MATES ED2MIT

Education and Training for Data Driven Maritime Industry

Tutorial A01.01

Big Data Technologies: Concepts and Algorithms

Yuri Demchenko MATES Project
University of Amsterdam

Maritime Alliance for fostering the European Blue economy through a Marine Technology Skilling Strategy

Co-funded by the
Erasmus+ Programme
of the European Union

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Outline

- Big Data definition and technology domain
  - 6V of Big Data
- Big Data use cases
- Big Data Reference Architecture
  - Organisational roles
- Data Lifecycle and data management
- Discussion

This work is licensed under the Creative Commons Attribution 4.0 International License.
To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

Co-funded by the
Erasmus+ Programme
of the European Union

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Multiple aspects of Big Data

Big Data is a complex of technologies to enable handling of Big Data (storage, processing, transfer, security)

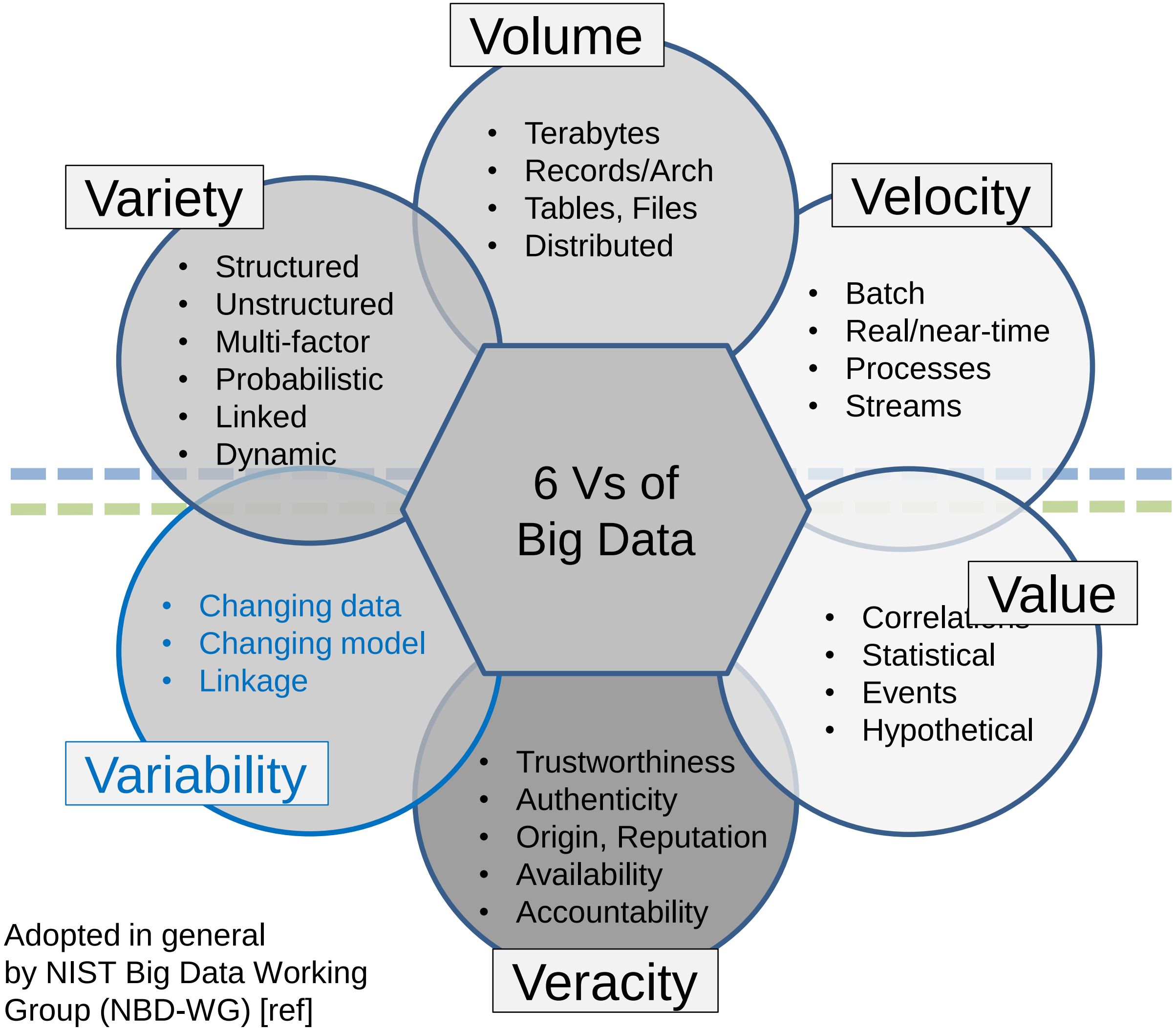
Big Data and multiple sources of data

- Social Media
- IoT
- Internet

- Science
- Industrial data
- Communication, voice

Data analytics blending with open and social media data

Big Data Properties: 6 (3+3) V's of Big Data



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

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

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

6 Vs of Big Data

Value
- Correlations
- Statistical
- Events
- Hypothetical

Variability
- Changing data
- Changing model
- Linkage

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

Adopted in general by NIST Big Data Working Group (NBD-WG) [ref]

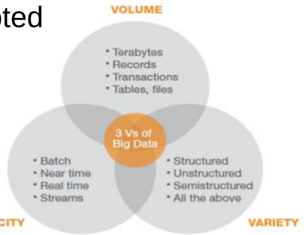
Generic Big Data Properties

- Volume
- Variety
- Velocity

Acquired Properties (after entering system)

- Value
- Veracity
- Variability

Commonly accepted 3V's of Big Data by Gartner [ref]



Big Data Definition: From 6V to 5 Parts (1)

(1) Big Data Properties: 6V

- Volume, Variety, Velocity, Value, Veracity, Variability

(2) New Data Models

- SQL and NoSQL
- Data Lifecycle management: Data linking, provenance and referral integrity

(3) New Analytics

- Real-time/streaming analytics, interactive and machine learning analytics
- Domain specific data analytics methods (e.g. bioinformatics, UX/user experience)

(4) New Infrastructure and Tools

- High performance Computing, Storage, Network – cloud based
- 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 data sources and sensors/IoT
- Data delivery to different visualisation and actionable systems and consumers
- Fully digitised input and output, (ubiquitous) sensor networks, full digital control

Big Data Nature: Origin and consumers (target)

Big Data Origin

- **Science, bioinformatics**
- Internet, Web
- Industry
- Business
- Living Environment, Smart Cities
- Social media and networks
- Healthcare
- **Telecom/Infrastructure**

