Cloud and Big Data Standardisation

EuroCloud Symposium
ICS Track: Standards for Big Data in the Cloud
15 October 2013, Luxembourg

Yuri Demchenko
System and Network Engineering Group, University of Amsterdam

Outline

- Standardisation on Big Data – Overview
- Research Data Alliance (RDA) and related initiatives PID and ORCID
- Overview NIST Big Data Working Group (NBD-WG) activities and deliverables
- Conceptual approach: Big Data Architecture Framework (BDAF) by UvA

Big Data Standardisation Initiatives

- First attempts by industry associations: ODCA, TMF
- Big Data and Data Analytics architectures
  - By the major providers IBM, LexisNexis
  - By the major Cloud Service Providers: AWS Big Data Services, Microsoft Azure HDInsight, LexisNexis HPCC Systems
- Research Data Alliance (RDA)
  - Valuable work on Data Models, Metadata Registries, Trusted Registries and Metadata
- Research community initiatives
  - PID (Persistent Identifier)
  - ORCID (Open Researcher and Contributor ID)
- NIST Big Data Working Group (NBD-WG)
  - Big Data Reference Architecture
  - Big Data technology roadmap

Standardisation goals

- Common vocabulary
- Capabilities
- Stakeholders and actors
- Technology Roadmap

Research Data Alliance – First Steps

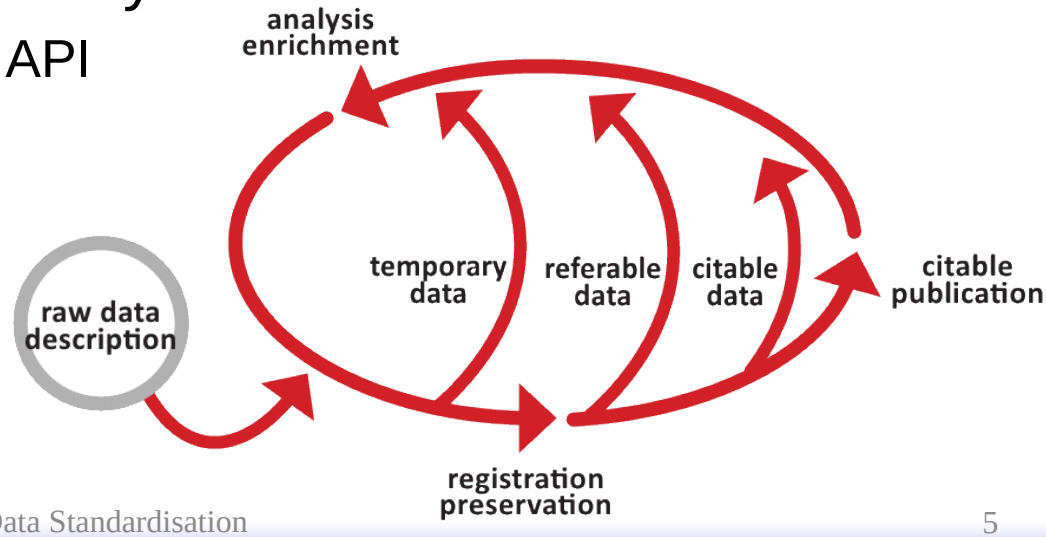
http://www.rd-alliance.org/

- Joint initiative EC, NSF, NIST: launched October 2012
  - RDA1 – March 2013 (Gothenburg), RDA2 – Sept 2013 (Washington), RDA3 – March 2014 (Dublin), RDA4 – Sept 2014 (Amsterdam)
  - Positioned as community forum and not standardisation body (currently)

- Working Groups created
  - *Data Foundation and Terminology*
  - *Harmonization and Use of PID Information Types*
  - *Data Type Registries*
  - *Metadata*
  - *Practical Policy (based on iRODS community practice)*
  - UPC (Universal Product Code) Code for Data
  - Publication/Data Citation/Linking
  - Repository Audit and Certification, Legal Interoperability
  - Big Data Analytics (evaluation and study)
  - Data Intensive Science Education and Skills development
  - Number of application domains

