Defining the Big Data Architecture Framework (BDAF)

Outcome of the Brainstorming Session at the University of Amsterdam

Yuri Demchenko (facilitator, reporter),
SNE Group, University of Amsterdam

17 July 2013, UvA, Amsterdam

Outline

- Big Data definition
  - 5 V's of Big Data: Volume, Velocity, Variety, Value, Veracity
  - Data Origin and Target
- From Big Data to All-Data – Paradigm change and New challenges
  - Big Data Infrastructure and Big Data Security
- Defining Big Data Architecture Framework (BDAF)
  - From Architecture to Ecosystem to Architecture Framework
  - Developments at NIST, ODCA, TMF, RDA
- Data Models and Big Data Lifecycle
- Big Data Infrastructure (BDI)
- Brainstorming: new features, properties, components, missing things, definition, directions

Big Data Research at SNE

- Focus on Infrastructure definition and services
  - Including Big Data Security
  - Software Defined Infrastructure based on Cloud/Intercloud technologies
- Papers published and submitted
  - ***Addressing Big Data Issues in Scientific Data Infrastructure***, by Demchenko, Y., P.Membrey, P.Grosso, C. de Laat.
    First International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2013). Part of The 2013 International Conference on Collaboration Technologies and Systems (CTS 2013), May 20-24, 2013, San Diego, California, USA
  - ***Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure***, by Y.Demchenko, P.Membrey, C.Ngo, C. de Laat, D.Gordijenko *Submitted* to Secure Data Management (SDM'13) Workshop. Part of VLDB2013 conference, 26-30 August 213, Trento, Italy
  - 科研信息化基础设施的大数据挑战 (***Big Data Challenges for e-Science Infrastructure***) by Demchenko, Y., Z.Zhao, P.Grosso, A.Wibisono, C. de Laat, In China Science and Technology Resources Review, Vol.45 No.1 30-35,40 Jan. 2013.

Big Data Architecture Framework (BDAF) - Proposed Context for the discussion

- Data Models, Structures, Types
  - Data formats, non/relational, file systems, etc.
- Big Data Management
  - Big Data Lifecycle (Management) Model
    - Big Data transformation/staging
  - Provenance, Curation, Archiving
- Big Data Analytics and Tools
  - Big Data Applications
    - Target use, presentation, visualisation
- Big Data Infrastructure (BDI)
  - Storage, Compute, (High Performance Computing,) Network
  - Big Data Operational support
- Big Data Security
  - Data security in-rest, in-move, trusted processing environments

Big Data Definition (1)

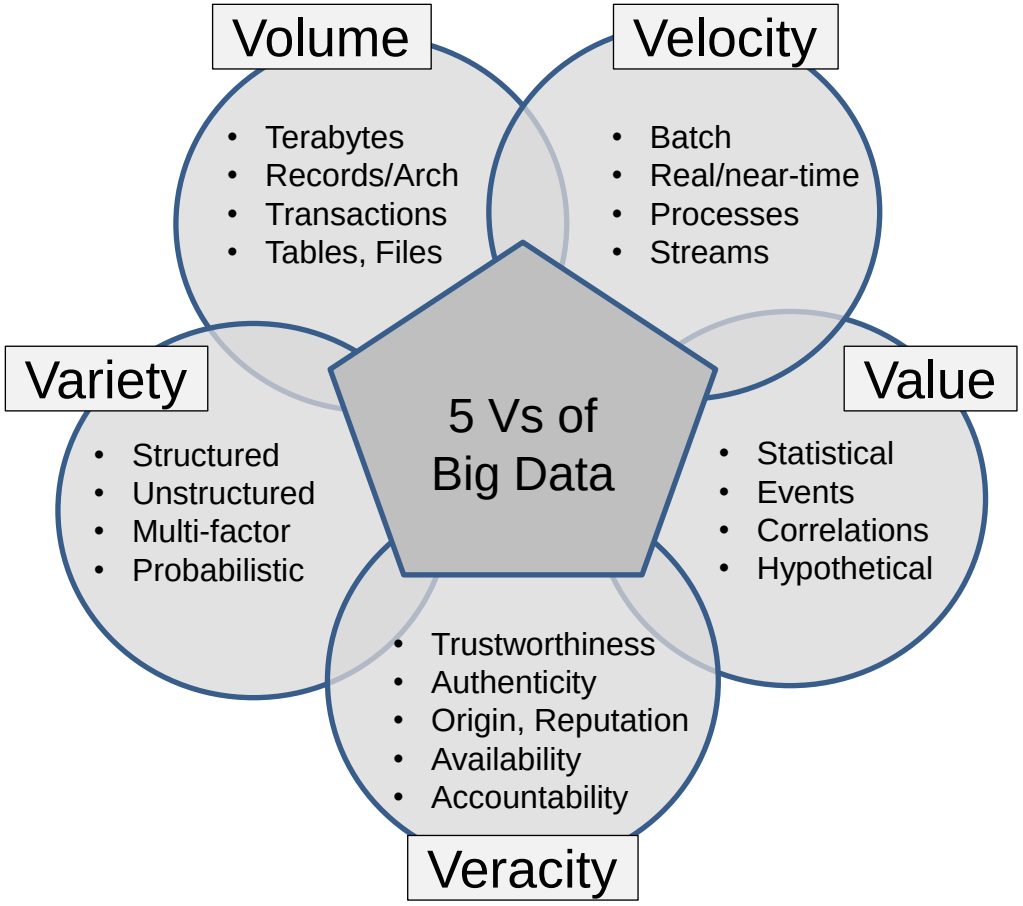
- IDC definition (conservative and strict approach) of Big Data:
  "A new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis"

- Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. Gartner, http://www.gartner.com/it-glossary/big-data/
  - Termed as 3 parts definition, not 3V definition
- Big Data: a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques.
  - From "The Big Data Long Tail" blog post by Jason Bloomberg (Jan 17, 2013). http://www.devx.com/blog/the-big-data-long-tail.html
- "Data that exceeds the processing capacity of conventional database systems. *The data is too big, moves too fast, or doesn't fit the structures of your database architectures*. To gain value from this data, you must choose an alternative way to process it."
  - Ed Dumbill, program chair for the O'Reilly Strata Conference

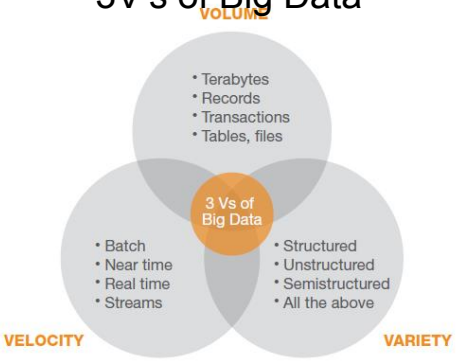
- Termed as the Fourth Paradigm *)
  "The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration." (Jim Gray, computer scientist)

*) *The Fourth Paradigm: Data-Intensive Scientific Discovery. Edited by Tony Hey, Stewart Tansley, and Kristin Tolle. Microsoft, 2009.*