Data Transformation

Big Data Target Use

- **Scientific discovery**
- New technologies
- Manufacturing, processes, transport
- Living environment support
- Healthcare support
- Personal services, campaigns, media
- Social Networks
- **Intelligence**

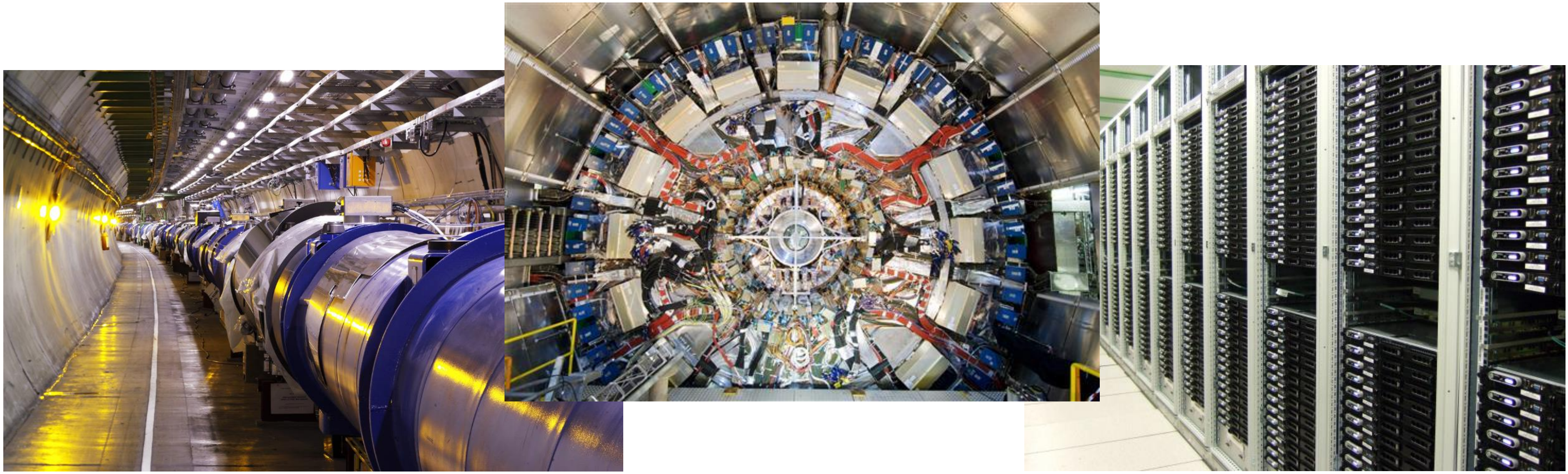
Volume, Velocity, Variety – Examples Science

- Volume – Terabyte records, transactions, tables, files.
  - LHC (Large Hadron Collider)
    - 5 PB a month (now is under re-construction to increase beam energy)
  - LOFAR (Low Frequency Array), SKA (Square Kilometer Array)
    - 5 PB every hour, requires processing asap to discard non-informative data
  - Large Synoptic Survey Telescope (LSST)
    - 10 Petabytes per year of the ***complex interlinked hierarchical data***
  - Genomic research – x10 TB per individual
  - Earth, climate and weather data
- Velocity – batch, near-time, real-time, streams.
  - LHC ATLAS detector generates about 1 Petabyte raw data per second, during the collision time about 1 ms
- Variety – structures, unstructured, semi-structured, and all the above in a mix
  - Biodiversity, Biological and medical, facial research
  - Human, psychology and behavior research
  - History, archeology and artifacts