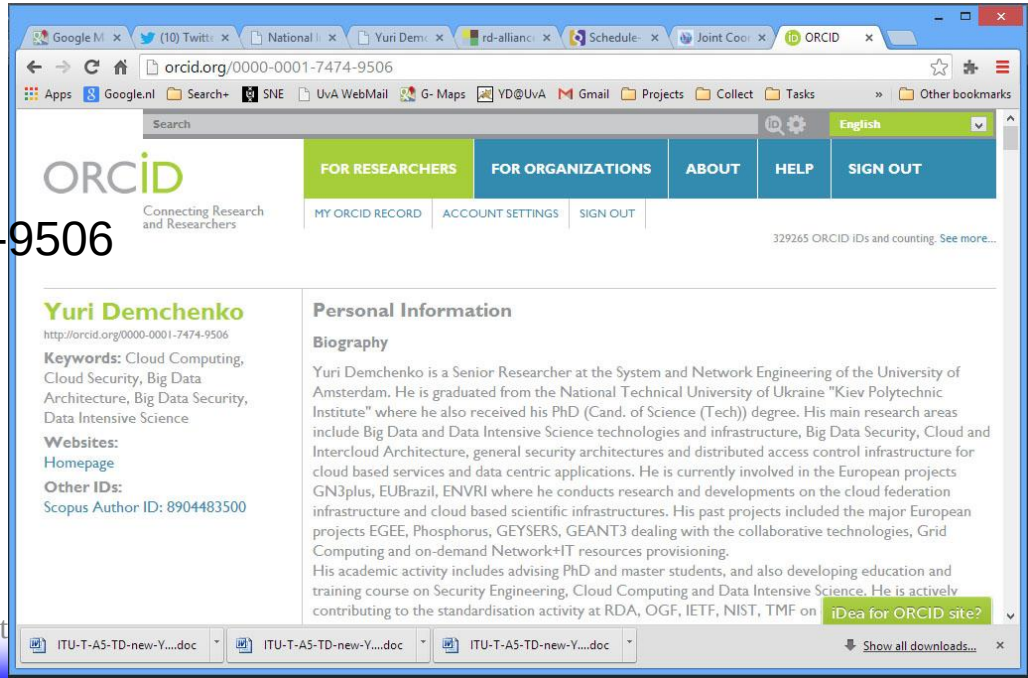
Persistent Identifier (PID)

- PID – Persistent Identifier for Digital Objects
  - Managed by European PID Consortium (EPIC) http://www.pidconsortium.eu/
  - Superset of DOI - Digital Object Identifier (http://www.doi.org/)
  - Handle System by CNRI (Corporation for National Research Initiatives) for resolving DOI (http://www.handle.net/)
- PID provides a mechanism to link data during the whole research data transformation cycle
  - EPIC RESTful Web Service API published May 2013



ORCID (Open Researcher and Contributor ID)

