Overview NIST Big Data Working Group Activities

and

Big Data Architecture Framework (BDAF) by UvA

Yuri Demchenko

SNE Group, University of Amsterdam

Big Data Analytics Interest Group

17 September 2013, 2nd RDA Plenary

Outline

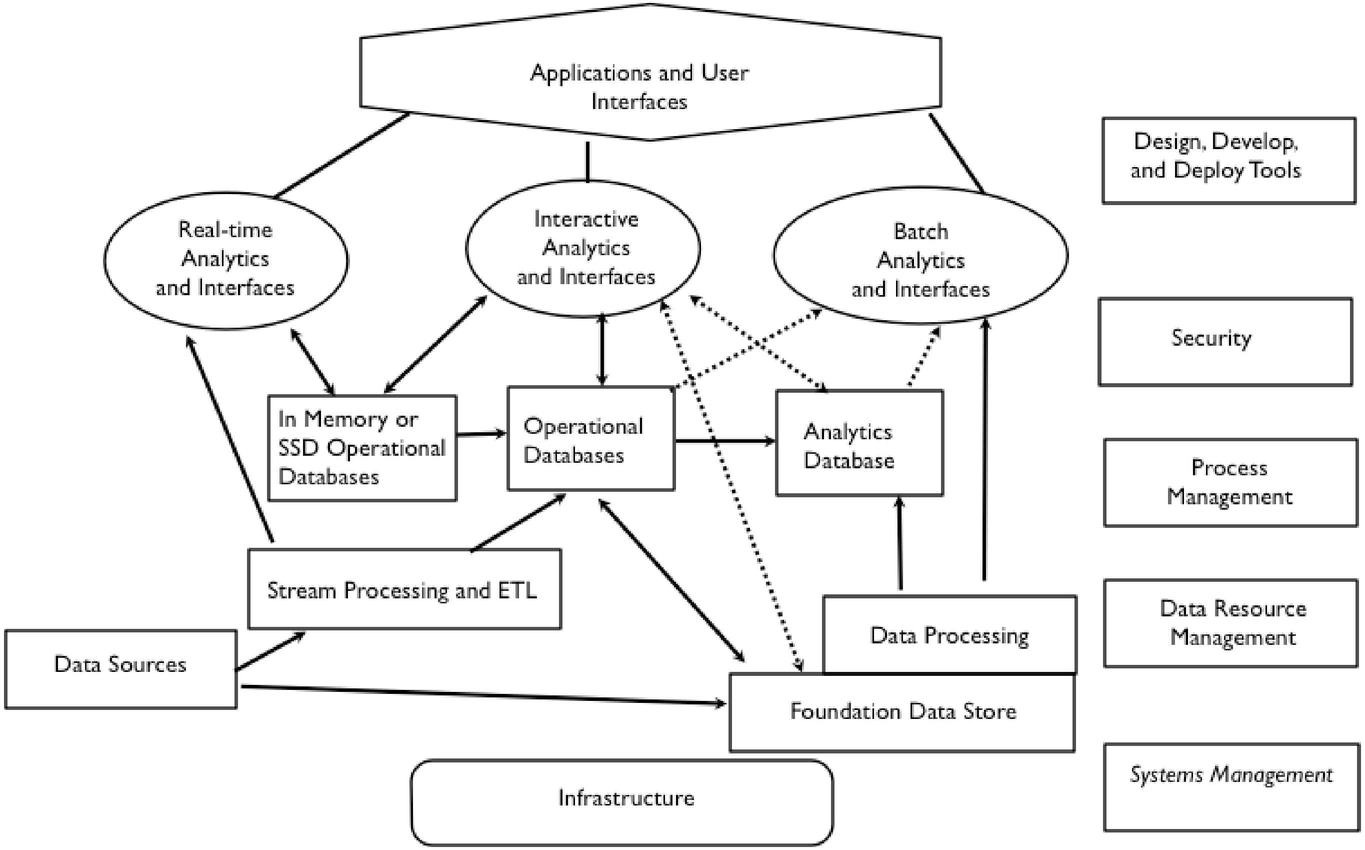
- Overview NIST Big Data Working Group (NBD-WG) activities and deliverables
- Proposed Big Data Architecture Framework (BDAF)
  - Data Models and Big Data Lifecycle
  - Big Data Infrastructure (BDI)
- Discussion: Liaison and information exchange with NIST BD-WG

Disclaimer: Presented here information about NIST Big Data Working Group (NBD-WG) and images from the NBD-WG working documents are not official position of the NBD-WG and are solely the authors opinion.

NIST Big Data Working Group (NBD-WG)

- Deliverables target – September 2013
  - 26 September – initial draft documents
  - 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 usecases http://bigdatawg.nist.gov/_uploadfiles/M0224_v1_1076079077.xlsx

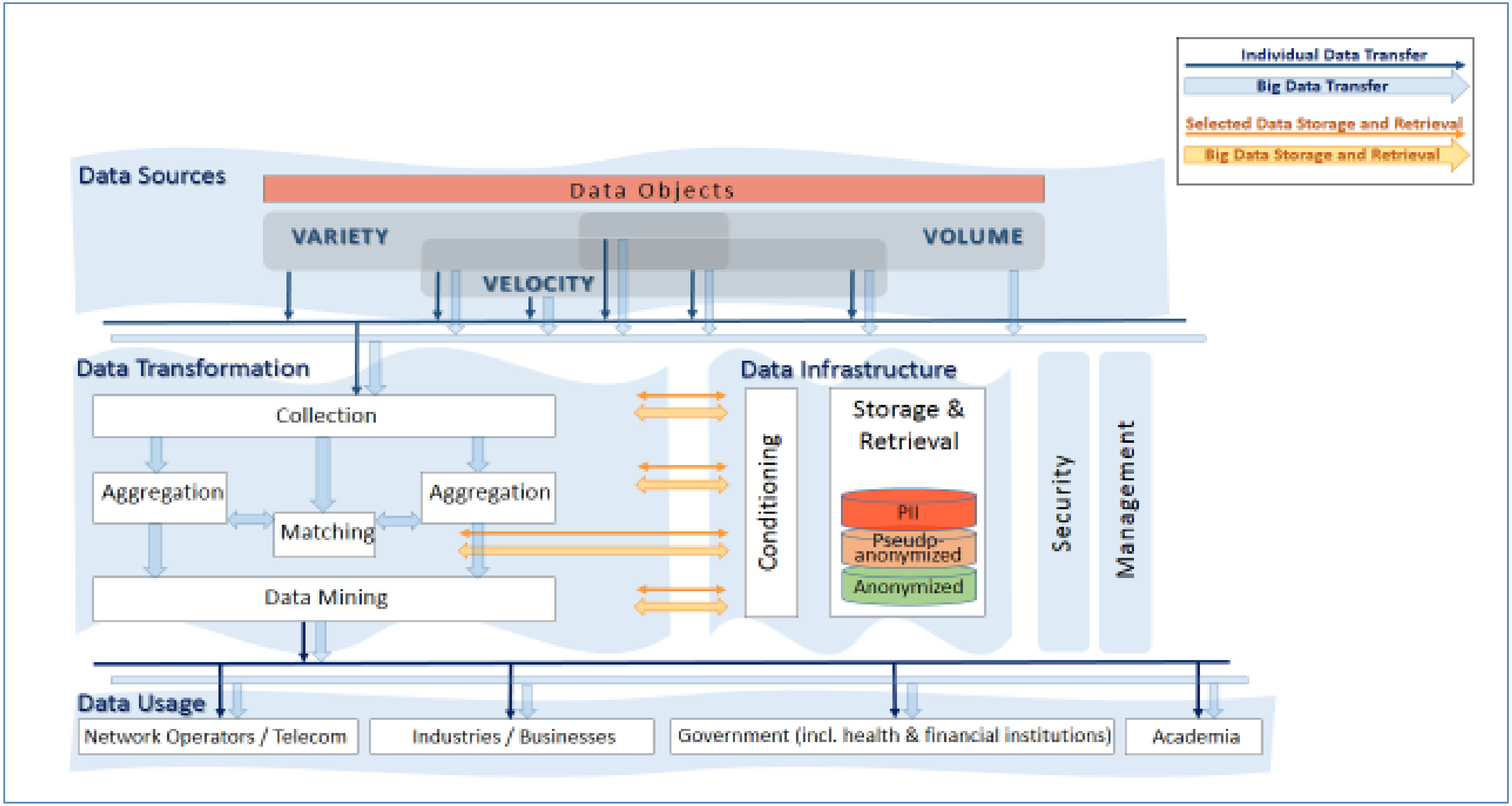
NIST Proposed Reference Architecture (before July 2013)


Applications and User Interfaces
Design, Develop, and Deploy Tools
Real-time Analytics and Interfaces
Interactive Analytics and Interfaces
Batch Analytics and Interfaces
Security
In Memory or SSD Operational Databases
Operational Databases
Analytics Database
Process Management
Stream Processing and ETL
Data Processing
Data Resource Management
Data Sources
Foundation Data Store
Infrastructure
Systems Management