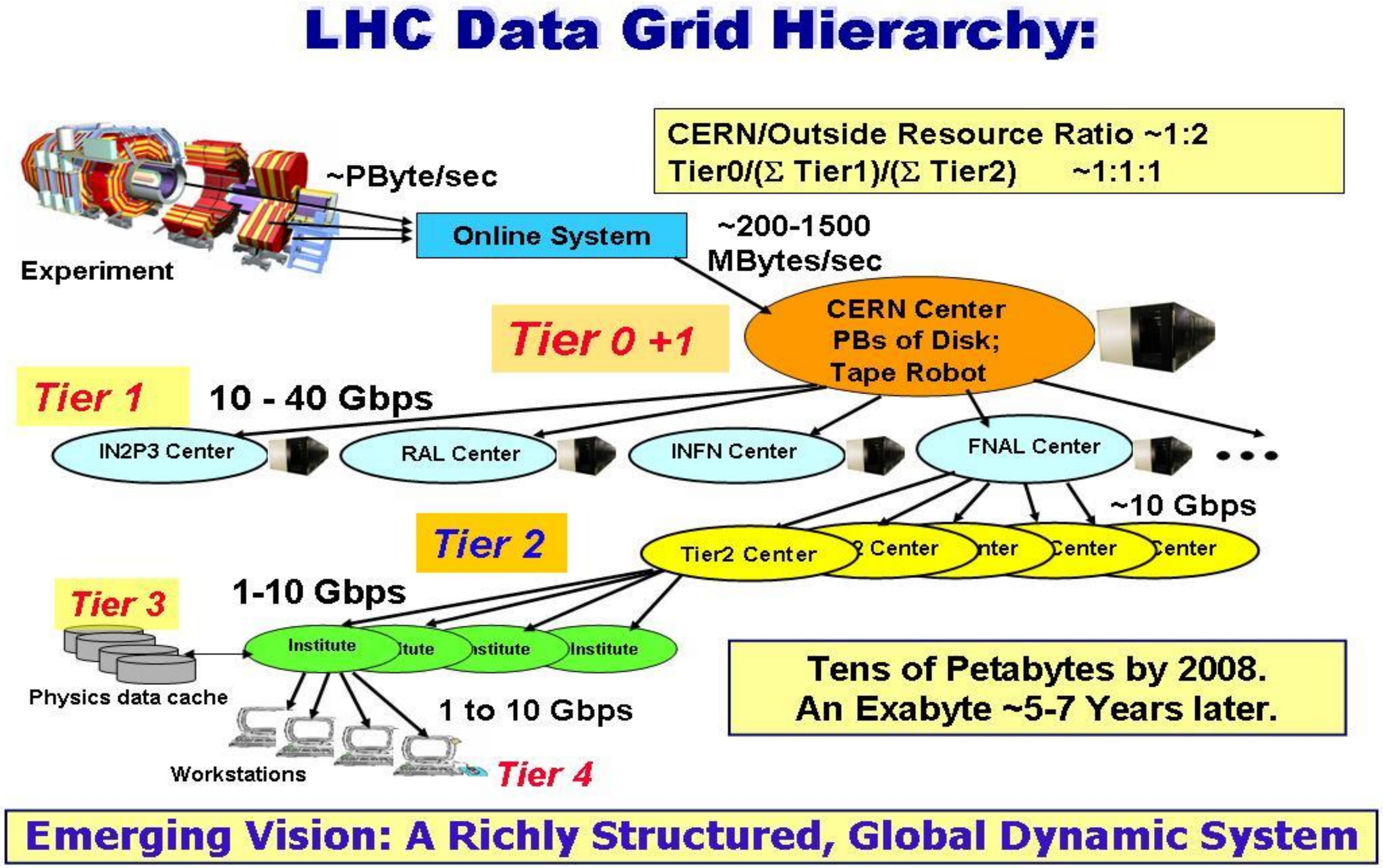
LHC and Atlas: Volume and Velocity

ALICE
CMS
ATLAS
LHCb

- The LHC Collider location at CERN (CERN-LHC, 2014).
- LHC contains 2 accelerator rings built in the tunnels 100 m underground: SPS (Super Proton Synchrotron) with the diameter 2 km and LHC having circumference 27 km
- Atlas is a collision detector



LHC Data Grid Hierarchy



2005 – 2010

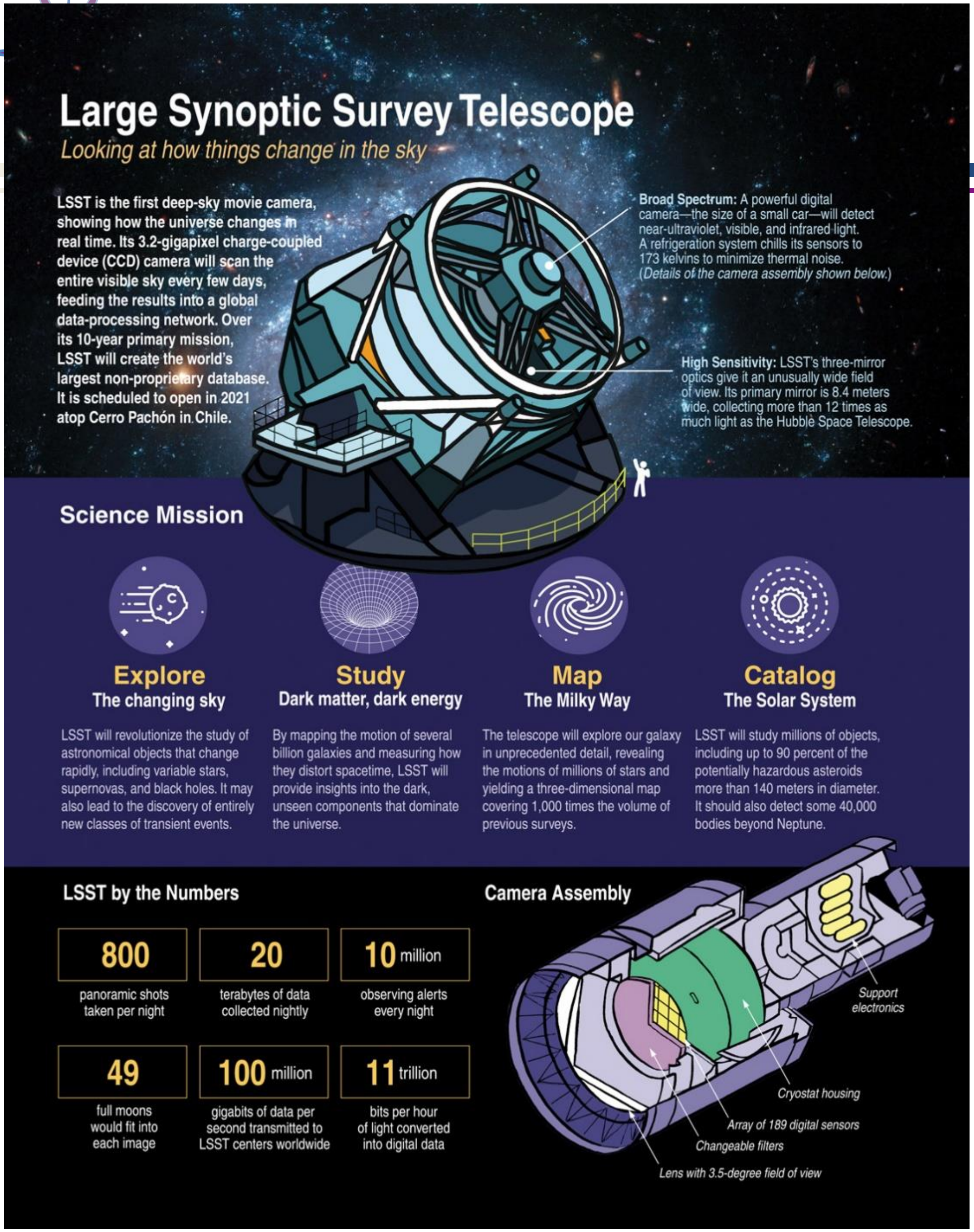
EGEE project

Operational

WLCG – Worldwide LHC Computer Grid

https://wlcg.web.cern.ch/

LSST: Volume and Variability



Large Synoptic Survey Telescope

Looking at how things change in the sky

LSST is the first deep-sky movie camera, showing how the universe changes in real time. Its 3.2-gigapixel charge-coupled device (CCD) camera will scan the entire visible sky every few days, feeding the results into a global data-processing network. Over its 10-year primary mission, LSST will create the world's largest non-proprietary database. It is scheduled to open in 2021 atop Cerro Pachón in Chile.

Broad Spectrum: A powerful digital camera—the size of a small car—will detect near-ultraviolet, visible, and infrared light. A refrigeration system chills its sensors to 173 kelvins to minimize thermal noise. (*Details of the camera assembly shown below.*)

High Sensitivity: LSST's three-mirror optics give it an unusually wide field of view. Its primary mirror is 8.4 meters wide, collecting more than 12 times as much light as the Hubble Space Telescope.

Science Mission

Explore
The changing sky

LSST will revolutionize the study of astronomical objects that change rapidly, including variable stars, supernovas, and black holes. It may also lead to the discovery of entirely new classes of transient events.

Study
Dark matter, dark energy

By mapping the motion of several billion galaxies and measuring how they distort spacetime, LSST will provide insights into the dark, unseen components that dominate the universe.