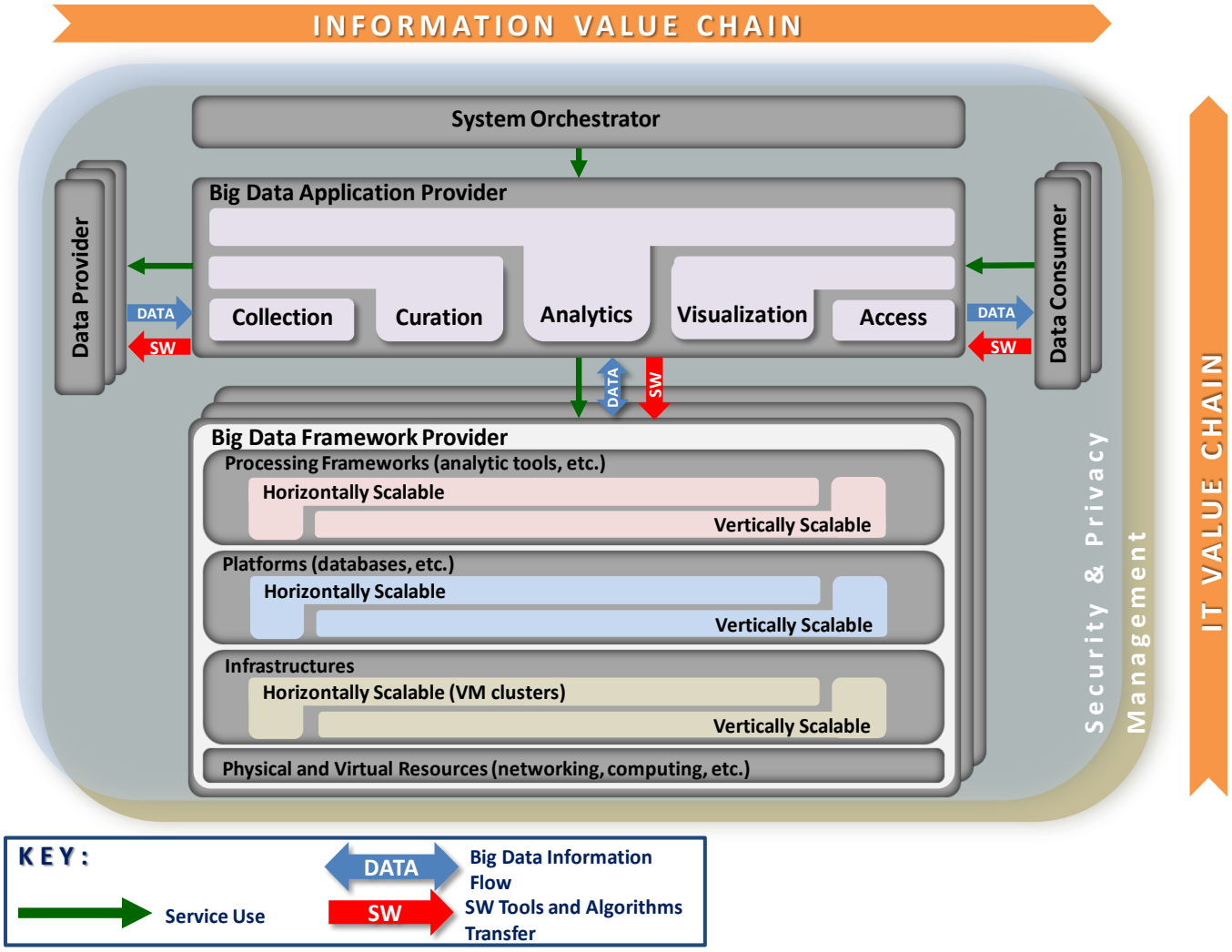
- ORCID is a nonproprietary alphanumeric code to uniquely identify scientific and other academic authors
  - Launched October 2012
- ORCID Statistics – October 2013
  - Live ORCID iDs 329,265
  - ORCID iDs with at least one work 79,332
  - Works 2,205,971
  - Works with unique DOIs 1,267,083
- Personal ORCID
  - ORCID 0000-0001-7474-9506
  - http://orcid.org/0000-0001-7474-9506
  - Scopus Author ID 8904483500



NIST Big Data Working Group (NBD-WG)

- First deliverables target – September 2013
  - 30 September – Workshop and F2F meeting
- Activities: Conference calls every day 17-19:00 (CET) by subgroup - http://bigdatawg.nist.gov/home.php
  - Big Data Definition and Taxonomies
  - Requirements (chair: Geoffrey Fox, Indiana Univ)
  - Big Data Security
  - Reference Architecture
  - Technology Roadmap
- BigdataWG mailing list and useful documents
  - Input documents http://bigdatawg.nist.gov/show_InputDoc2.php
  - Big Data Reference Architecture http://bigdatawg.nist.gov/_uploadfiles/M0226_v2_1885676266.docx
  - Requirements for 21 use cases http://bigdatawg.nist.gov/_uploadfiles/M0224_v1_1076079077.xlsx
- Prospective ISO Big Data Study Committee to be started