- Obviously not data centric
- Doesn’t make data (lifecycle) management clear

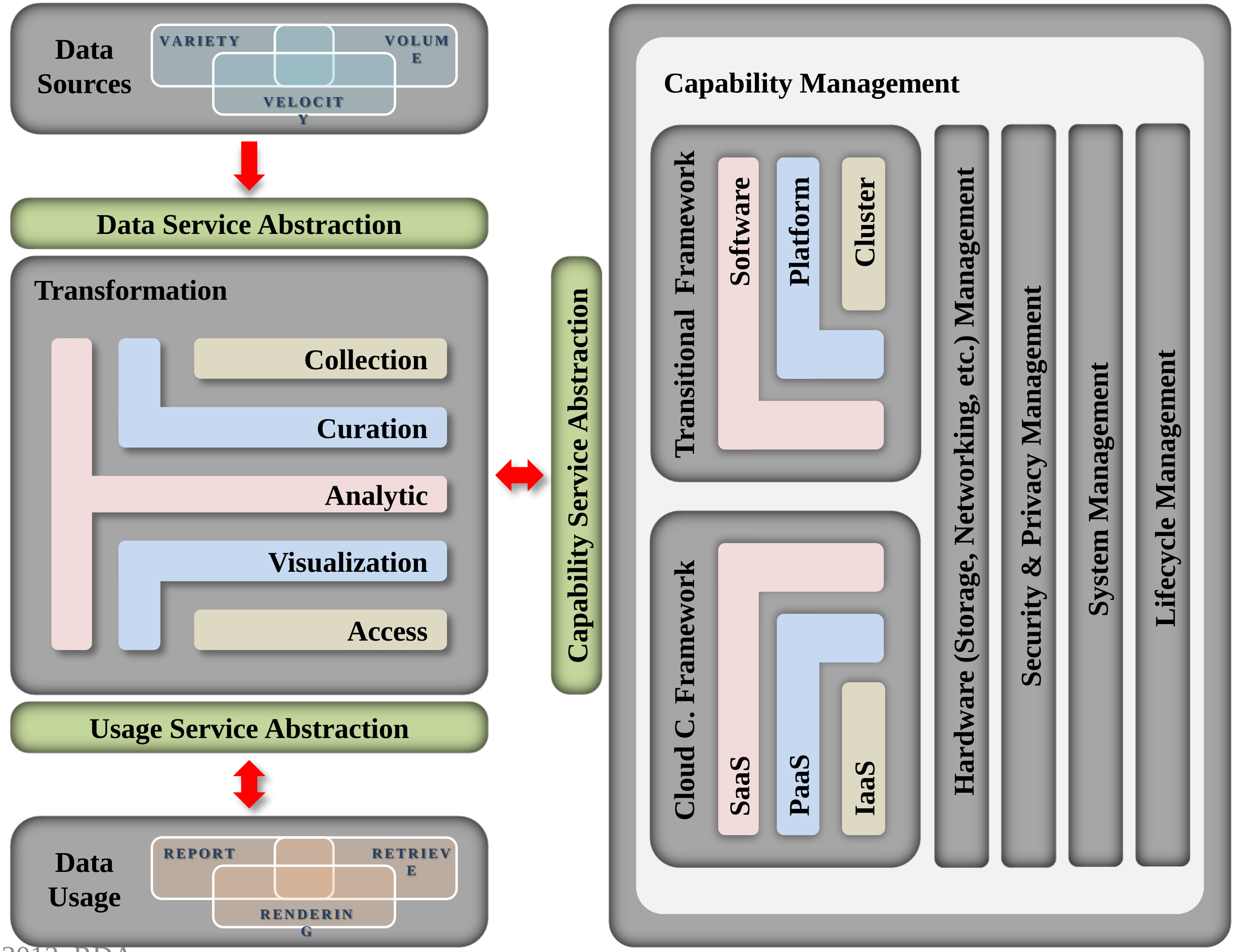
[ref] NIST Big Data WG mailing list discussion
http://bigdatawg.nist.gov/_uploadfiles/M0010_v1_6762570643.pdf

Big Data Ecosystem Reference Architecture (By Microsoft) [ref] – Initial contribution July 2013

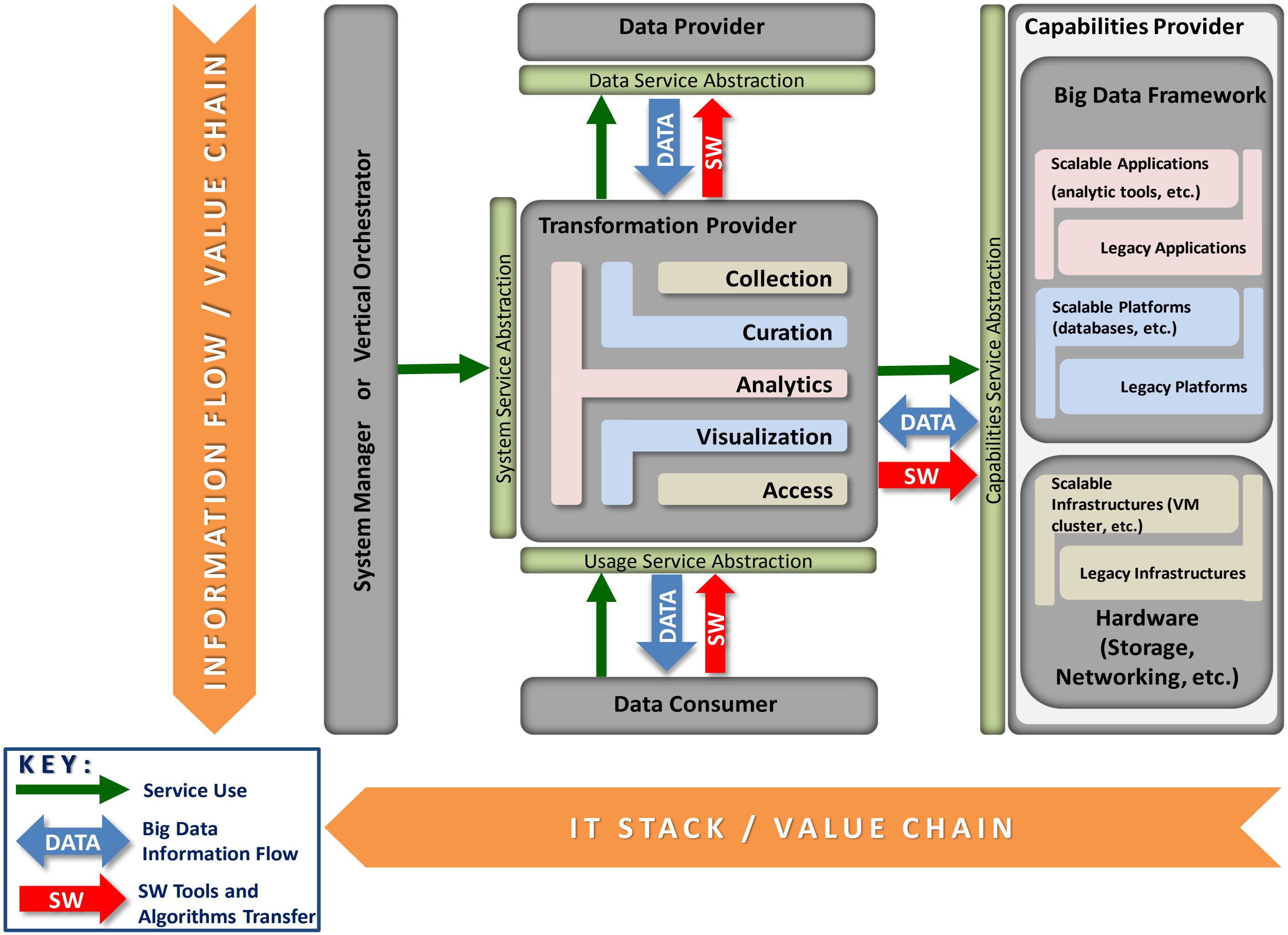


[ref] Big Data Ecosystem Reference Architecture (Microsoft)
http://bigdatawg.nist.gov/_uploadfiles/M0015_v1_1596737703.docx

NIST Reference Architecture version 0.0 (August 2013)



