Big Data Analytics

Course Objectives

- glean system’s view of big data analytics
- understand the technological challenges
- develop familiarity with the state of the art

Course Structure

Theory

- lectures
- readings

Lab

- hands-on practice

Course Content

Theory

- original research publications
- articles from tech journals - ACM, Spinger, IEEE, …
- textbooks and reference books

Course Content

Practice

- Hadoop and associated technology stack
- official documentation and reference guides
- reference books and online resources

Evaluation

Two internal tests

Lab assignments

End-semester written theory exam

End-semester lab exam

What is Big Data?

Big data technologies describe 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.

- *IDC, 2011*

What is Big Data?

Big data is where the data volume, acquisition velocity, or data representation limits the ability to perform effective analysis using traditional relational approaches or requires the use of significant horizontal scaling for efficient processing.

- *NIST*

Big Data Science

The study of techniques covering the acquisition, conditioning, and evaluation of big data.

- *NIST*

Big Data Frameworks

Software libraries along with their associated algorithms that enable distributed processing and analysis of big data problems across clusters of computer units.

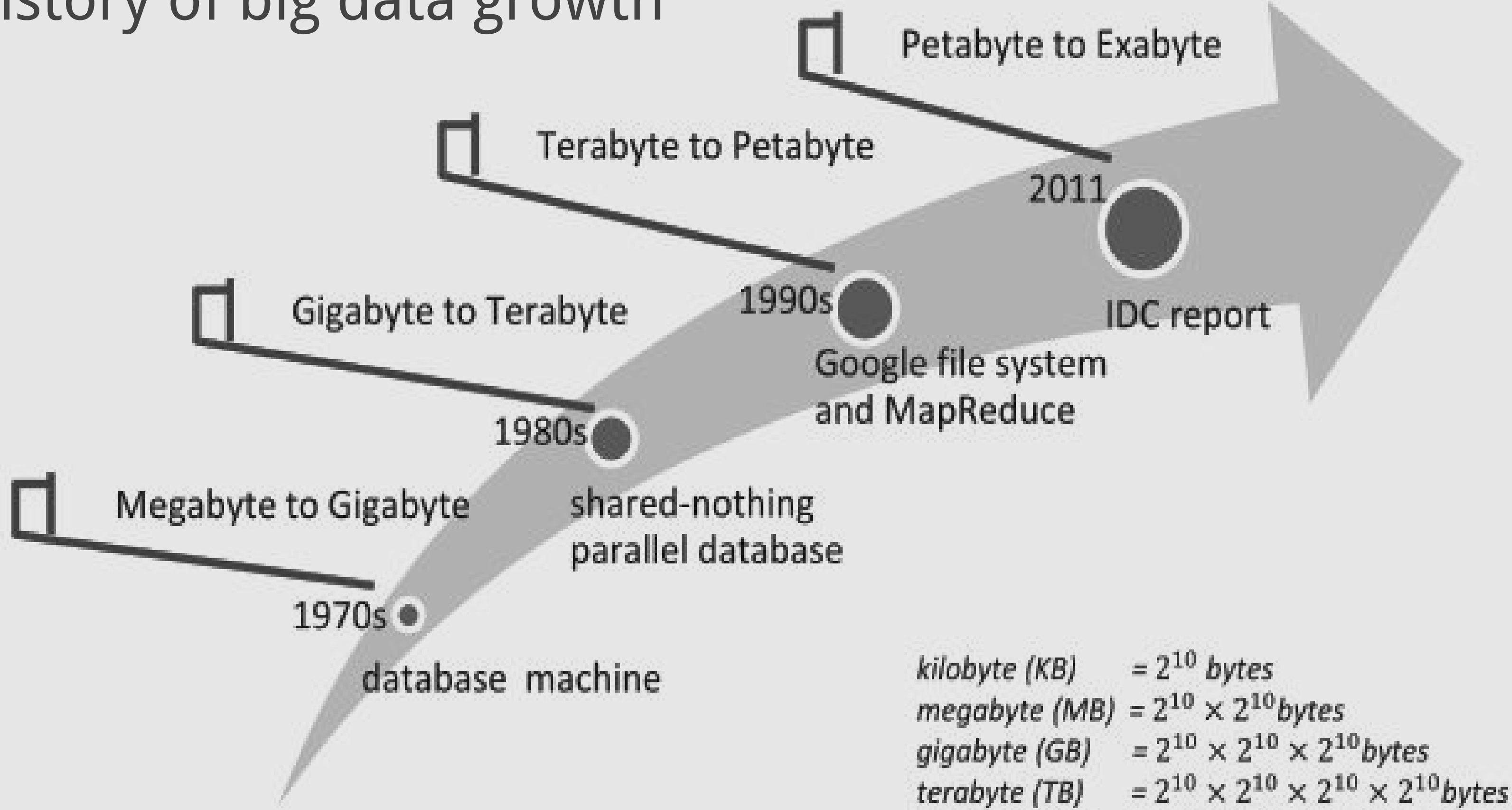
- *NIST*

Big Data Frameworks

“The reason big data is impacting every one of us is the data oozing out of everything… It’s like electricity flowing throughout an organization—everyone can tap into it on command to answer the individual questions their jobs demand”

- *Pat Hanrahan*

History of big data growth



kilobyte (KB) $= 2^{10}$ *bytes*

megabyte (MB) $= 2^{10} \times 2^{10}$*bytes*

gigabyte (GB) $= 2^{10} \times 2^{10} \times 2^{10}$*bytes*

terabyte (TB) $= 2^{10} \times 2^{10} \times 2^{10} \times 2^{10}$*bytes*