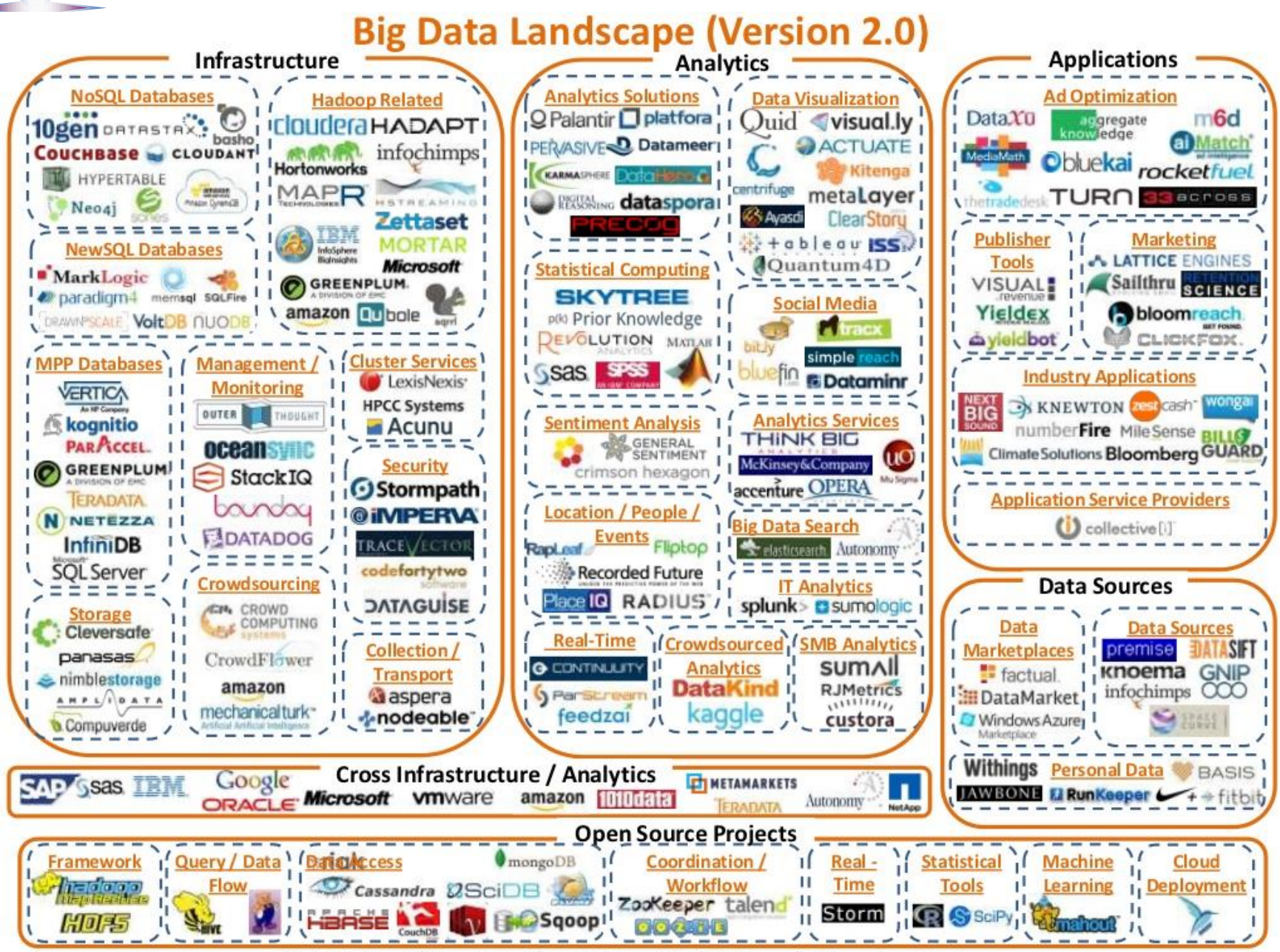
NIST Big Data Reference Architecture – Draft version 0.7, 26 Sept 2013



Main Component

- Data Provider
- Big Data Application Provider
- Big Data Framework Provider
- Data Consumer
- System Orchestrator