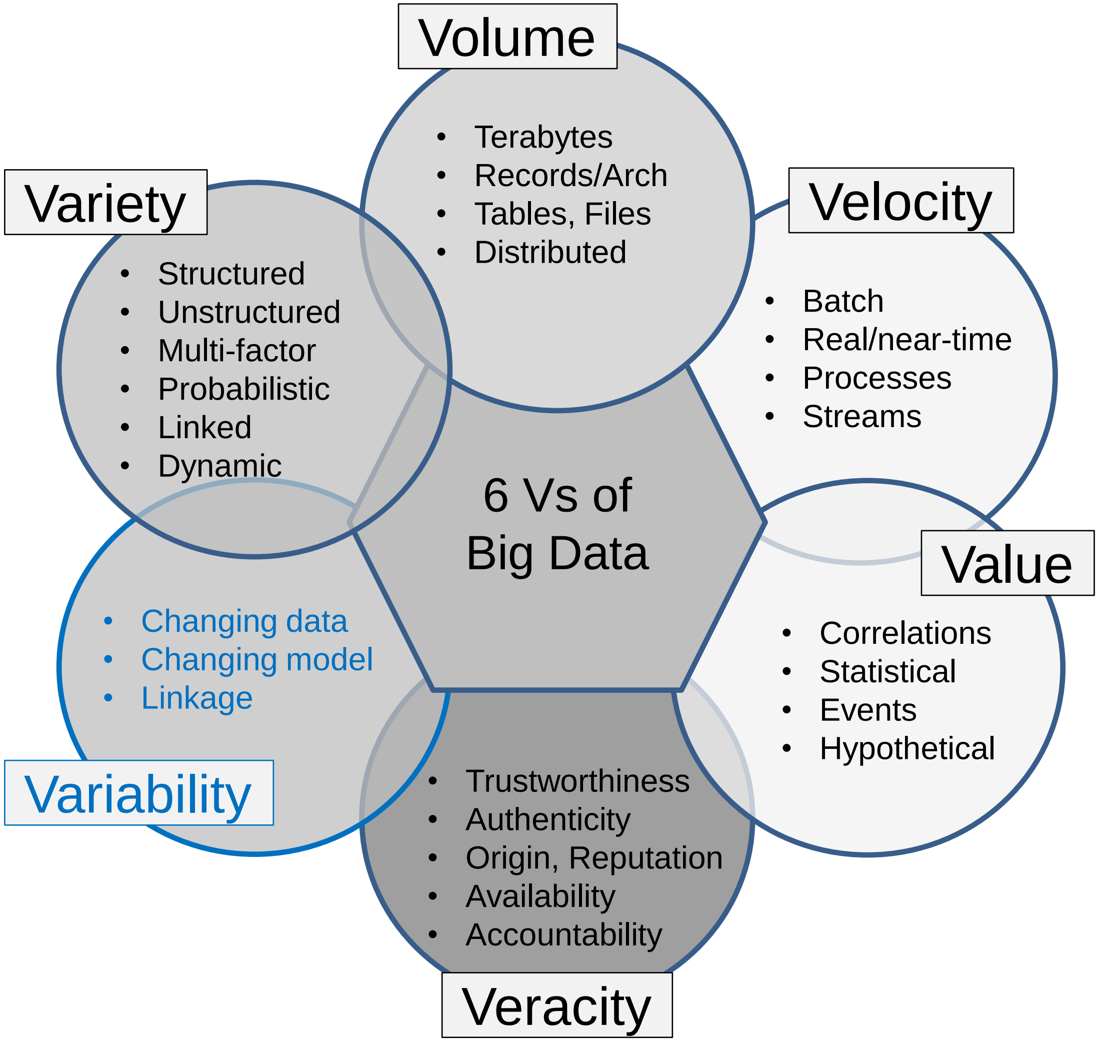
NIST Reference Architecture version 0.1 (September 2013)


INFORMATION FLOW / VALUE CHAIN
System Manager or Vertical Orchestrator
Data Provider
Data Service Abstraction
DATA
SW
System Service Abstraction
Transformation Provider
Collection
Curation
Analytics
Visualization
Access
Usage Service Abstraction
DATA
SW
Data Consumer
DATA
SW
Capabilities Service Abstraction
Capabilities Provider
Big Data Framework
Scalable Applications (analytic tools, etc.)
Legacy Applications
Scalable Platforms (databases, etc.)
Legacy Platforms
Scalable Infrastructures (VM cluster, etc.)
Legacy Infrastructures
Hardware (Storage, Networking, etc.)
KEY:
Service Use
DATA
Big Data Information Flow
SW
SW Tools and Algorithms Transfer
IT STACK / VALUE CHAIN

Big Data Architecture Framework (BDAF) by the University of Amsterdam

- Big Data definition: from 5+1Vs to 5 parts
- Big Data Architecture Framework (BDAF) components

Improved: 5+1 V's of Big Data



Generic Big Data Properties

- Volume
- Variety
- Velocity

Acquired Properties (after entering system)

- Value
- Veracity
- Variability

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 5V to 5 Parts (2)

Refining Gartner definition

- 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

Big Data Nature: Origin and consumers (target)

Big Data Origin

- Science
- Telecom
- Industry
- Business
- Living Environment, Cities
- Social media and networks
- Healthcare

Big Data Target Use

- Scientific discovery
- New technologies
- Manufacturing, processes, transport
- Personal services, campaigns
- Living environment support
- Healthcare support

Big Data Nature: Origin and consumers (targets)

	Scietific Discovery	New Technology	Manufactur Transport	Personal services, campaigns	Living Environmnt, Infrastruct, Utility	Healthcare support
Science	+++++	++++	+	-	++	+++
Telecom	+	++++	++	+	++++	+
Industry	++	++++	+++++	-	-	++
Business	+	+++	++	-	+	++
Living environment, Cities	++	++	++	++	+++++	+
Social media, networks	+	++	-	++++	++	-
Healthcare	+++	++	-	-	++	+++++

Rich information on usecases is available from the NIST document store
http://bigdatawg.nist.gov/show_InputDoc.php

Moving to Data-Centric Models and Technologies

- 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 variability and lifecycle
  - Data integrity and identification
  - Data centric security and access control

Defining Big Data Architecture Framework

- Existing attempts don't converge to consistent view: ODCA, TMF, NIST
  - See http://bigdatawg.nist.gov/_uploadfiles/M0055_v1_7606723276.pdf
- Big Data Architecture Framework (BDAF) by UvA
  Architecture Framework and Components for the Big Data Ecosystem. Draft Version 0.2
  http://www.uazone.org/demch/worksinprogress/sne-2013-02-techreport-bdaf-draft02.pdf
- 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) 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

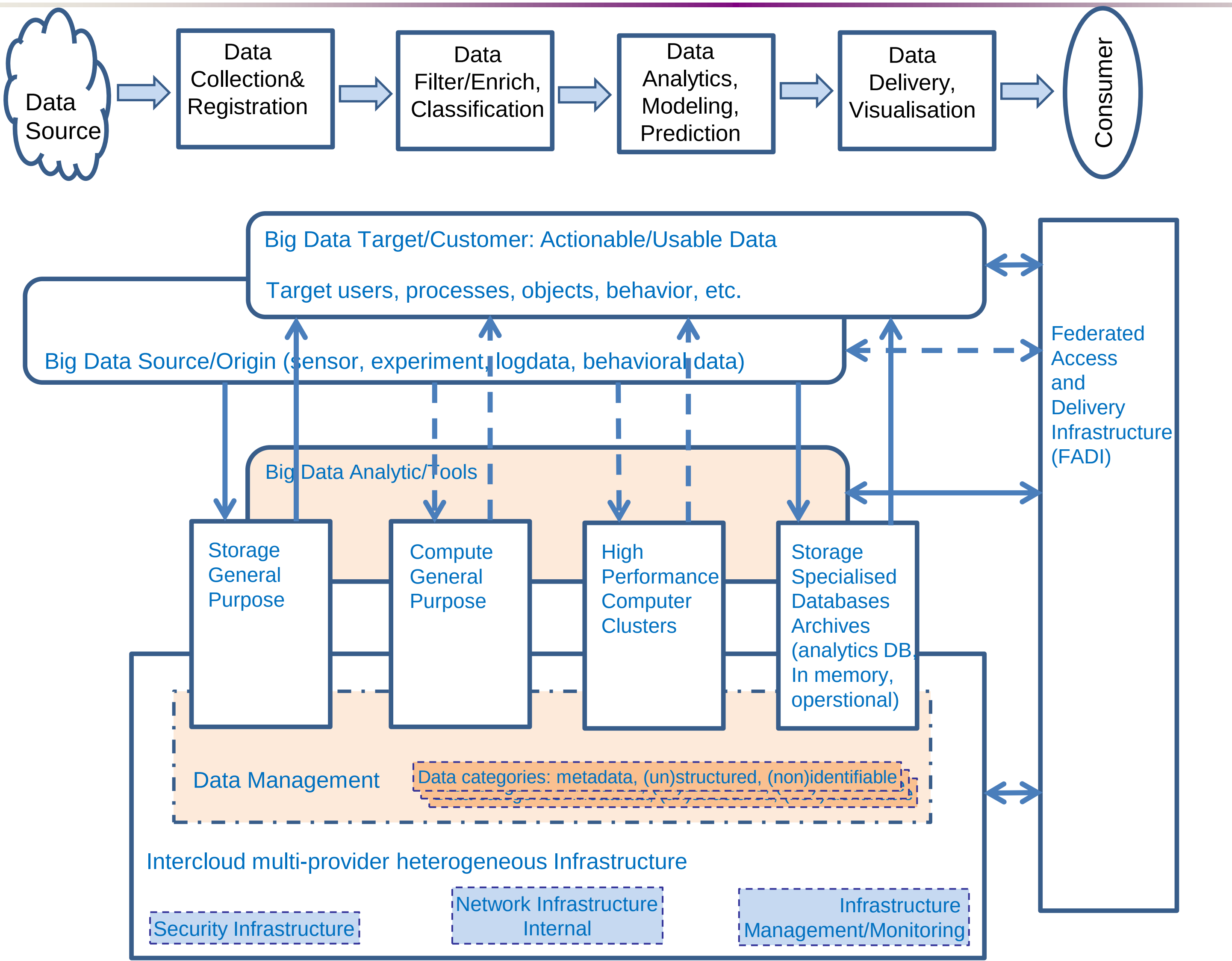
(5) Big Data Security

- 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 Ecosystem: Data, Lifecycle, 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)