Map
The Milky Way

The telescope will explore our galaxy in unprecedented detail, revealing the motions of millions of stars and yielding a three-dimensional map covering 1,000 times the volume of previous surveys.

Catalog
The Solar System

LSST will study millions of objects, including up to 90 percent of the potentially hazardous asteroids more than 140 meters in diameter. It should also detect some 40,000 bodies beyond Neptune.

LSST by the Numbers

800	panoramic shots taken per night
20	terabytes of data collected nightly
10 million	observing alerts every night
49	full moons would fit into each image
100 million	gigabits of data per second transmitted to LSST centers worldwide
11 trillion	bits per hour of light converted into digital data

Camera Assembly

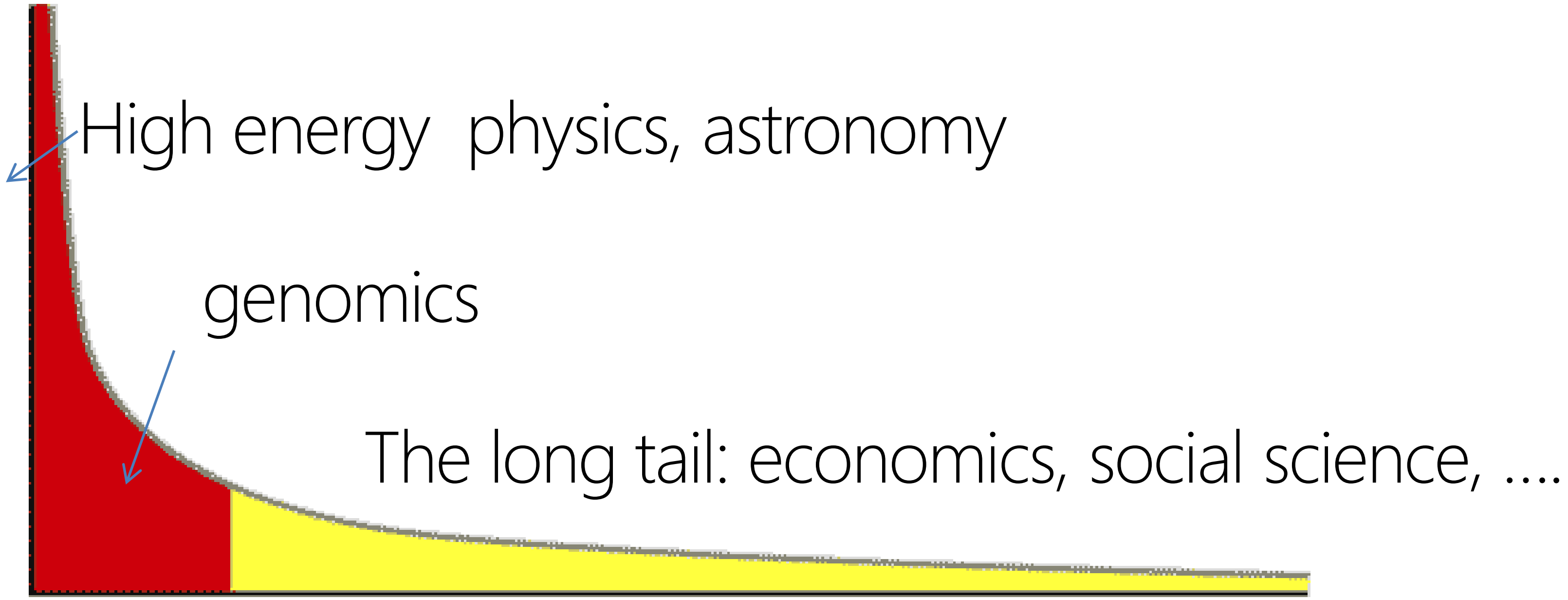
Support electronics
Cryostat housing
Array of 189 digital sensors
Changeable filters
Lens with 3.5-degree field of view

- Solar system inventory to discover and track moving objects, asteroids, Near Earth Objects (NEOs)
- Optical transients of all kinds , including alert notification within 60 seconds
- Milky Way observation including star streams, motion, estimated dark matter
- World largest camera 3.200 MPix

Length	12.25 ft (3.73 m)
Height	5.5 ft (1.65 m)
Weight	6200 lbs (2800 kg)
Pixel Count	3200 megapixel
Wavelength Range	320–1050 nm

https://www.bnl.gov/newsroom/news.php?a=216631

The Long Tail of Science (aka "Dark Data")



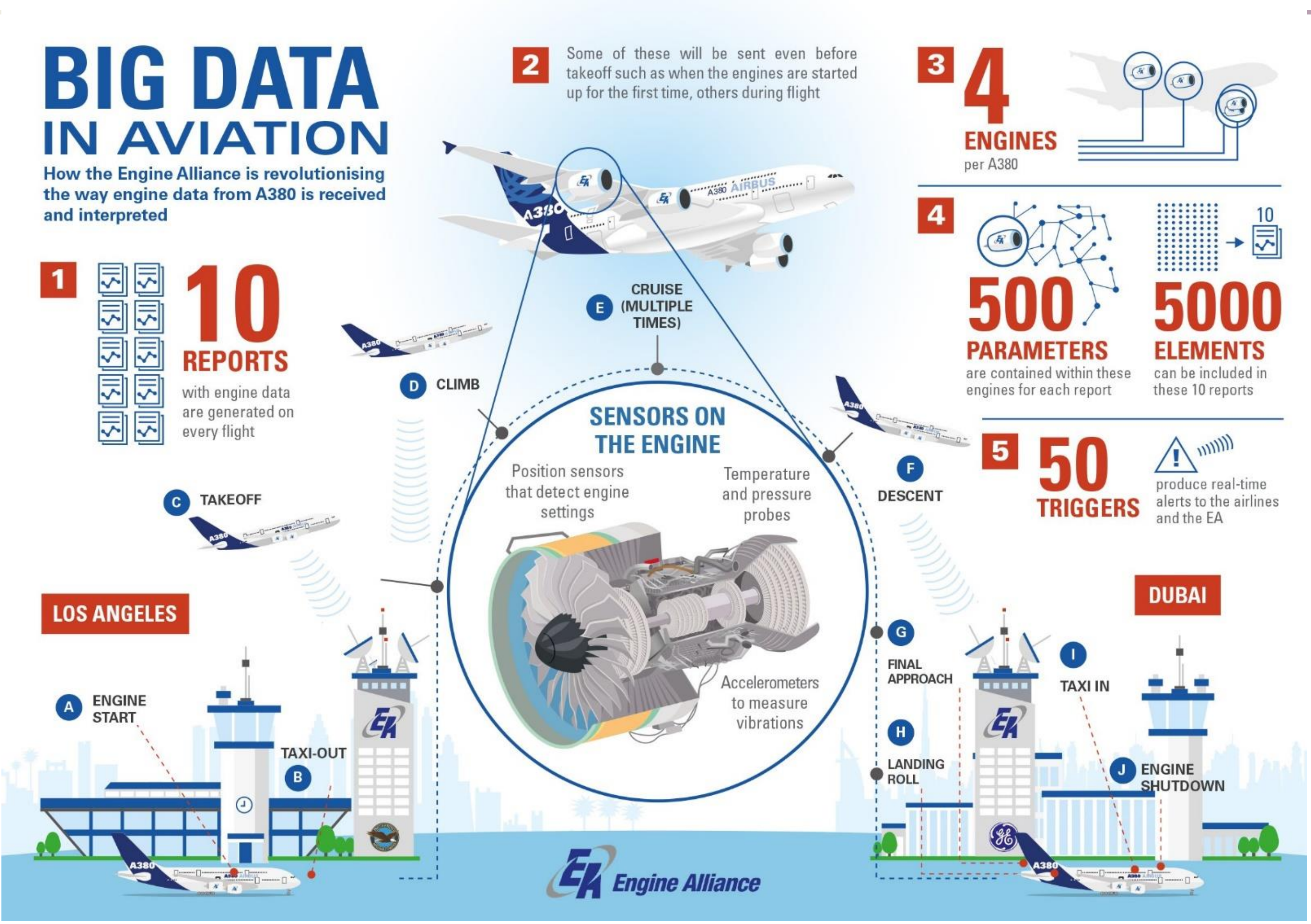
- Collectively "Long Tail" science is generating a lot of data
  - Estimated as over 1PB per year and it is growing fast with the new technology proliferation
  - Big Data and Data Science technologies development facilitates collecting more data and using Big Data analytics tools
- 80-20 rule: 20% users generate 80% data but not necessarily 80% knowledge