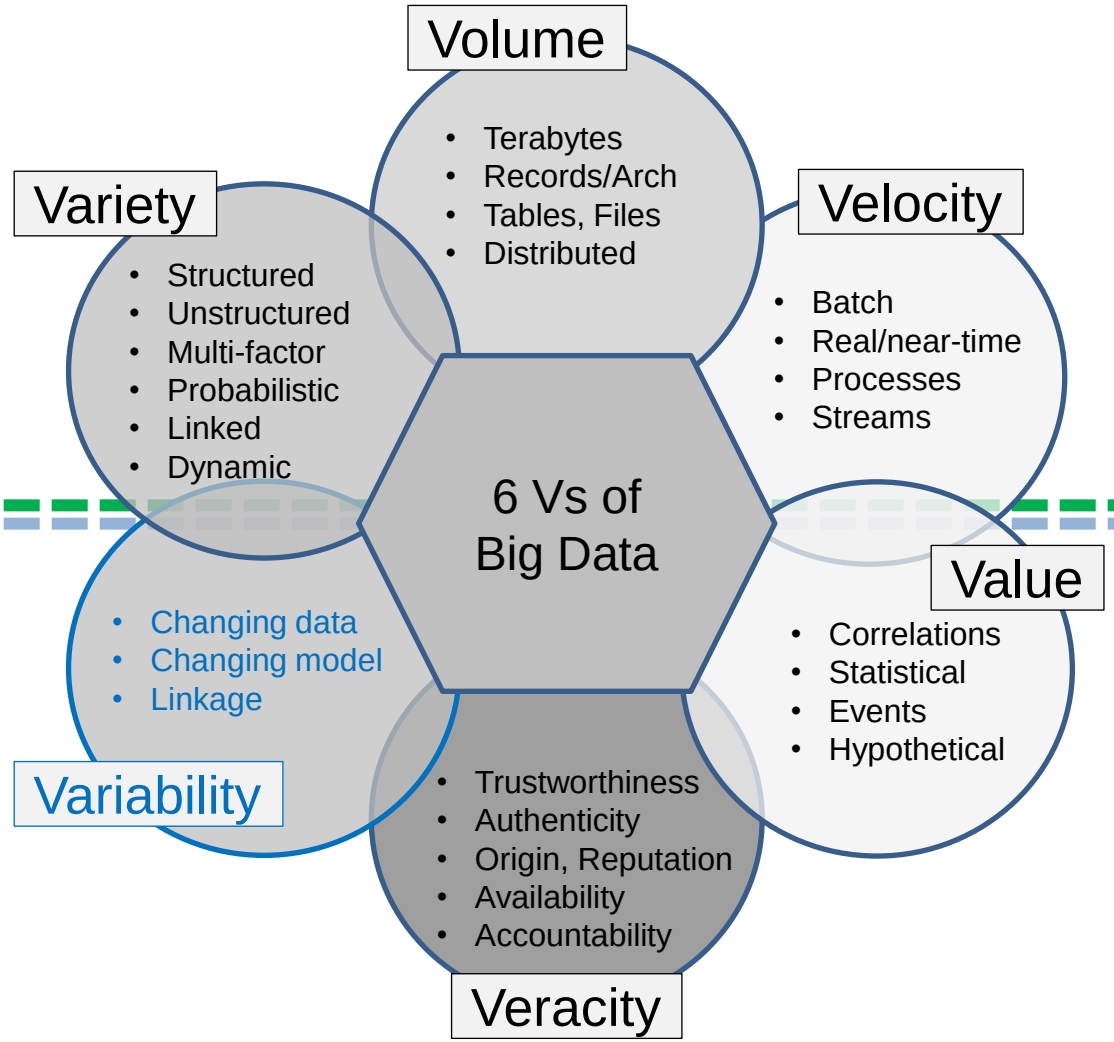
http://mattturck.com/2012/10/15/a-chart-of-the-big-data-ecosystem-take-2/ [Dec 2012]



Conceptual approach: Big Data Architecture Framework (BDAF) by UvA

- Big Data definition: From 5+1Vs to 5 parts
- Big Data Architecture Framework (BDAF) components
- Data Lifecycle Management model

Improved: 5+1 V's of Big Data



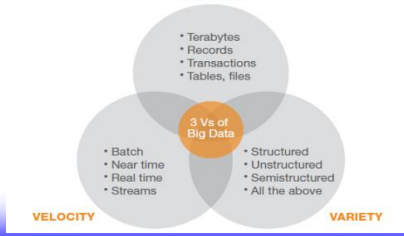
Generic Big Data Properties

- Volume
- Variety
- Velocity

Acquired Properties (after entering system)

- Value
- Veracity
- Variability

Commonly accepted 3V's of Big Data



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

Big Data Definition: From 5V to 5 Parts (2)

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 (3) 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) (2) new data models (supporting all data states and stages during the whole data lifecycle) and (4) new infrastructure services and tools that allows also obtaining (and processing data) from (5) 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

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 (Stack)
  - Separates concerns and factors
    - Control and Management functions, orthogonal factors
  - Architecture Framework components are inter-related

- Big Data Architecture Framework (BDAF) by UvA
  Architecture Framework and Components for the Big Data Ecosystem. SNE Technical Report 2013-02, Version 0.2, 12 September
  http://www.uazone.org/demch/worksinprogress/sne-2013-02-techreport-bdaf-draft02.pdf