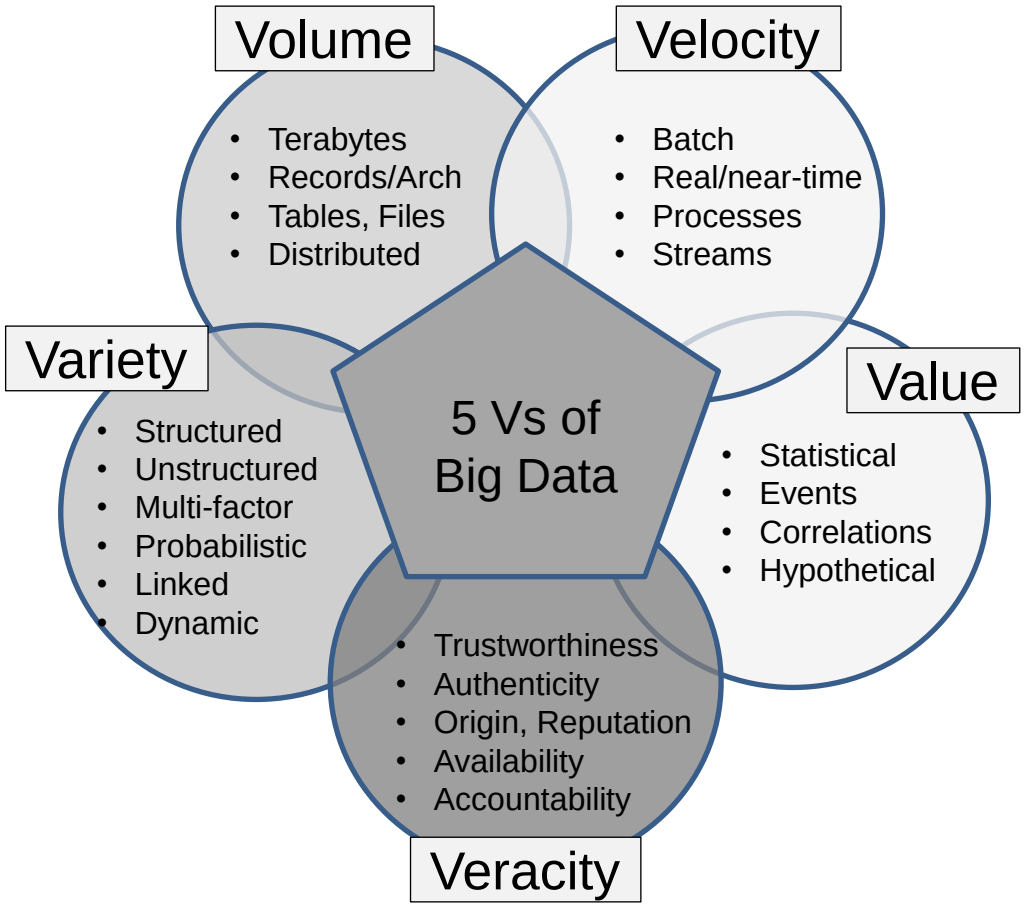
5 V's of Big Data



Commonly accepted 3V's of Big Data



Big Data Security: Veracity and other factors



- Trustworthiness and Reputation -> Integrity
- Origin, Authenticity and Identification
  - Identification both Data and Source
  - Source: system/domain and author
  - Data linkage (for complex hierarchical data, data provenance)
- Availability
  - Timeliness
  - Mobility (mobile/remote access; from other domain – roaming; federation)
- Accountability
  - As pro-active measure to ensure data veracity
- Data Dynamicity (i.e. Variability as 6th V)
  - As an additional property reflecting data change during their processing or lifecycle

Big Data Definition: From 5V 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

Overview: Technology Definitions and Timeline

- Service Oriented Architecture (SOA): First proposed in 1996 and revived with the Web Services advent in 2001-2002
  - Currently standard for industry, and widely used
  - Provided a conceptual basis for Web Services development
- Computer Grids: Initially proposed in 1998 and finally shaped in 2003 with the Open Grid Services Architecture (OGSA) by Open Grid Forum (OGF)
  - Currently remains as a collaborative environment
  - Migrates to cloud and inter-cloud platform
- Cloud Computing: Initially proposed in 2008
  - Defined new features, capabilities, operational/usage models and actually provided a guidance for the new technology development
  - Originated from the Service Computing domain and service management focused
- Big Data: Yet to be defined
  - Involves more components and processes to be included into the definition
  - Can be better defined as ecosystem where data are the main driving factor/component
  - Need to define the Big Data properties, expected technology capabilities and provide a guidance/vision for future technology development

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

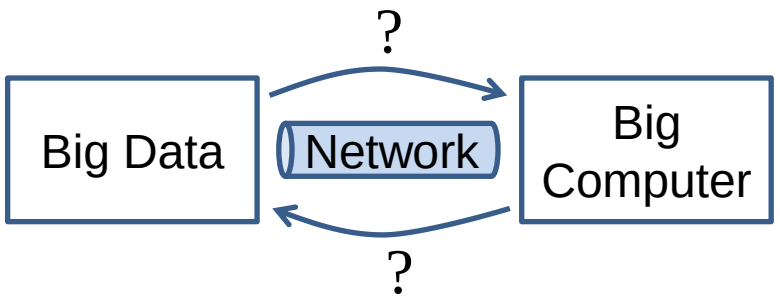
- Scientific discovery
- New technologies
- Manufacturing, processes, transport
- Personal services, campaigns
- Living environment support
- Healthcare support