petabyte (PB) $= 2^{10} \times 2^{10} \times 2^{10} \times 2^{10} \times 2^{10}$*bytes*

exabyte (EB) $= 2^{10} \times 2^{10} \times 2^{10} \times 2^{10} \times 2^{10} \times 2^{10}$*bytes*

Big Data Analytics

- Volume
- Velocity
- Value
- Variety
- Veracity

Big data analytics is a workflow that distills terabytes of low-value data down to, in some cases, a single bit of high-value data.

The goal is to see the big picture from the minutia of our digital lives.

Big Data Analytics - Limitations

Just because analysts have big data to work with doesn’t guarantee the sample they need is sufficiently representative of their entire user population (bigger is not better)

Big Data Analytics - Limitations

Working with big data is still subjective and that automated data collection is not self-explanatory — it requires selection and interpretation.

Data sampling and cleaning processes in particular are prone to potential error and bias.

- boyd and Crawford

The Nature of Analytics Work

Corporate analytics teams

The analytics team uses their expertise in statistics, data mining, machine learning, and visualization to answer questions that corporate leaders pose.

The Nature of Analytics Work

- The work is exploratory and demand-driven
- The ultimate goal of the work is clear communication
- The work must produce high-confidence results
- The work creates a strong need to preserve institutional memory

Big Data Paradigms

Batch processing

Stream processing

- complex event processing
- stream data processing

Examples of Real time Analytics

Academic research scientists

- Research scientists analyze data to test hypotheses and form theories.
- Scientists typically choose their own research questions, exercise more control over the source data, and report results to knowledgeable peers

Examples of Real time Analytics

- Capital markets
  - algorithmic trading
  - smart order routing
- Banking
  - credit card fraud detection