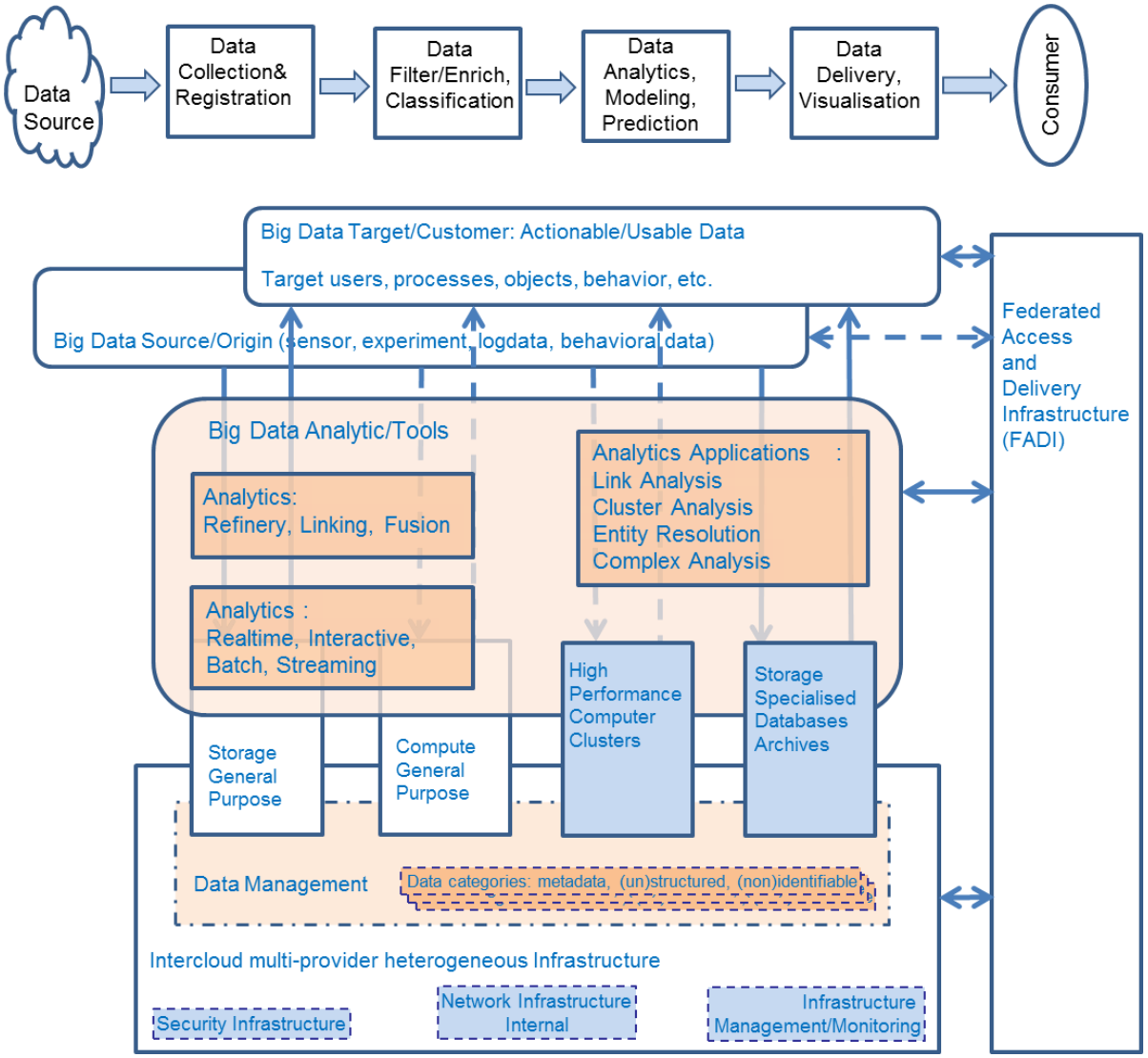
Big Data Architecture Framework (BDAF) for Big Data Ecosystem (BDE)

- (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

Big Data Architecture Framework (BDAF) – Aggregated – Relations between components (2)

Col: Used By Row: Requires This	Data Models Structrs	Data Managmnt & Lifecycle	BigData Infrastr & Operations	BigData Analytics & Applicatn	BigData Security
Data Models & Structures		+	++	+	++
Data Managmnt & Lifecycle	++		++	++	++
BigData Infrastruct & Operations	+++	+++		++	+++
BigData Analytics & Applications	++	+	++		++
BigData Security	+++	+++	+++	+	