Source: Dennis Gannon (Microsoft)
NIST Big Data Workshop, 2012

Volume, Velocity, Variety – Examples Industry

- Volume – Terabyte records, transactions, tables, files.
  - A Boeing 4-engine Jumbo jet aircraft can create 640TB on one Atlantic crossing. Multiply that to 25,000 flights flown each day
  - Network monitoring, logging, intrusion detection
- Velocity – batch, near-time, real-time, streams.
  - Today's on-line ads serving requires *40ms to respond with a decision* what relevant to user information can be displayed on the page
  - Financial services (i.e., stock quotes feed) need near *1ms to calculate* customer scoring probabilities
  - Stream data, such as movies, need to travel at high speed for proper rendering
- Variety – structures, unstructured, semi-structured, and all the above in a mix
  - WalMart processes 1Mln customer transactions per hour and feeds information to a database estimated at 2.5PB (petabytes)
  - Old and new data sources like RFID, sensors, mobile payments, in-vehicle tracking, etc.

Example Industry; Aviation, Predictive Maintenance


BIG DATA
IN AVIATION
How the Engine Alliance is revolutionising the way engine data from A380 is received and interpreted
1
10
REPORTS
with engine data are generated on every flight
2
Some of these will be sent even before takeoff such as when the engines are started up for the first time, others during flight
3
4
ENGINES
per A380
4
500
PARAMETERS
are contained within these engines for each report
10
5000
ELEMENTS
can be included in these 10 reports
5
50
TRIGGERS
produce real-time alerts to the airlines and the EA
SENSORS ON THE ENGINE
Position sensors that detect engine settings
Temperature and pressure probes
Accelerometers to measure vibrations
A ENGINE START
B TAXI-OUT
C TAKEOFF
D CLIMB
E CRUISE (MULTIPLE TIMES)
F DESCENT
G FINAL APPROACH
H LANDING ROLL
I TAXI IN
J ENGINE SHUTDOWN
LOS ANGELES
DUBAI
A380
AIRBUS
Engine Alliance

- Flight data collection and analysis for Predictive Maintenance
  - Data Quality
- Total flight data volume **640TB** on one Atlantic crossing for a Boeing 4-engine Jumbo jet aircraft.
- Multiply that to 25,000 flights flown each day

Use case: Big Data in Aviation: Infographics by Engine Alliance

http://hub.enginealliance.com/res/images/infographics.jpg -- Volume, Variety, Value, Variability

Targeted Ads Service

- Today's on-line ads serving requires *40ms to respond with a decision* what relevant to user information can be displayed on the page
- What technology is used
  - Technological cookies (formally are not subject to GDPR)
  - Website tracking cookies
  - Google Search: aggregates your search website analytics by google (also treated technological cookies)
  - Webshop items viewing, bank transactions
- And still timely ads placing is critical

Big Data technology drivers - Examples

- Modern e-Science in search for new knowledge
  - Scientific experiments and tools are becoming bigger and heavily based on data processing and mining
- Traditional data intensive industry
  - Genomic research, drugs development, Healthcare
  - High-tech industry, CAD/CAM, weather/climate, etc.
- AI, IoT and Industry 4.0
  - Data and Analytics are in foundation
- Network/infrastructure management
  - Network monitoring, Intrusion detection, troubleshooting
- Intelligence and security
- Consumer facing companies like Google and Facebook have driven many of the recent advances in Big Data efficiency
  - Facebook has some 1.74+ Billion users and is still growing
  - Google handles number of search queries at 3 billion per day
  - Twitter handles some 400 million tweets per day count for 12 terabytes per day
    - Twitter data are widely used to add sentiments to market analysis and prediction
  - Power companies: process up to 350 billion annual meter readings to better predict power consumption
- Individually targeted online advertisement and campaigns

NIST Big Data Working Group (NBD-WG) and ISO/IEC JTC1 Study Group on Big Data (SGBD)

- NIST Big Data Working Group (NBD-WG) is leading the development of the Big Data Technology Roadmap - http://bigdatawg.nist.gov/home.php
  - Built on experience of developing the Cloud Computing standards
- Published as NIST Special Publication 1500 Volumes 1-7 in 2015
- New revision V3 published 2020 - https://bigdatawg.nist.gov/V3_output_docs.php