- Healthcare
  - patient monitoring
  - fraud detection

Examples of Real time Analytics

Public sector

- surveillance
- emergency response
- security

Energy

- energy trading
- pipeline monitoring
- power grid control

Examples of Real time Analytics

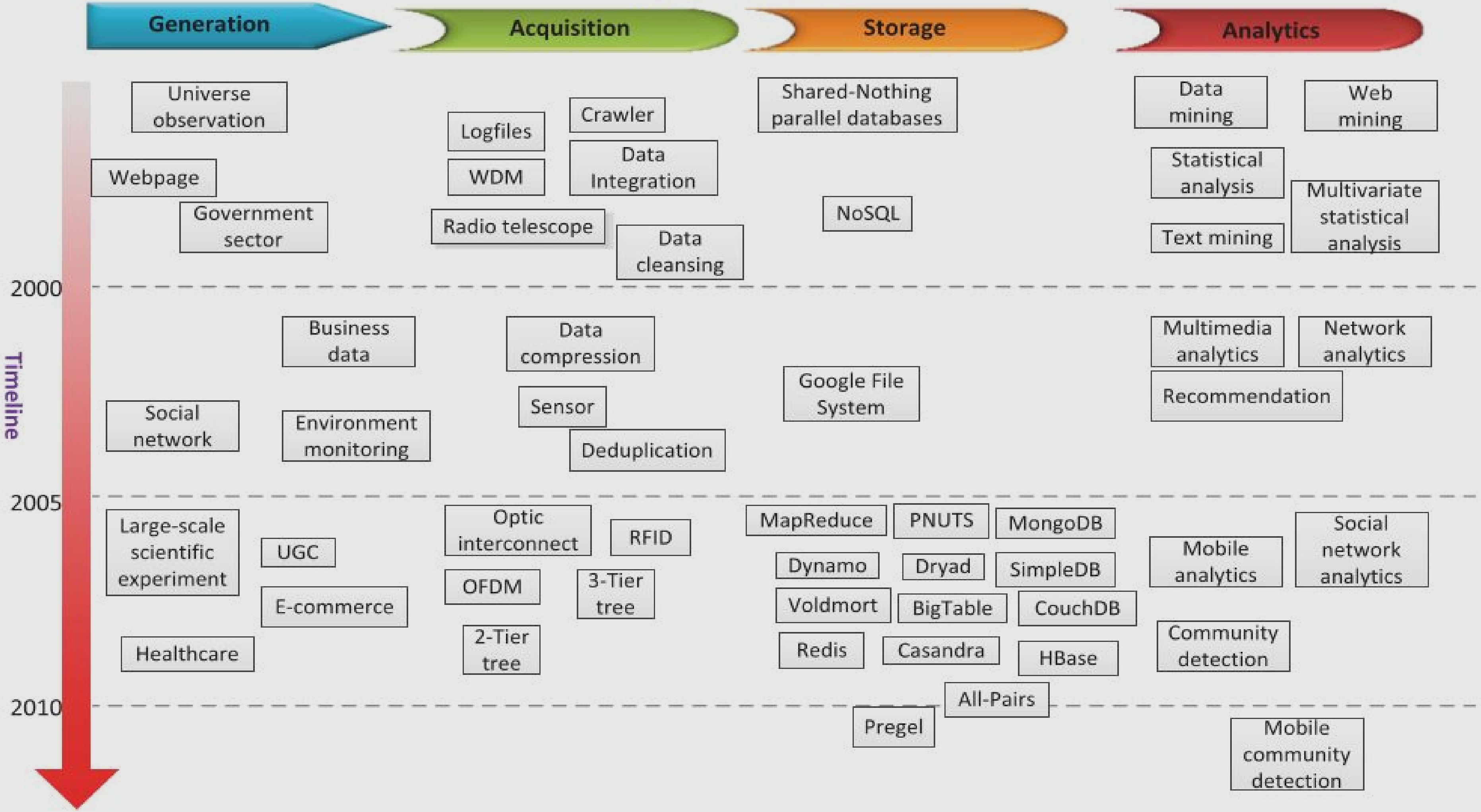
Web

- click-stream analysis
- resource monitoring
- fraud detection

Energy

- energy trading
- pipeline monitoring
- power grid control

Big Data Value Chain



Generation

Acquisition

Storage

Analytics

Timeline

Universe observation

Webpage

Government sector

Logfiles

WDM

Radio telescope

Crawler

Data Integration

Data cleansing

Shared-Nothing parallel databases

NoSQL

Data mining