Big Data Infrastructure and Analytics Tools



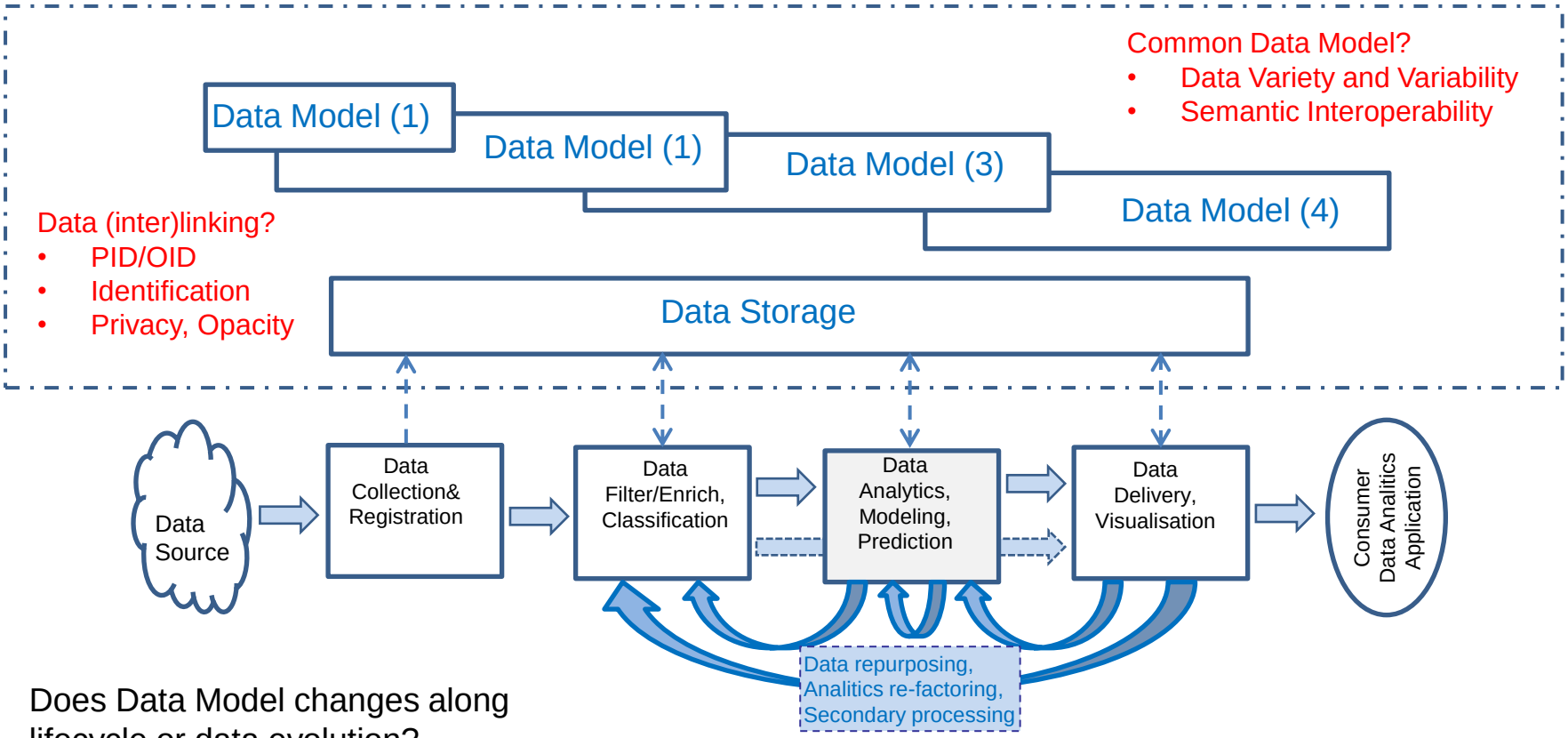
Big Data Infrastructure

- Heterogeneous multi-provider inter-cloud infrastructure
- Data management infrastructure
- Collaborative Environment (user/groups managements)
- Advanced high performance (programmable) network
- Security infrastructure