  Volume 1: Definitions

  Volume 2: Taxonomies

  Volume 3: Use Case & Requirements

  Volume 4: Security & Privacy

  Volume 5: Reference Architecture White Paper

  Volume 6: Reference Architecture

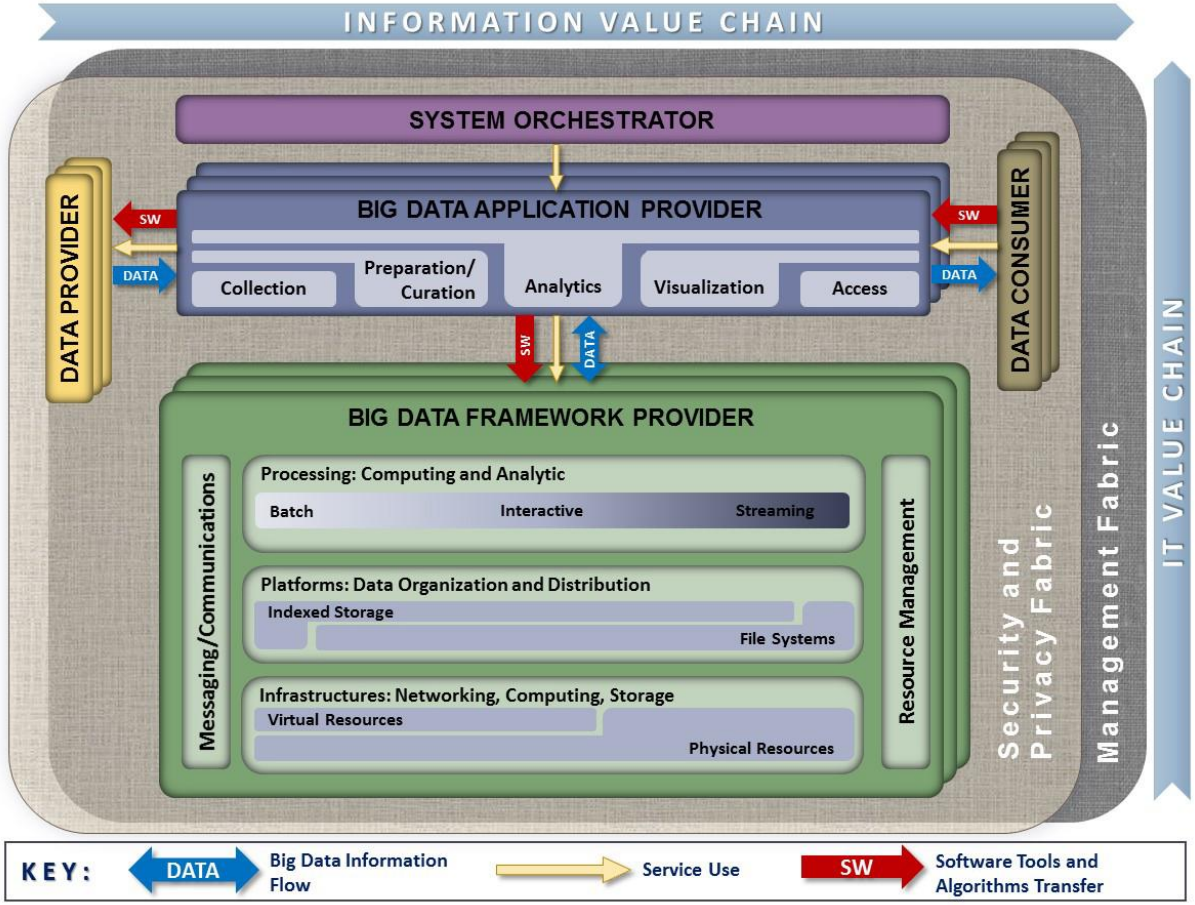
  Volume 7: Standards Roadmap

  Volume 9: Reference Architecture Interface

  Volume 10: Adoption and Modernization
- NBD-WG defined 3 main components of the new technology:
  - Big Data Paradigm
  - Big Data Science and Data Scientist as a new profession
  - Big Data Architecture

The **Big Data Paradigm** consists of the distribution of data systems across horizontally-coupled independent resources to achieve the scalability needed for the efficient processing of extensive datasets.

NIST Big Data Reference Architecture (2020)



Main components of the Big Data ecosystem

- Data Provider
- Big Data Applications Provider
- Big Data Framework Provider
- Data Consumer
- Service Orchestrator

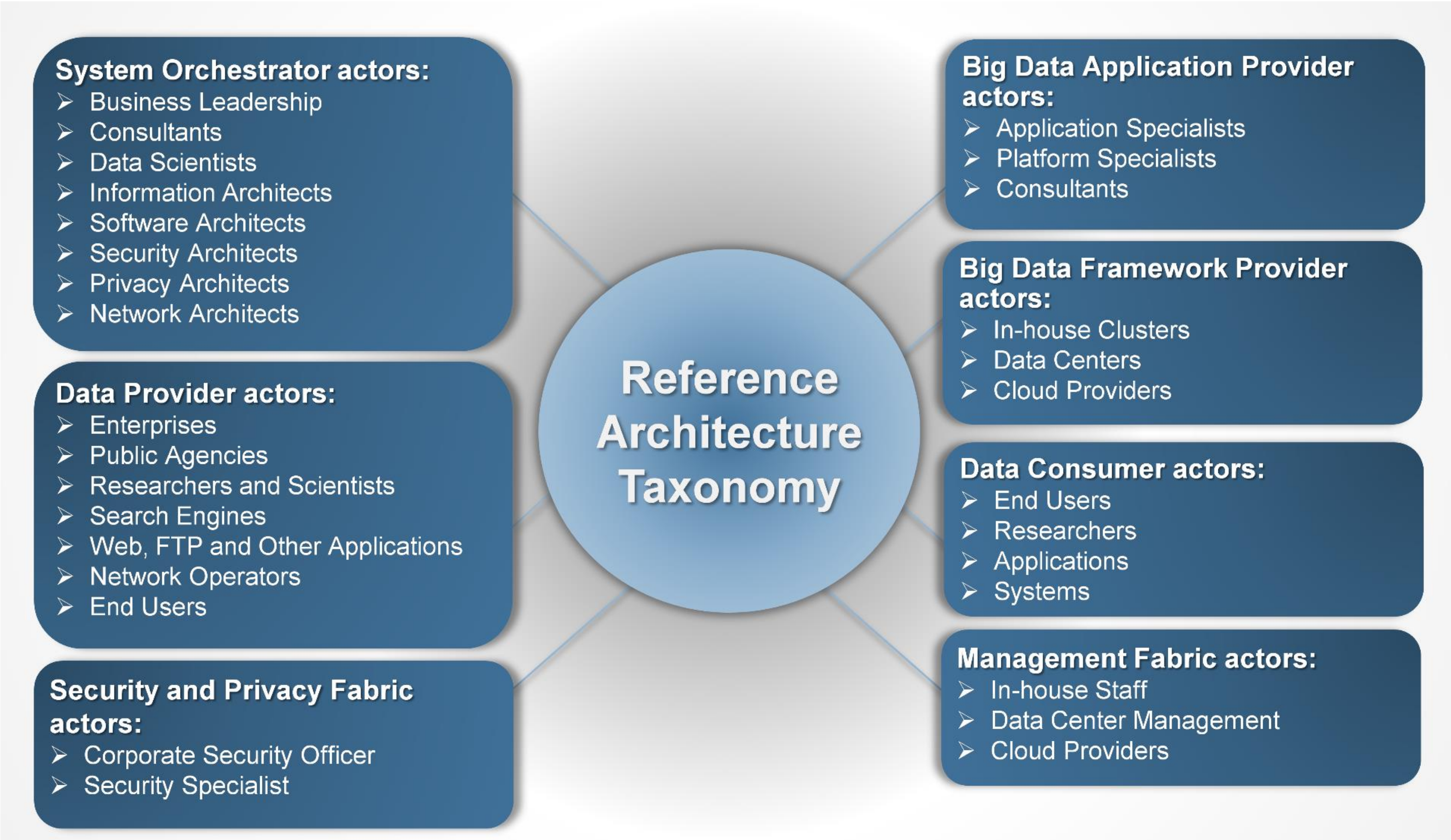
Big Data Lifecycle and Applications Provider activities