Web mining

Statistical analysis

Text mining

Multivariate statistical analysis

2000

Business data

Social network

Environment monitoring

Data compression

Sensor

Deduplication

Google File System

Multimedia analytics

Network analytics

Recommendation

2005

Large-scale scientific experiment

UGC

E-commerce

Healthcare

Optic interconnect

RFID

OFDM

3-Tier tree

2-Tier tree

MapReduce

PNUTS

MongoDB

Dynamo

Dryad

SimpleDB

Voldmort

BigTable

CouchDB

Redis

Casandra

HBase

All-Pairs

Pregel

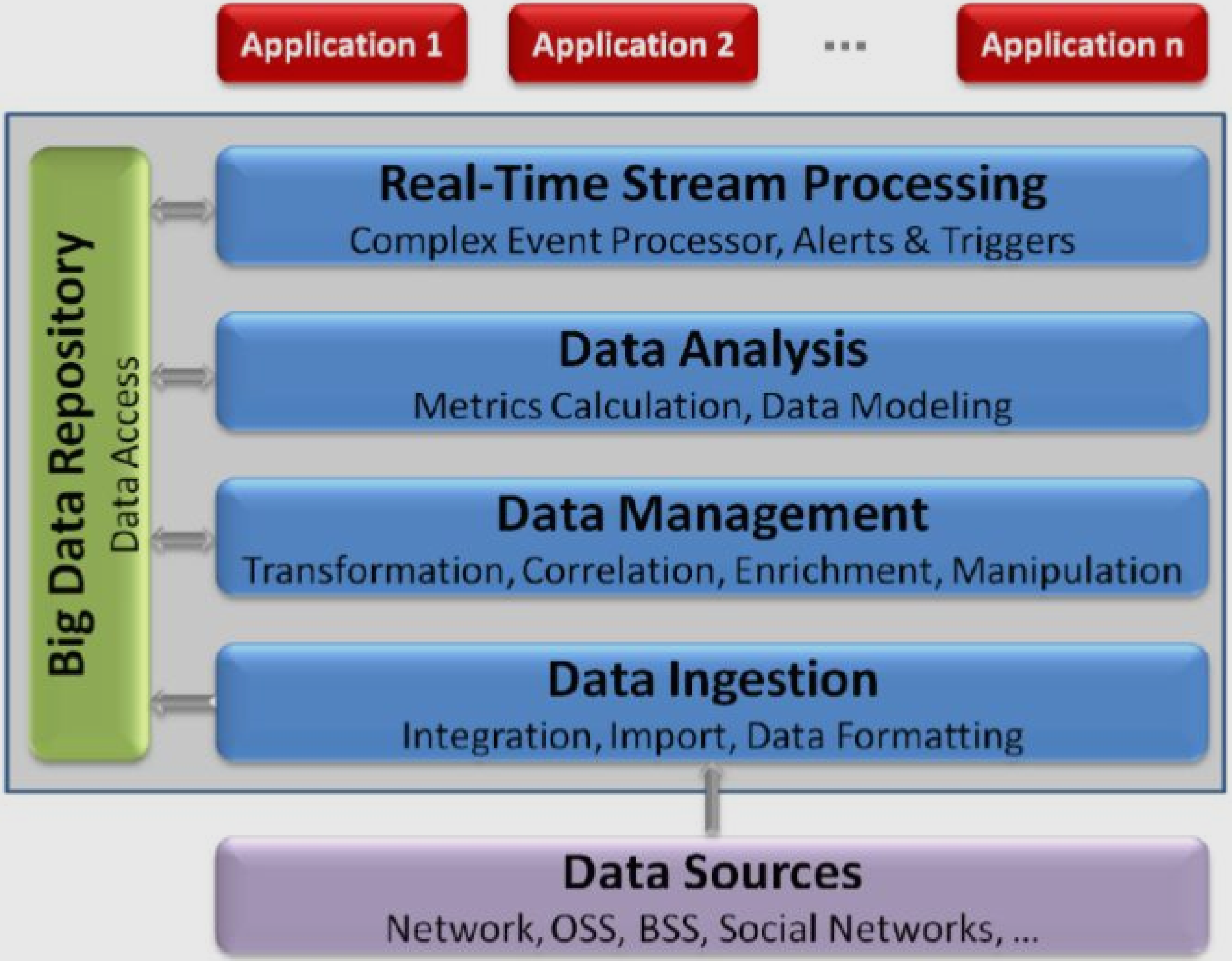
Mobile analytics

Social network analytics

Community detection

2010

Mobile community detection



Application 1

Application 2

...