Big Data Analytics

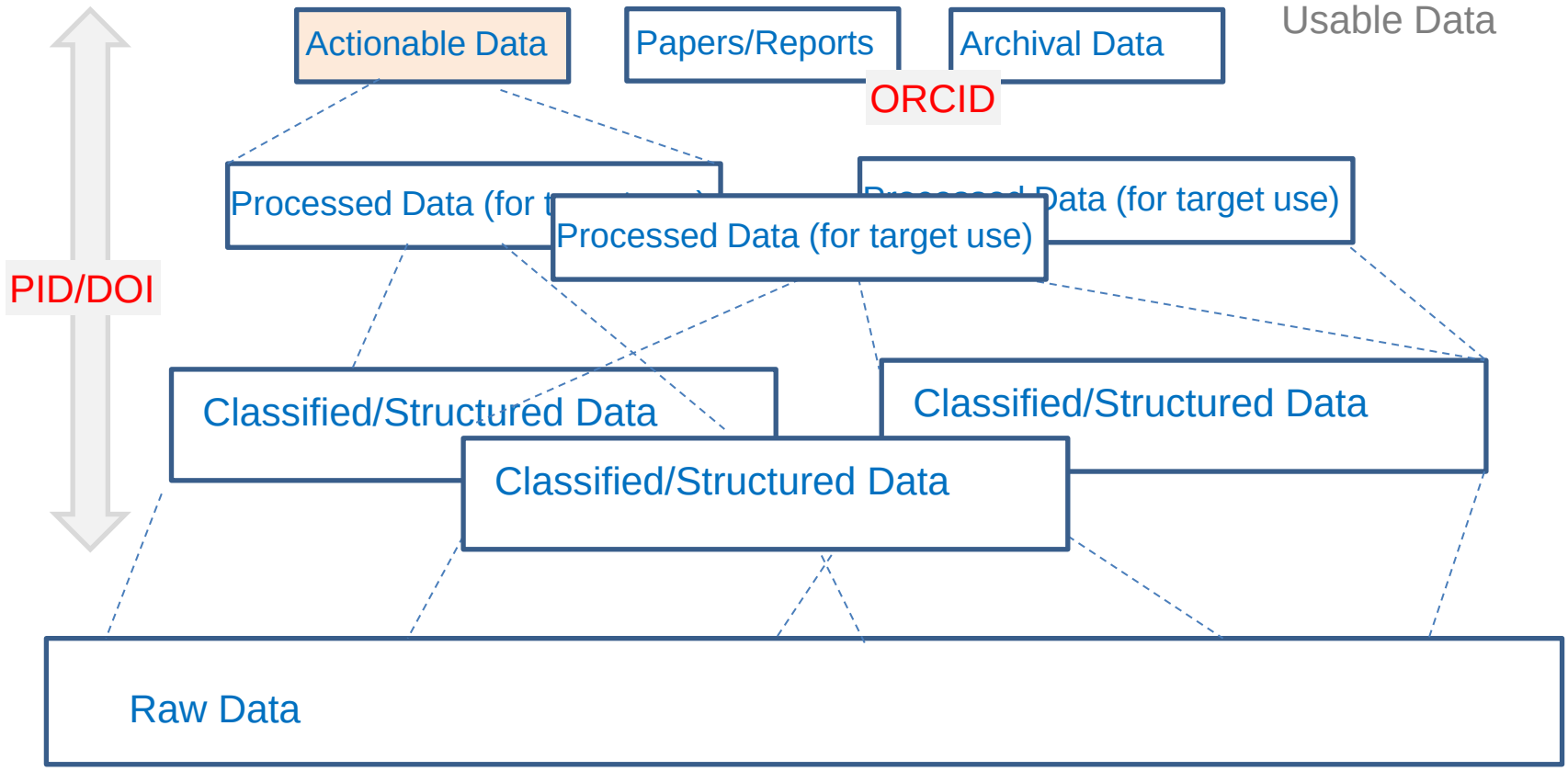
- High Performance Computer Clusters (HPCC)
- Analytics/processing: Real-time, Interactive, Batch, Streaming
- Big Data Analytics tools and applications

Data Transformation/Lifecycle Model



- Does Data Model changes along lifecycle or data evolution?
- Identifying and linking data
  - Persistent identifier
  - Traceability vs Opacity
  - Referral integrity

Evolutional/Hierarchical Data Model



Topics for discussion, research and standardisation

- Common Data Model?
- Data interlinking?
- Fits to Graph data type?
- Metadata
- Referrals
- Control information
- Policy
- Data patterns