- Collection
- Preparation
- Analysis and Analytics
- Visualization
- Access

Big Data Ecosystem includes all components that are involved into Big Data production, processing, delivery, and consuming

[ref] Volume 6: NIST Big Data Reference Architecture. http://bigdatawg.nist.gov/V1_output_docs.php

NIST Big Data Reference Architecture Taxonomy – Roles and actors



Reference Architecture Taxonomy

System Orchestrator actors:
- Business Leadership
- Consultants
- Data Scientists
- Information Architects
- Software Architects
- Security Architects
- Privacy Architects
- Network Architects

Data Provider actors:
- Enterprises
- Public Agencies
- Researchers and Scientists
- Search Engines
- Web, FTP and Other Applications
- Network Operators
- End Users

Security and Privacy Fabric actors:
- Corporate Security Officer
- Security Specialist

Big Data Application Provider actors:
- Application Specialists
- Platform Specialists
- Consultants

Big Data Framework Provider actors:
- In-house Clusters
- Data Centers
- Cloud Providers

Data Consumer actors:
- End Users
- Researchers
- Applications
- Systems

Management Fabric actors:
- In-house Staff
- Data Center Management
- Cloud Providers

[ref] Volume 6: NIST Big Data Reference Architecture. http://bigdatawg.nist.gov/V1_output_docs.php

Big Data Architecture Framework (BDAF)

(1) Big Data Management

- Big Data Governance and FAIR (Findable, Accessible, Interoperable, Re-usable) data principles
- Big Data Lifecycle (Management) Model
- Provenance, Curation, Archiving

(2) Data Models, Structures, Types

- Data formats, relational/non-relational, SQL/NoSQL, file systems, etc.

(3) Big Data Analytics and Tools (BDA)

- Big Data Analytics and Machine Learning methods/algorithms
- Target use, presentation, visualisation

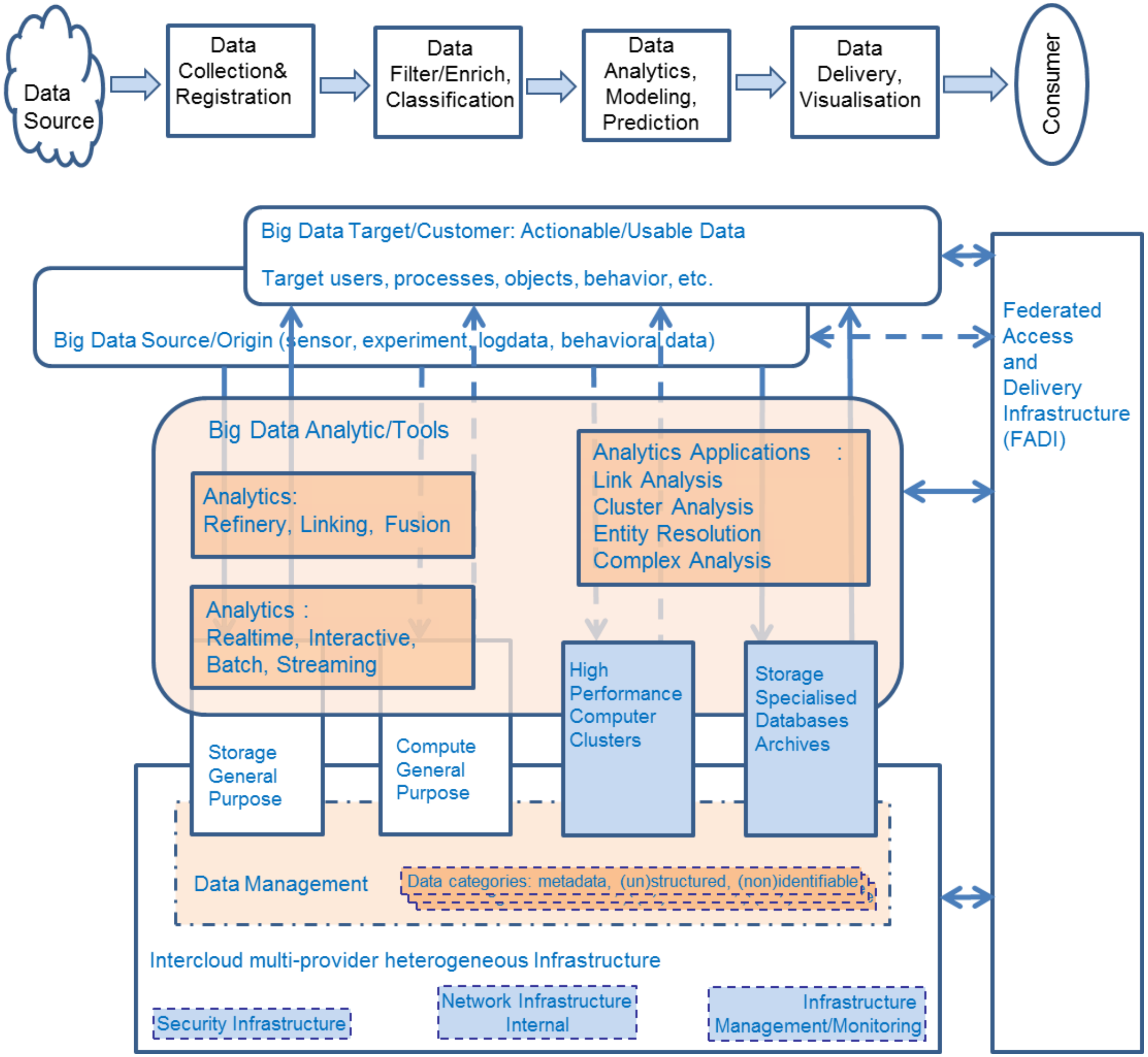
(4) Big Data Infrastructure (BDI)

