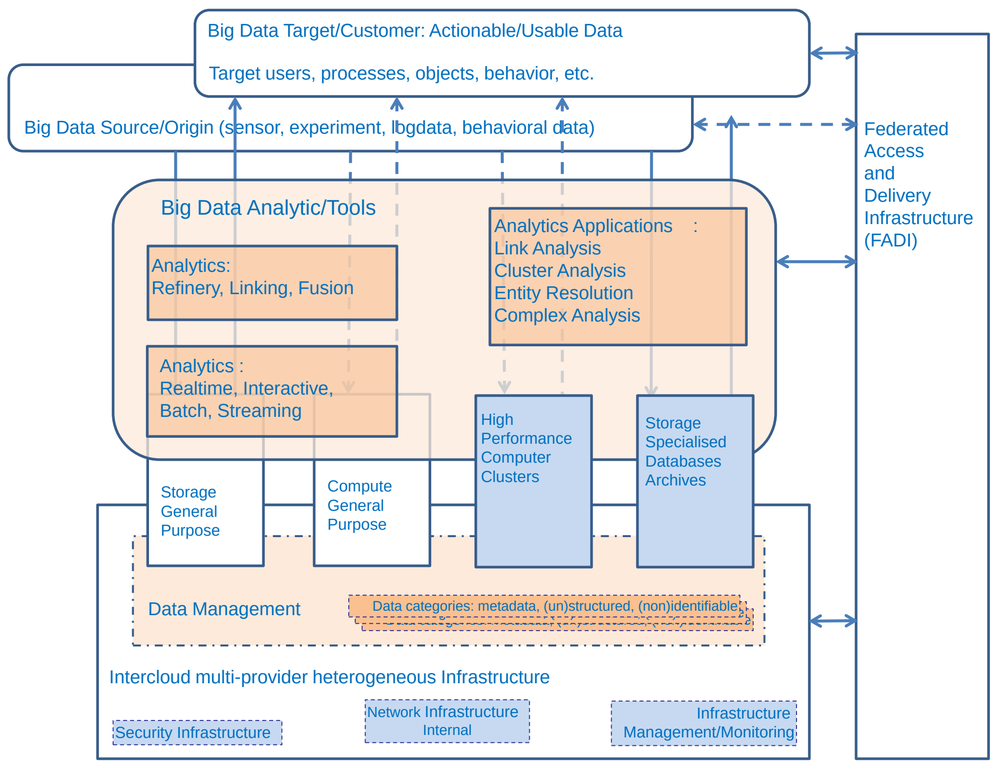
Big Data Analytics Tools/Applications

- Real-time, Interactive, Batch, Streaming
- Link Analysis, Graph analysis
- Cluster Analysis
- Entity Resolution
- Complex Analysis

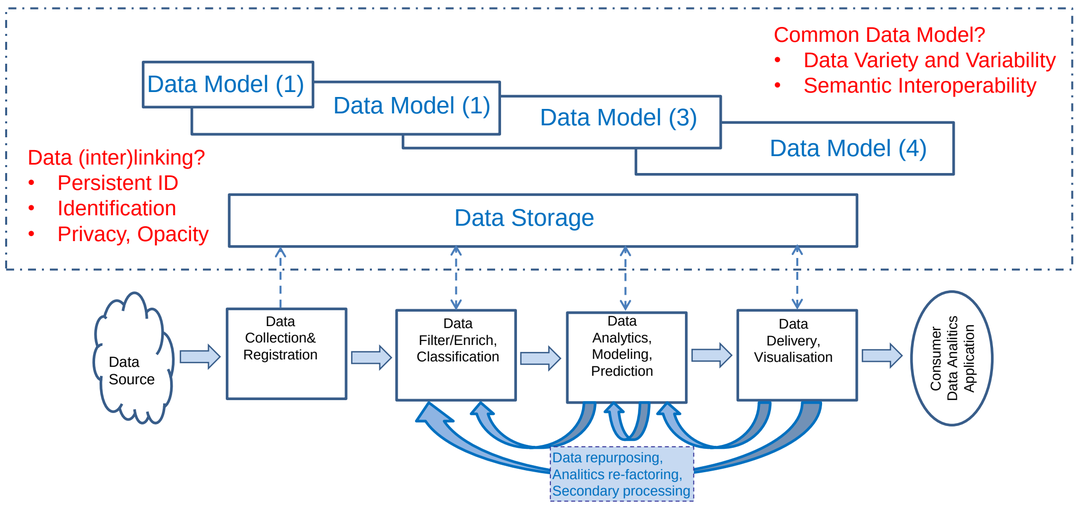
Big Data Source and Target

- Scientific Instruments, Sensor network, Experiments, Technological processes
- Logdata, web/online activity, social networks
- Human activity and input (crowdsourcing)
- Actionable data, reporting, visualisation

Big Data Analytics Infrastructure



Big Data Lifecycle Management (BDLM) Model



Related links

[1] Riding the wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. October 2010. [online] Available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf
[2] Big Data Ecosystem: Architecture Framework and Infrastructure Components. SNE Technical Report SNE-2013-02, 12 September 2013. [online] http://www.uazone.org/demch/worksinprogress/sne-2013-02-techreport-bdaf-draft02.pdf
[3] Demchenko, Y., P.Membrey, P.Grosso, C. de Laat, Addressing Big Data Issues in Scientific Data Infrastructure. First International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2013). Part of The 2013 Int. Conf. on Collaboration Technologies and Systems (CTS 2013), May 20-24, 2013, San Diego, California, USA.
[4] NIST Big Data Working Group (NBD-WG). [online] http://bigdatawg.nist.gov/home.php

Contributing Projects

GEANT3plus JRA1 Task 2 – Network Architectures for Cloud Services - http://www.geant.net/
COMMIT Project - http://www.commit-nl.nl/

Credits: Yuri Demchenko, Marcel Worring, Wouter Los, Cees de Laat
Contact: Yuri Demchenko <y.demchenko@uva.nl>