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

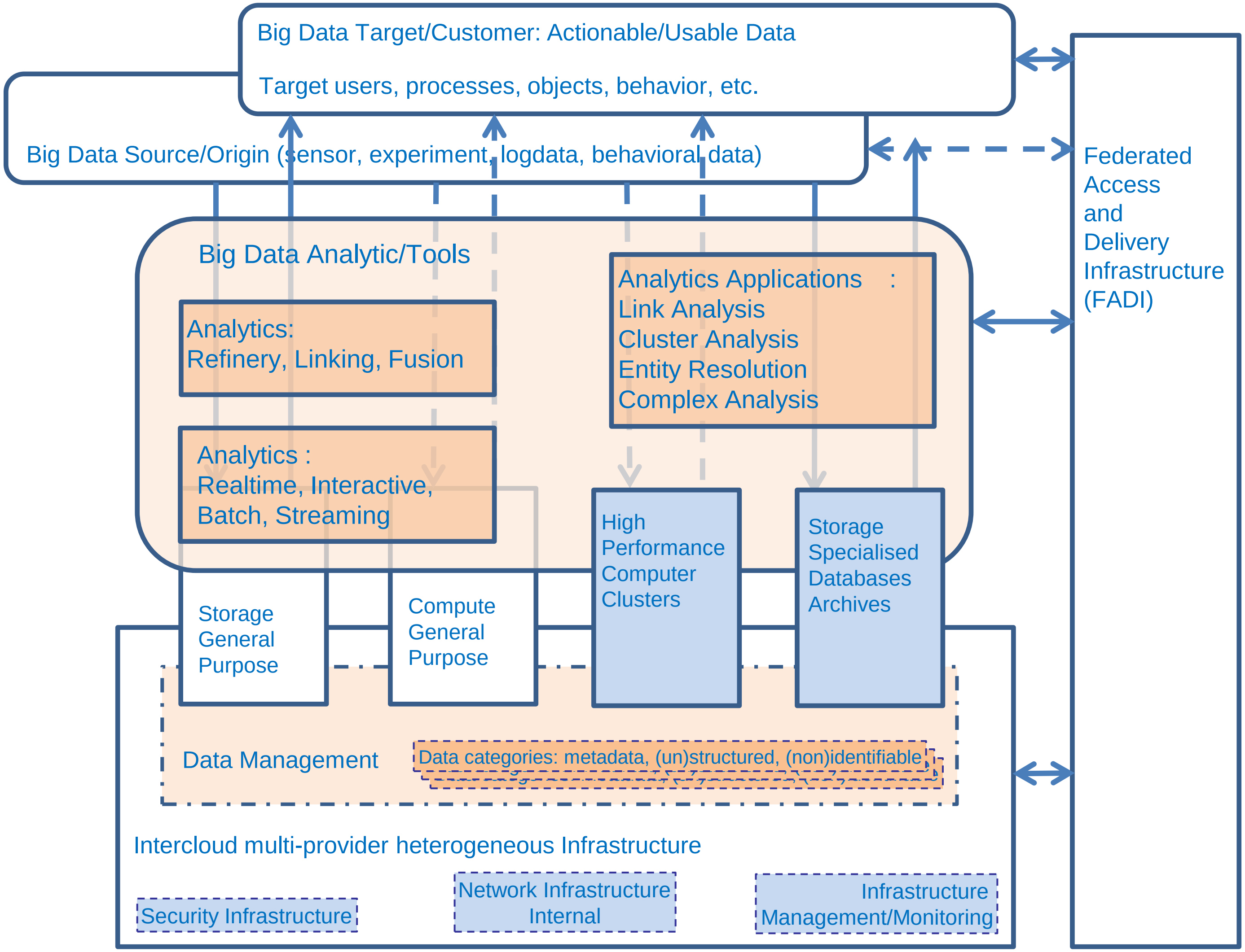
Intercloud multi-provider heterogeneous Infrastructure

Security Infrastructure

Network Infrastructure Internal

Infrastructure Management/Monitoring

Big Data Infrastructure and Analytic Tools



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

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

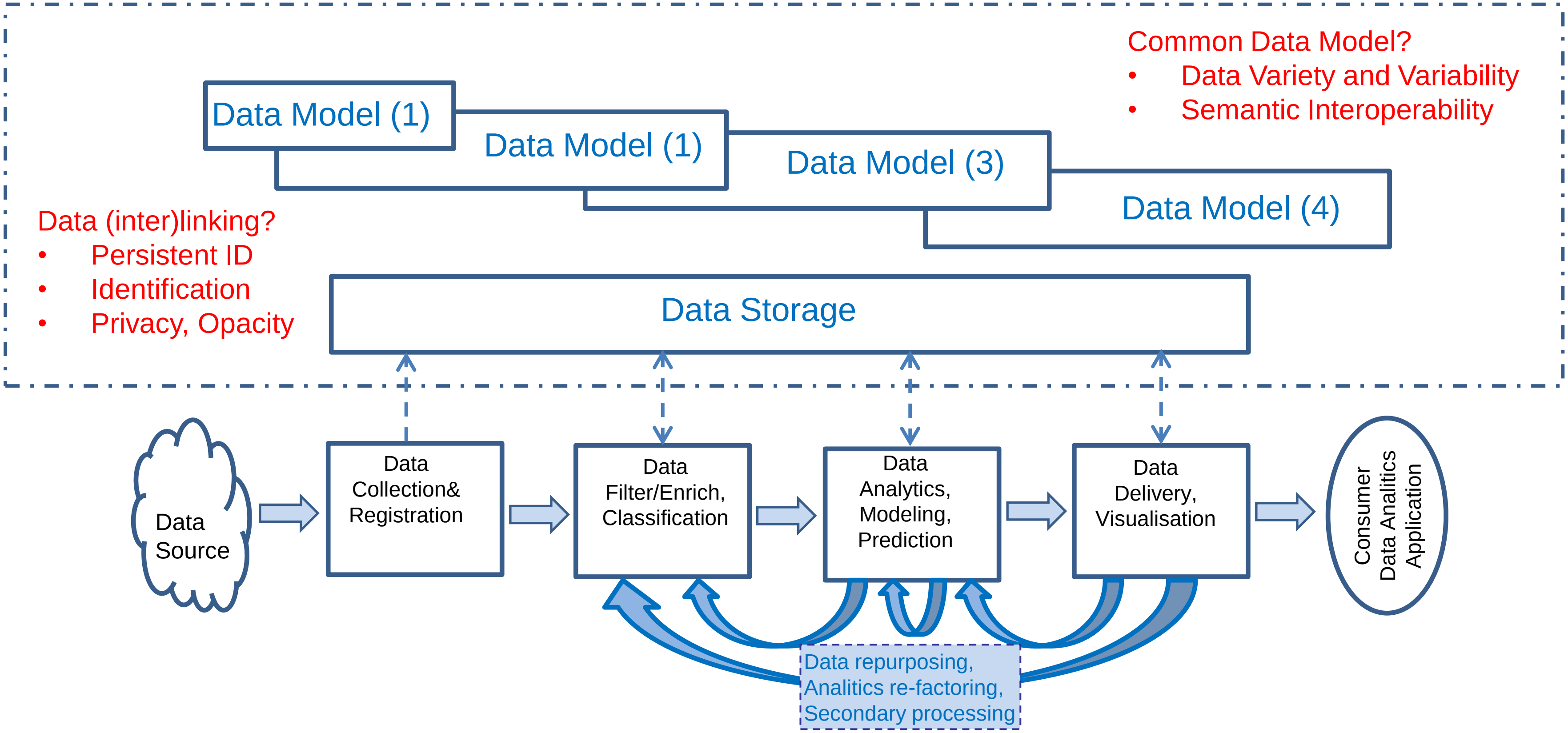
Intercloud multi-provider heterogeneous Infrastructure