Application n

Real-Time Stream Processing

Complex Event Processor, Alerts & Triggers

Data Analysis

Metrics Calculation, Data Modeling

Data Management

Transformation, Correlation, Enrichment, Manipulation

Data Ingestion

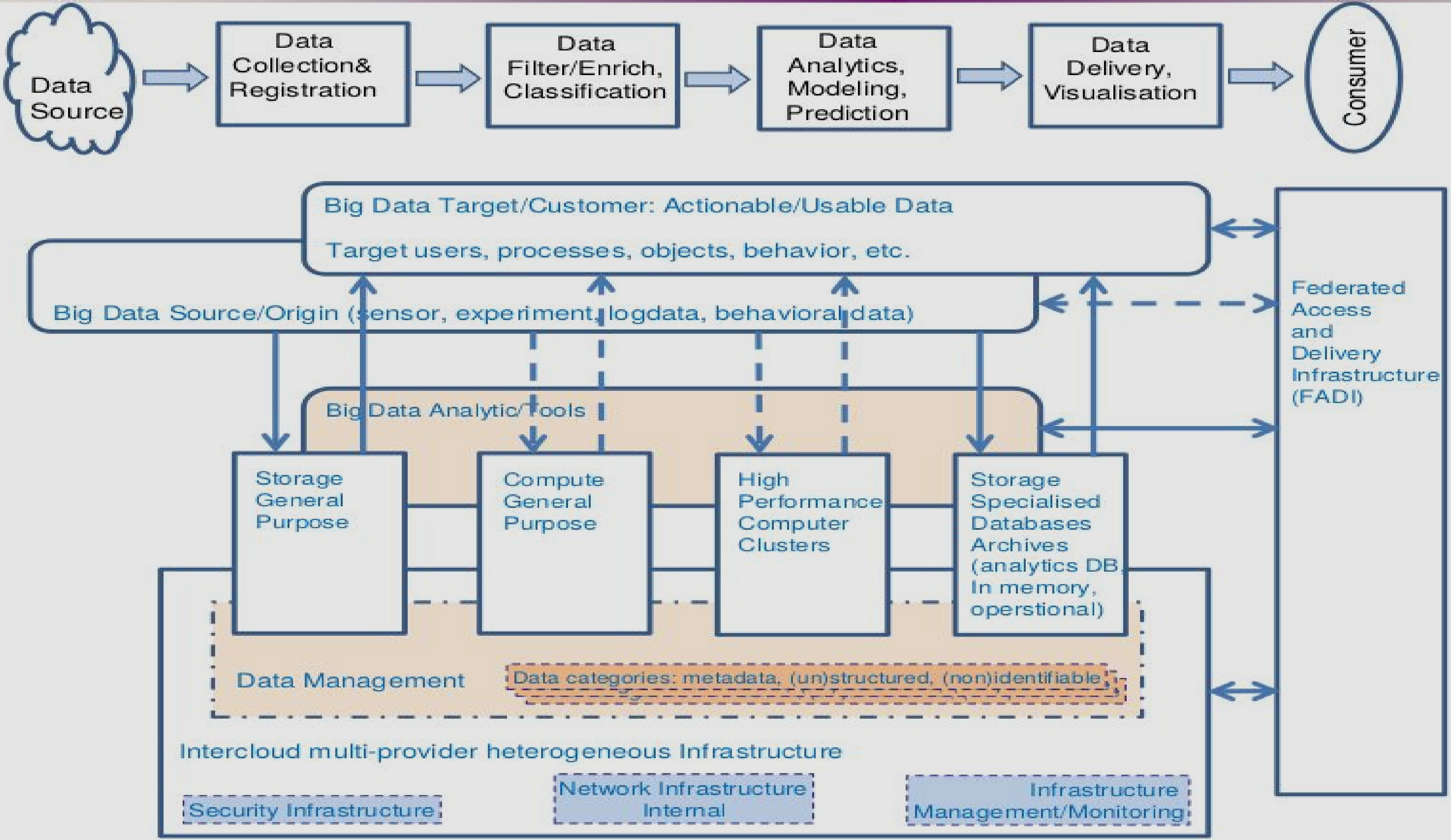
Integration, Import, Data Formatting

Big Data Repository

Data Access

Data Sources

Network, OSS, BSS, Social Networks, ...



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)

Federated Access and Delivery Infrastructure (FADI)

Big Data Analytic/Tools

Storage General Purpose

Compute General Purpose

High Performance Computer Clusters

Storage Specialised Databases Archives (analytics DB, In memory, operstional)

Data Management

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

Intercloud multi-provider heterogeneous Infrastructure

Security Infrastructure

Network Infrastructure Internal

Infrastructure Management/Monitoring