- Highly scalable Storage, Compute, High Performance Network
- Big Data Analytics platforms
- Sensor network, target/actionable devices

(5) Big Data Security

- Data security in-rest, in-move, trusted processing environments
- Data Sovereignty
- Big Data compliance, data verifiability and trustworthiness
- Digital rights protection
- Privacy and personal information protection

Big Data Infrastructure and Analytics Tools



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

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)

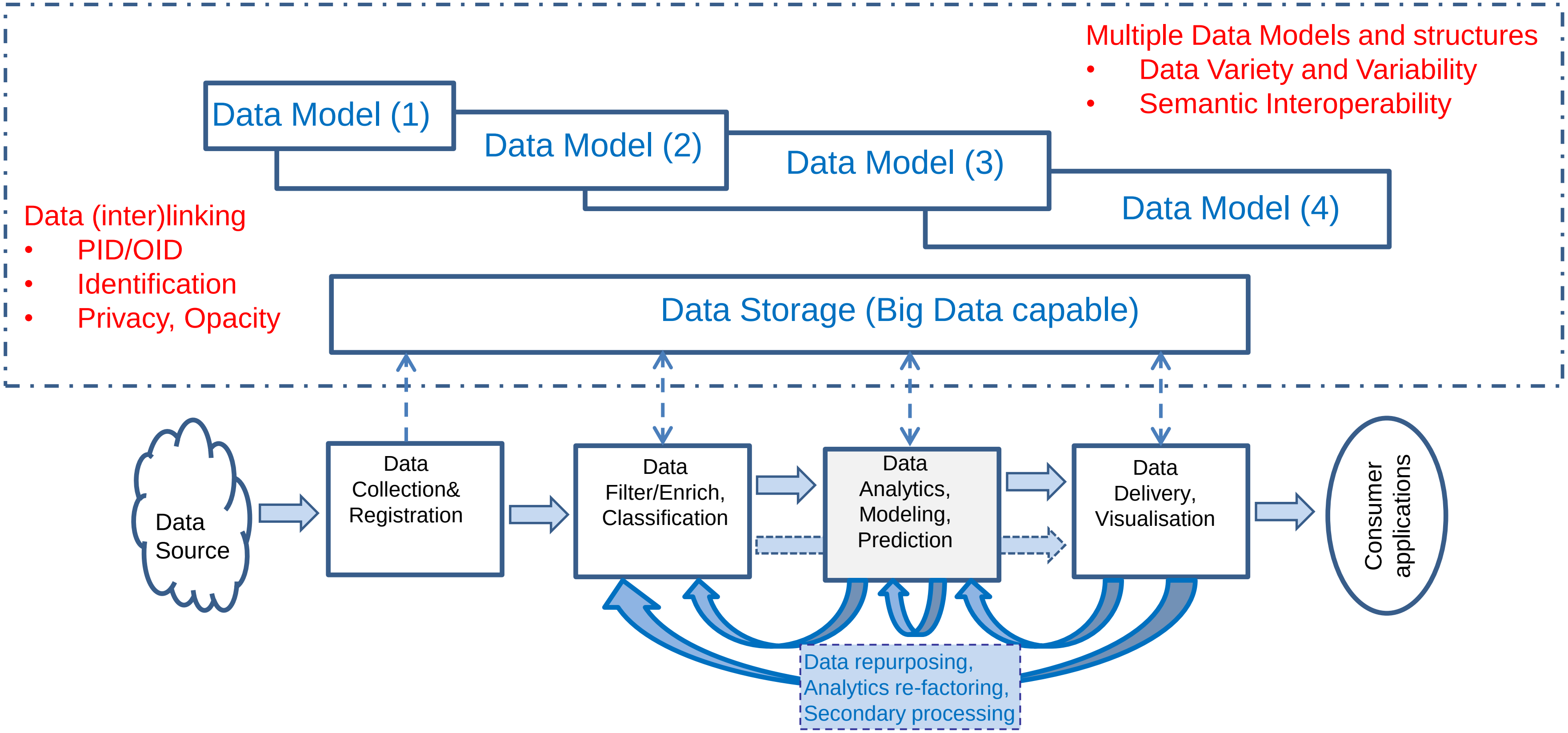
Big Data Infrastructure

- Heterogeneous multi-provider inter-cloud infrastructure
- Data management infrastructure
- Collaborative Environment
- Advanced high performance (programmable) network
- Security infrastructure
- Federated Access and Delivery Infrastructure (FADI)

Big Data Analytics Infrastructure/Tools – Hadoop/Spark Platform based

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

Data Lifecycle/Transformation Model


Multiple Data Models and structures
• Data Variety and Variability
• Semantic Interoperability
Data Model (1)
Data Model (2)
Data Model (3)
Data Model (4)
Data (inter)linking
• PID/OID
• Identification
• Privacy, Opacity
Data Storage (Big Data capable)
Data Source
Data Collection& Registration
Data Filter/Enrich, Classification
Data Analytics, Modeling, Prediction
Data Delivery, Visualisation
Consumer applications
Data repurposing, Analytics re-factoring, Secondary processing

- Data Model changes along data lifecycle or evolution
- Data provenance is a discipline to track all data transformations along lifecycle

- Identifying and linking data
  - Persistent data/object identifiers (PID/OID)
  - Traceability vs Opacity
  - Referral integrity

Discussion Questions

- Big Data aspects in your organisation
  - Go to **www.menti.com** and use the code **on the screen**

- How to start building your organisation Big Data infrastructure and Big Data Analytics facilities?
  - Cloud is a solution for quick start and onboarding

- How to scale them to specific big and small tasks?

Acknowledgement

- This work is supported by the ERASMUS+ MATES project
- The work is committed to the Open Source under Creative Commons 4.0 CC BY license

This work is licensed under the Creative Commons Attribution 4.0 International License.
To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

Co-funded by the
Erasmus+ Programme
of the European Union

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.