Security Infrastructure

Network Infrastructure Internal

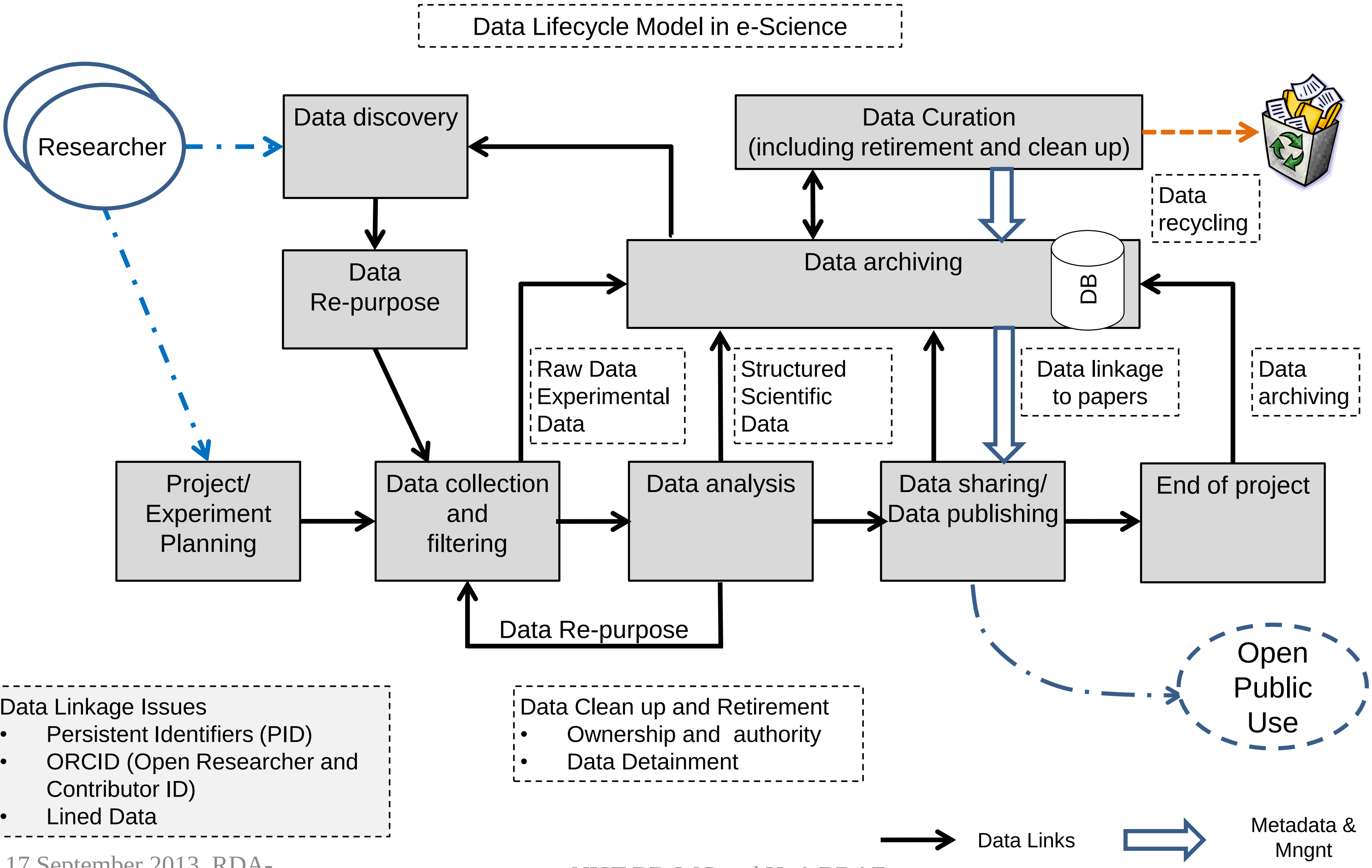
Infrastructure Management/Monitoring

Data Transformation/Lifecycle Model


Common Data Model?
• Data Variety and Variability
• Semantic Interoperability
Data Model (1)
Data Model (1)
Data Model (3)
Data Model (4)
Data (inter)linking?
• Persistent ID
• Identification
• Privacy, Opacity
Data Storage
Data Source
Data Collection& Registration
Data Filter/Enrich, Classification
Data Analytics, Modeling, Prediction
Data Delivery, Visualisation
Consumer Data Analitics Application
Data repurposing, Analitics re-factoring, Secondary processing

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

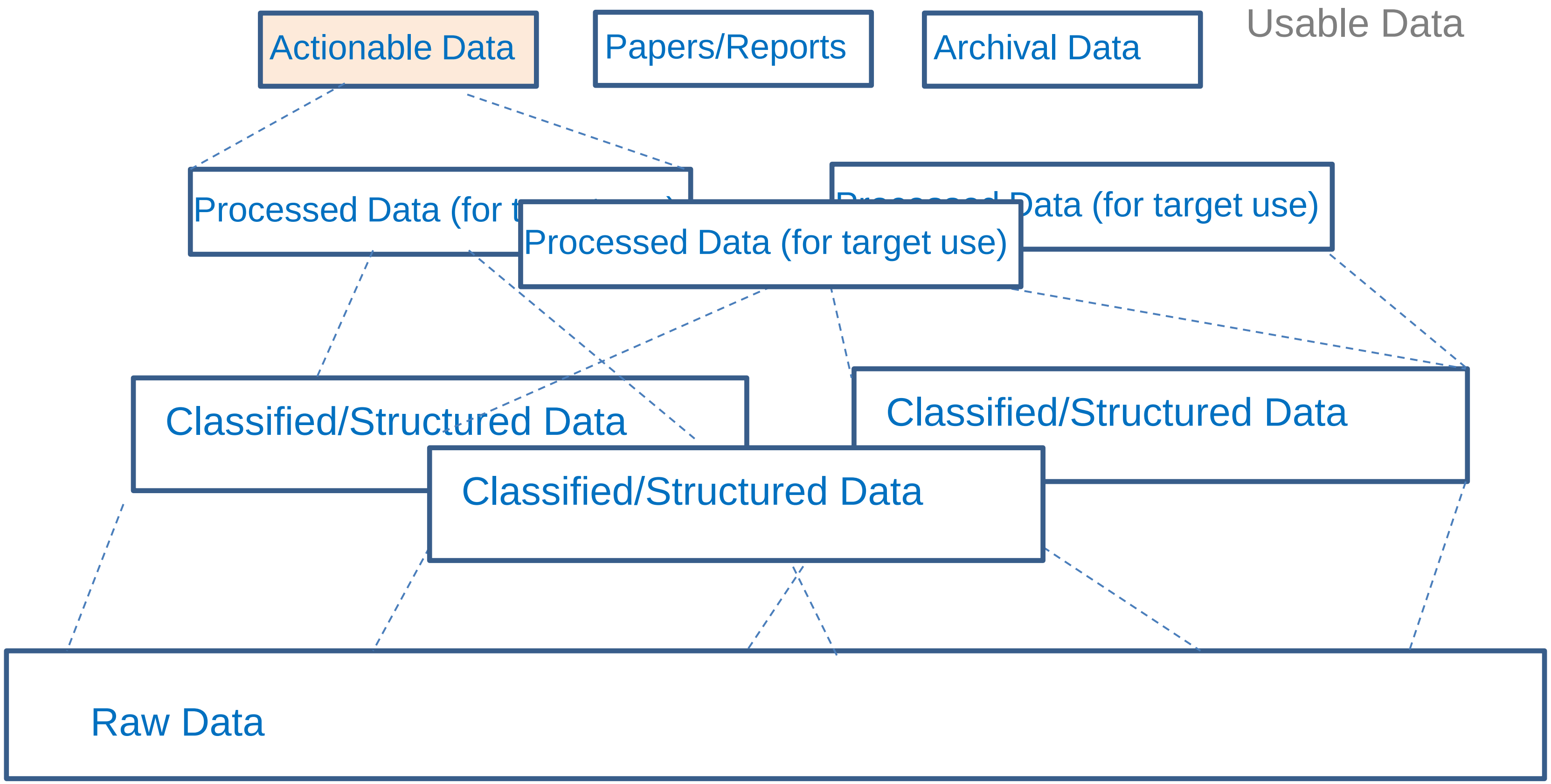
Scientific Data Lifecycle Management (SDLM) Model


Data Lifecycle Model in e-Science
Researcher
Data discovery
Data Curation (including retirement and clean up)
Data recycling
Data Re-purpose
Data archiving
DB
Raw Data Experimental Data
Structured Scientific Data
Data linkage to papers
Data archiving
Project/ Experiment Planning
Data collection and filtering
Data analysis
Data sharing/ Data publishing
End of project
Data Re-purpose
Open Public Use
Data Links
Metadata & Mngnt

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

Data Clean up and Retirement
- Ownership and authority
- Data Detainment

Evolutional/Hierarchical Data Model


Usable Data
Actionable Data
Papers/Reports
Archival Data
Processed Data (for t
Processed Data (for target use)
Processed Data (for target use)
Classified/Structured Data
Classified/Structured Data
Classified/Structured Data
Raw Data

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

- Referrals
- Control information
- Policy
- Data patterns

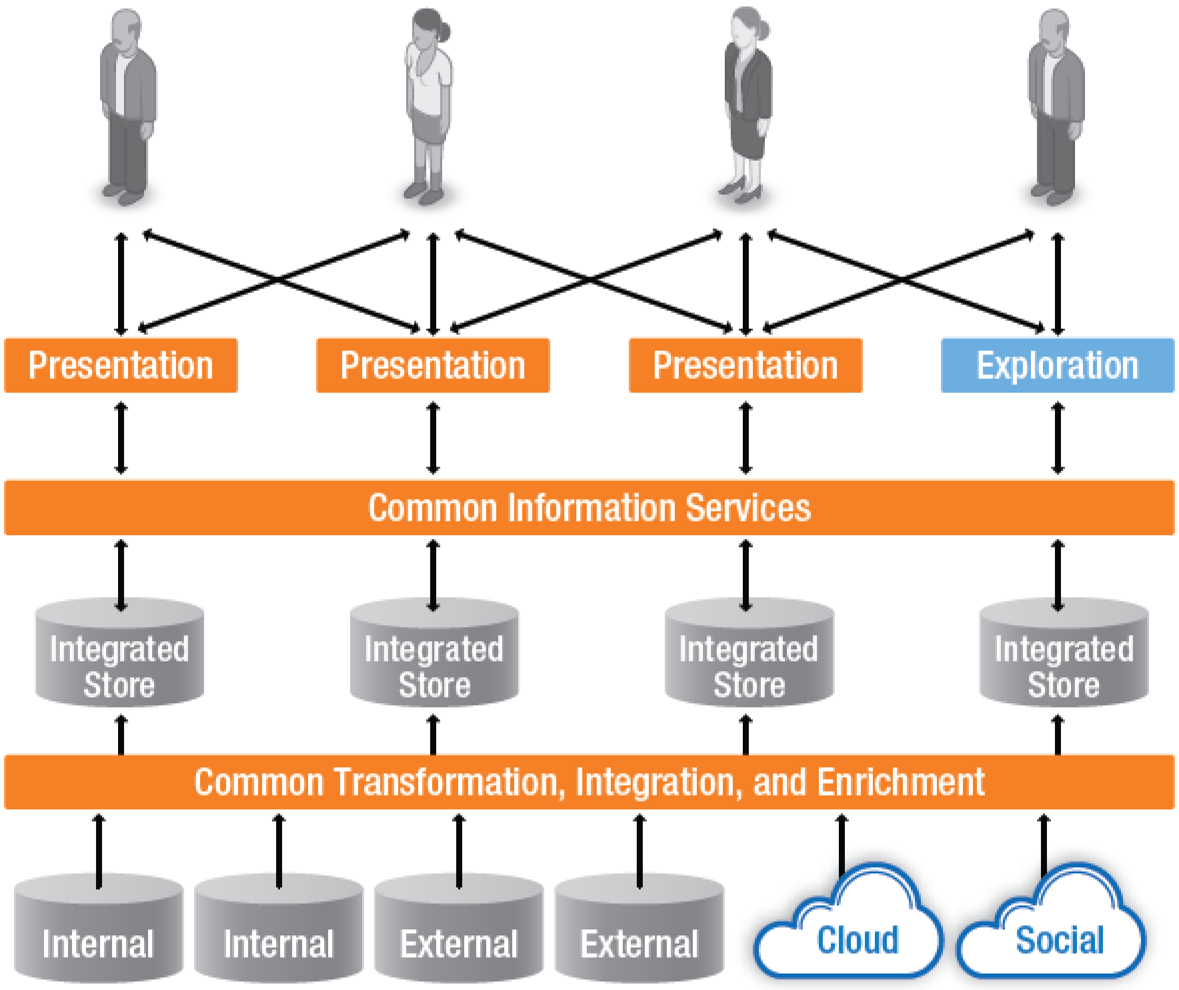
Additional Information

- Existing proposed Big Data architectures

Industry Initiatives to define Big Data (Architecture)

- Open Data Center Alliance (ODCA) Information as a Service (INFOaaS)
- TMF Big Data Analytics Reference Architecture
- Research Data Alliance (RDA)
  - All data related aspects, but not Infrastructure and tools
- LexisNexis HPCC Systems

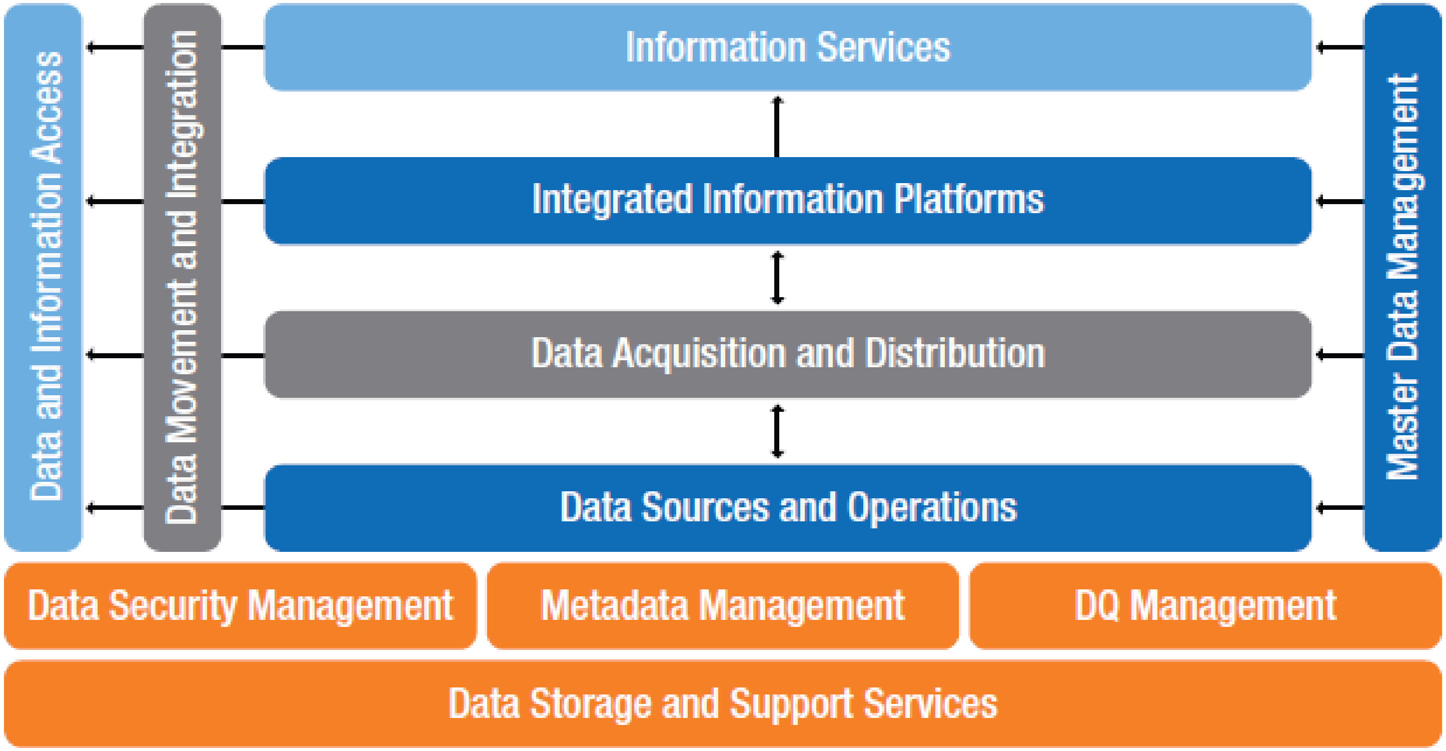
ODCA INFOaaS – Information as a Service



- Using integrated/unified storage
  - New DB/storage technologies allow storing data during all lifecycle

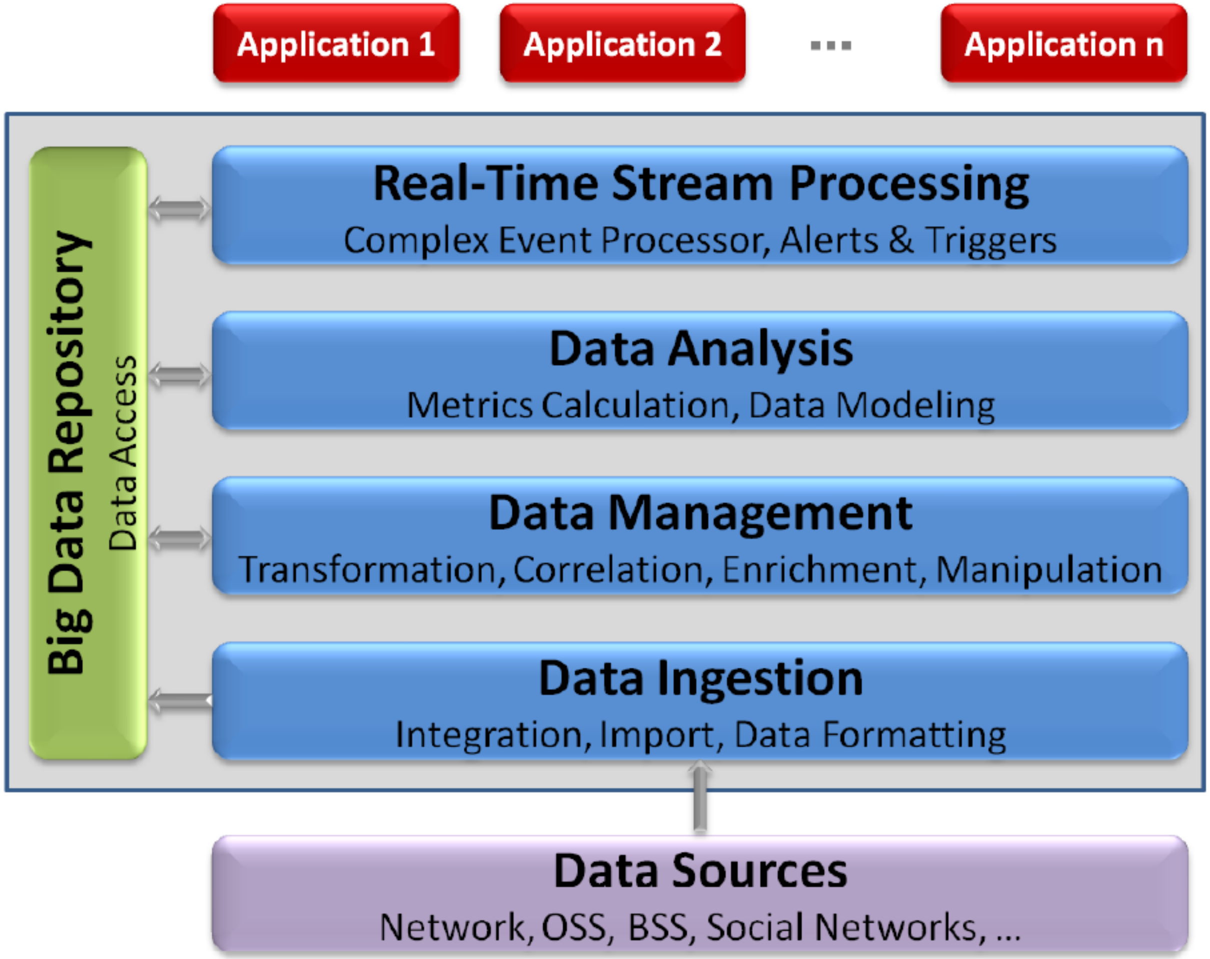
[ref] Open Data Center Alliance Master Usage model: Information as a Service, Rev 1.0. http://www.opendatacenteralliance.org/docs/Information_as_a_Service_Master_Usage_Model_Rev1.0.pdf

ODCA Example INFOaaS Architecture


Data and Information Access
Data Movement and Integration
Information Services
Integrated Information Platforms
Data Acquisition and Distribution
Data Sources and Operations
Master Data Management
Data Security Management
Metadata Management
DQ Management
Data Storage and Support Services

- Core Data and Information Components
- Data Integration and Distribution Components
- Presentation and Information Delivery Components
- Control and Support Components

TMF Big Data Analytics Architecture



Application 1

Application 2

...

Application n

Big Data Repository

Data Access

Real-Time Stream Processing
Complex Event Processor, Alerts & Triggers

Data Analysis
Metrics Calculation, Data Modeling

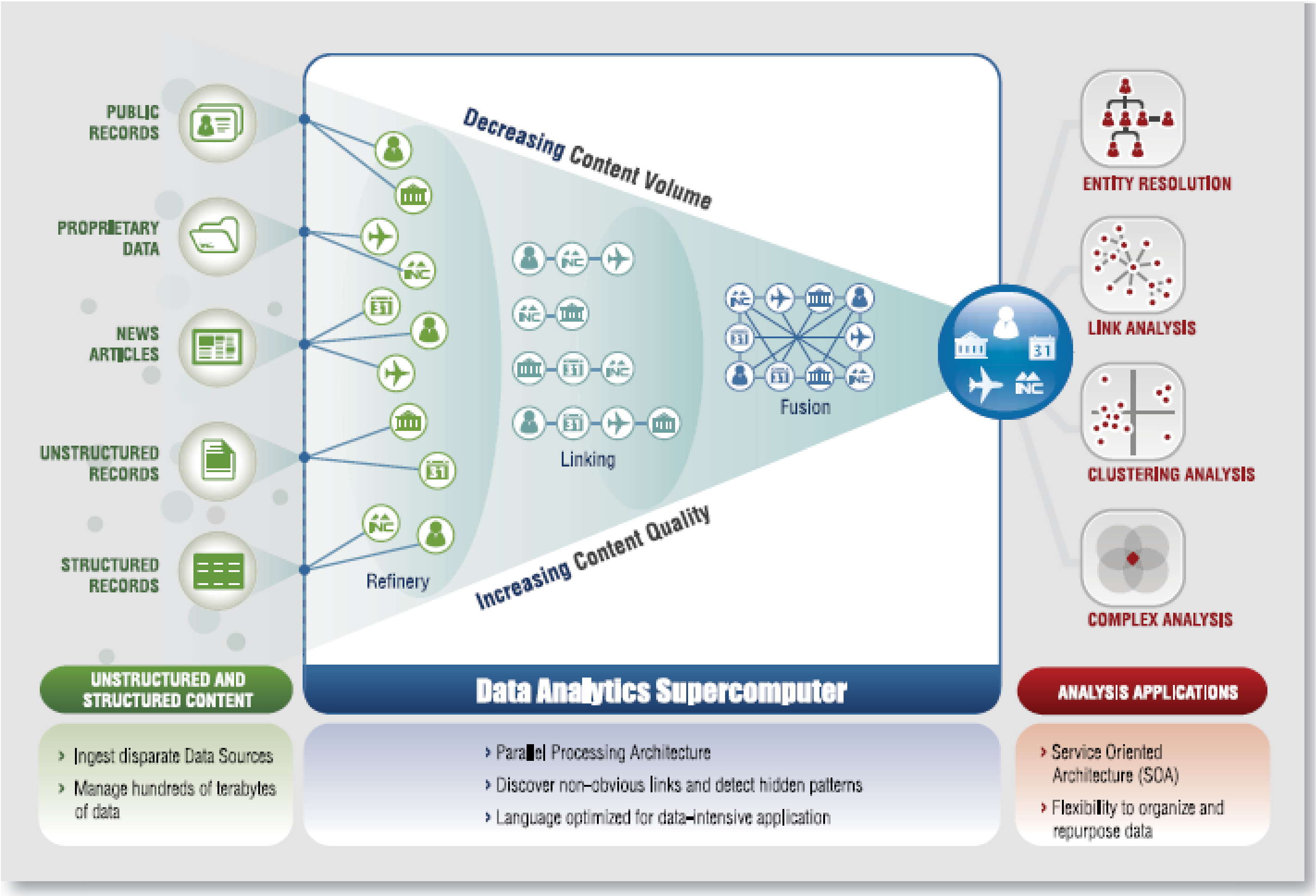
Data Management
Transformation, Correlation, Enrichment, Manipulation

Data Ingestion
Integration, Import, Data Formatting

Data Sources
Network, OSS, BSS, Social Networks, ...

[ref] TR202 Big Data Analytics Reference Model. Version 1.9, April 2013.

LexisNexis Vision for Data Analytics Supercomputer (DAS) [ref]



- Ingest disparate Data Sources
- Manage hundreds of terabytes of data

- Parallel Processing Architecture
- Discover non-obvious links and detect hidden patterns
- Language optimized for data-intensive application

- Service Oriented Architecture (SOA)
- Flexibility to organize and repurpose data

[ref] HPCC Systems: Introduction to HPCC (High Performance Computer Cluster), Author: A.M. Middleton, LexisNexis Risk Solutions, Date: May 24, 2011

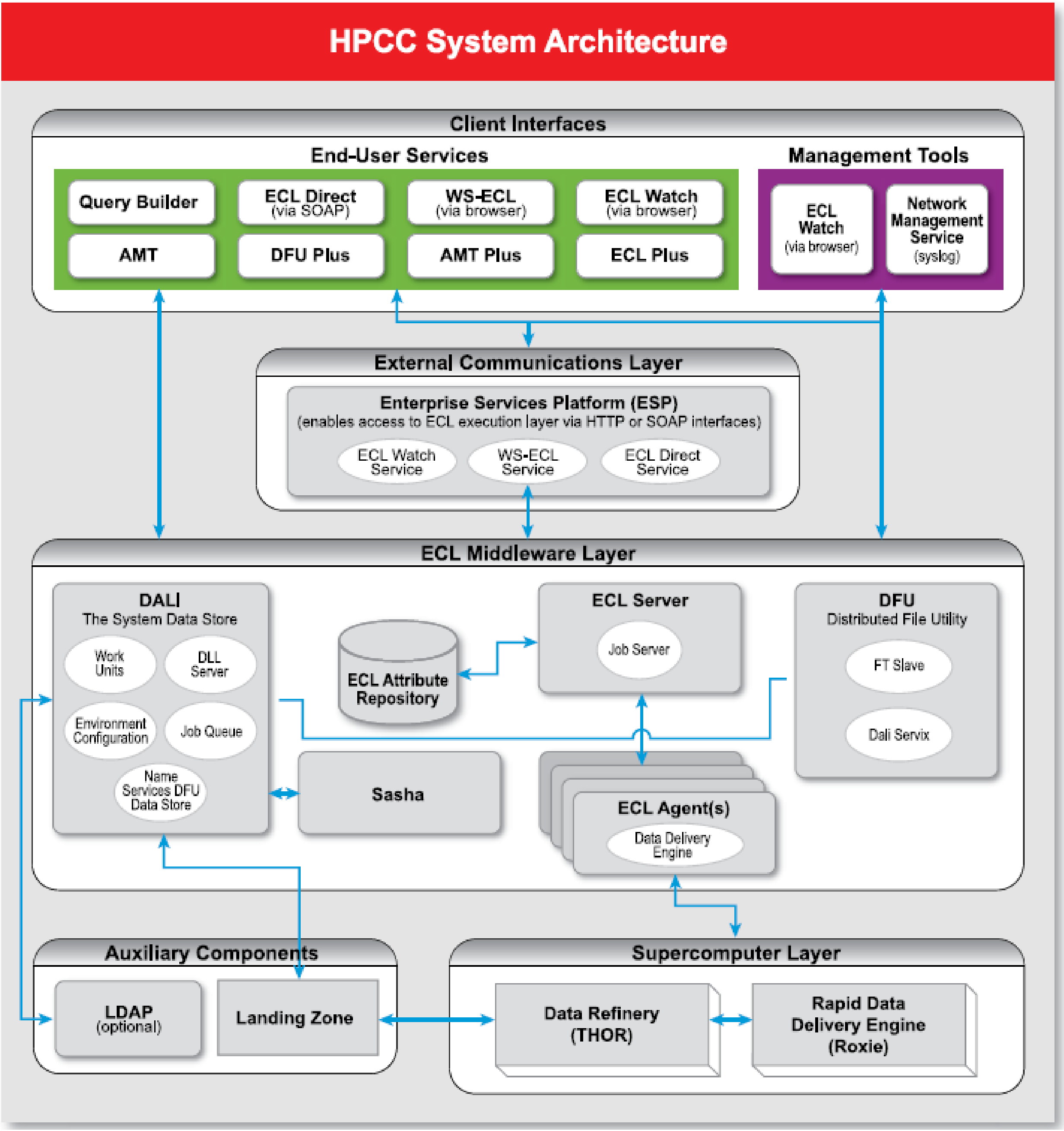
LexisNexis HPCC System Architecture

ECL – Enterprise Data Control Language

THOR Processing Cluster (Data Refinery)

Roxie Rapid Data Delivery Engine

[ref] HPCC Systems: Introduction to HPCC (High Performance Computer Cluster), Author: A.M. Middleton, LexisNexis Risk Solutions, Date: May 24, 2011



HPCC System Architecture

Client Interfaces

End-User Services

Query Builder

ECL Direct (via SOAP)

WS-ECL (via browser)

ECL Watch (via browser)

AMT

DFU Plus

AMT Plus

ECL Plus

Management Tools

ECL Watch (via browser)

Network Management Service (syslog)

External Communications Layer

Enterprise Services Platform (ESP)

(enables access to ECL execution layer via HTTP or SOAP interfaces)

ECL Watch Service

WS-ECL Service

ECL Direct Service

ECL Middleware Layer

DALI

The System Data Store

Work Units

DLL Server

Environment Configuration

Job Queue

Name Services DFU Data Store

ECL Attribute Repository

ECL Server

Job Server

DFU

Distributed File Utility

FT Slave

Dali Servix

Sasha

ECL Agent(s)

Data Delivery Engine

Auxiliary Components

LDAP (optional)

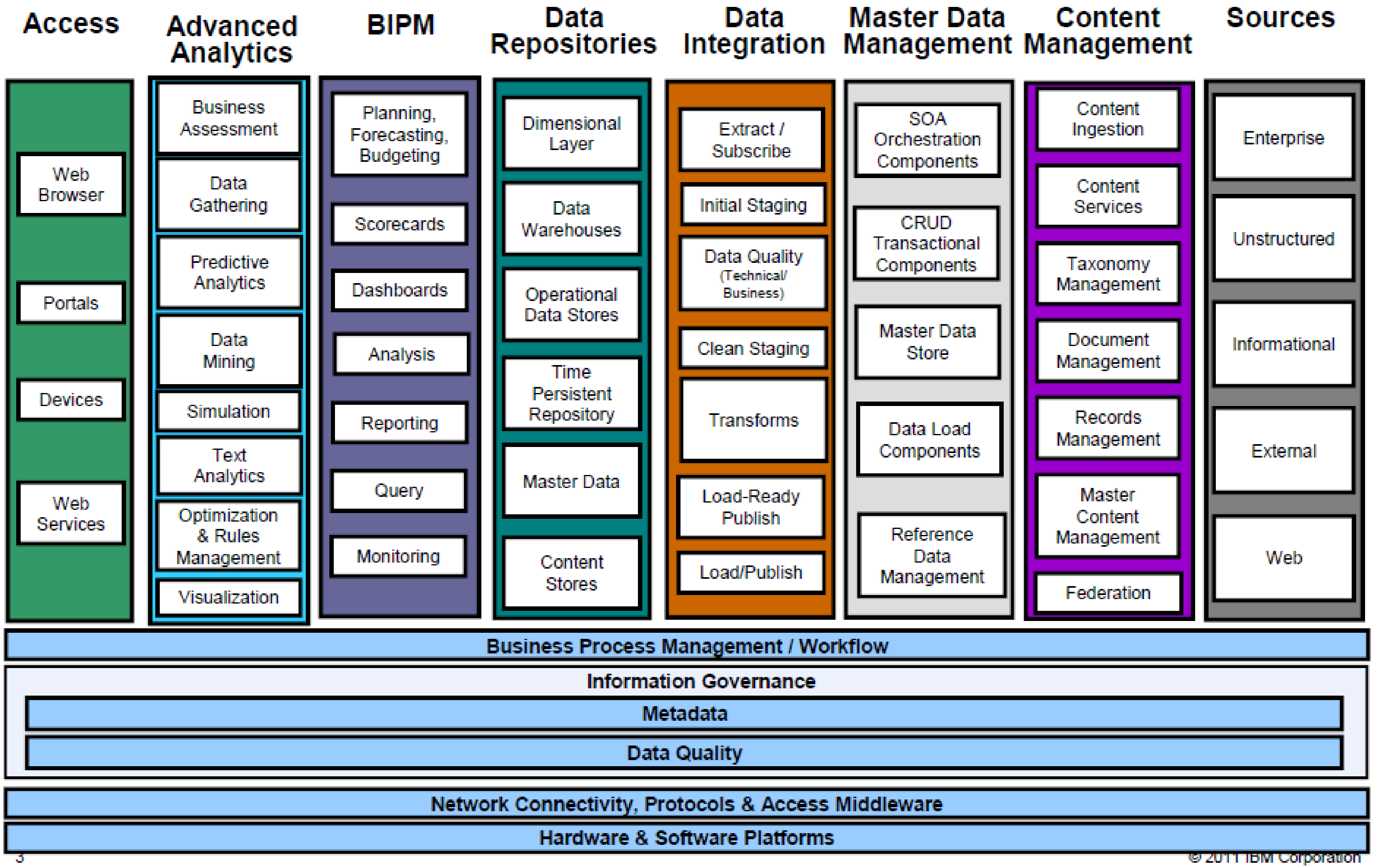
Landing Zone

Supercomputer Layer

Data Refinery (THOR)

Rapid Data Delivery Engine (Roxie)

The IBM Business Analytics and Optimization Reference Architecture Overview



Access
- Web Browser
- Portals
- Devices
- Web Services

Advanced Analytics
- Business Assessment
- Data Gathering
- Predictive Analytics
- Data Mining
- Simulation
- Text Analytics
- Optimization & Rules Management
- Visualization

BIPM
- Planning, Forecasting, Budgeting
- Scorecards
- Dashboards
- Analysis
- Reporting
- Query
- Monitoring

Data Repositories
- Dimensional Layer
- Data Warehouses
- Operational Data Stores
- Time Persistent Repository
- Master Data
- Content Stores

Data Integration
- Extract / Subscribe
- Initial Staging
- Data Quality (Technical/ Business)
- Clean Staging
- Transforms
- Load-Ready Publish
- Load/Publish

Master Data Management
- SOA Orchestration Components
- CRUD Transactional Components
- Master Data Store
- Data Load Components
- Reference Data Management

Content Management
- Content Ingestion
- Content Services
- Taxonomy Management
- Document Management
- Records Management
- Master Content Management
- Federation

Sources
- Enterprise
- Unstructured
- Informational
- External
- Web

Business Process Management / Workflow

Information Governance
- **Metadata**
- **Data Quality**

Network Connectivity, Protocols & Access Middleware

Hardware & Software Platforms

3

© 2011 IBM Corporation

IBM GBS Business Analytics and Optimisation (2011). https://www.ibm.com/developerworks/mydeveloperworks/files/basic/anonymous/api/library/48d92427-47d3-4e75-b54c-b6acfbd608c0/document/aa78f77c-0d57-4f41-a923-50e5c6374b6d/media&ei=yrknUbjMNM_liwKQhoCQBQ&usg=AFQjCNF_Xu6aifcAhlF4266xXNhKfKaTLw&sig2=j8JiFV_md5DnzfQl0spVrg&bvm=bv.42768644.d.cGE