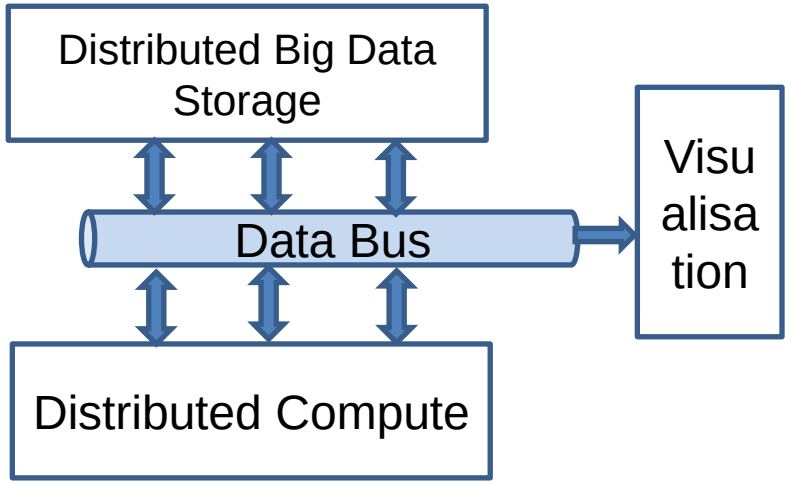
Big Data Nature: Origin and consumers (targets)

	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	+++	++	-	-	++	+++++

From Big Data to All-Data – Paradigm Change

- Really paradigm changing factor
  - Data storage and processing
  - Security
  - Identification and provenance
- Traditional model
  - BIG Storage and BIG computer with FAT pipe
  - Move compute to data *vs* Move data to compute
- New Paradigm
  - Continuous data production
  - Continuous data processing





Moving to Data-Centric Models and Technologies

- Current IT and communication technologies are host based or host centric
  - Any communication or processing is 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 security and access control
  - Data integrity and identifiability

Data Centric Security

- Paradigm shift to data centric security model
  - Previous and current security models are host or domain based
- New challenges and new security models
  - Data ownership
  - Data centric access control
    - Encryption enforced access control
  - Personally identified data, privacy, opacity
  - Trusted virtualisation platform
    - Dynamic trust bootstrapping

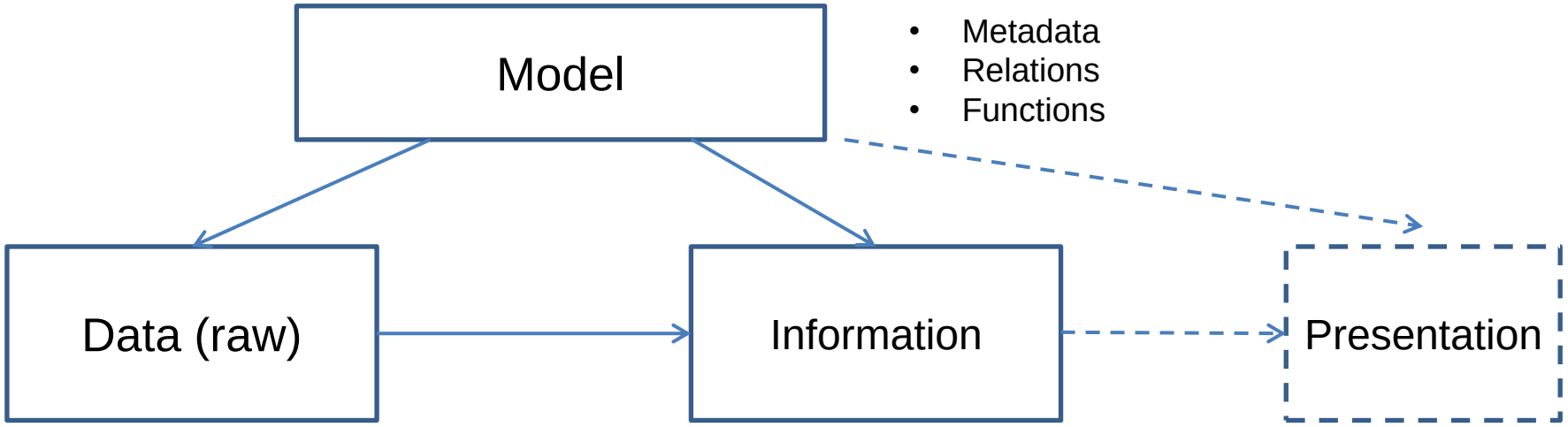
Defining Big Data Architecture Framework

- Existing attempts don't converge to something consistent: ODCA, TMF, NIST
  - See Appendix
- Architecture vs Ecosystem
  - Big Data undergo and number of transformation during their lifecycle
  - Big Data fuel the whole transformation chain
- Architecture vs Architecture Framework (Stack)
  - Separates concerns and factors
  - Architecture Framework components are inter-related

Missing Component – Data Model and Lifecycle

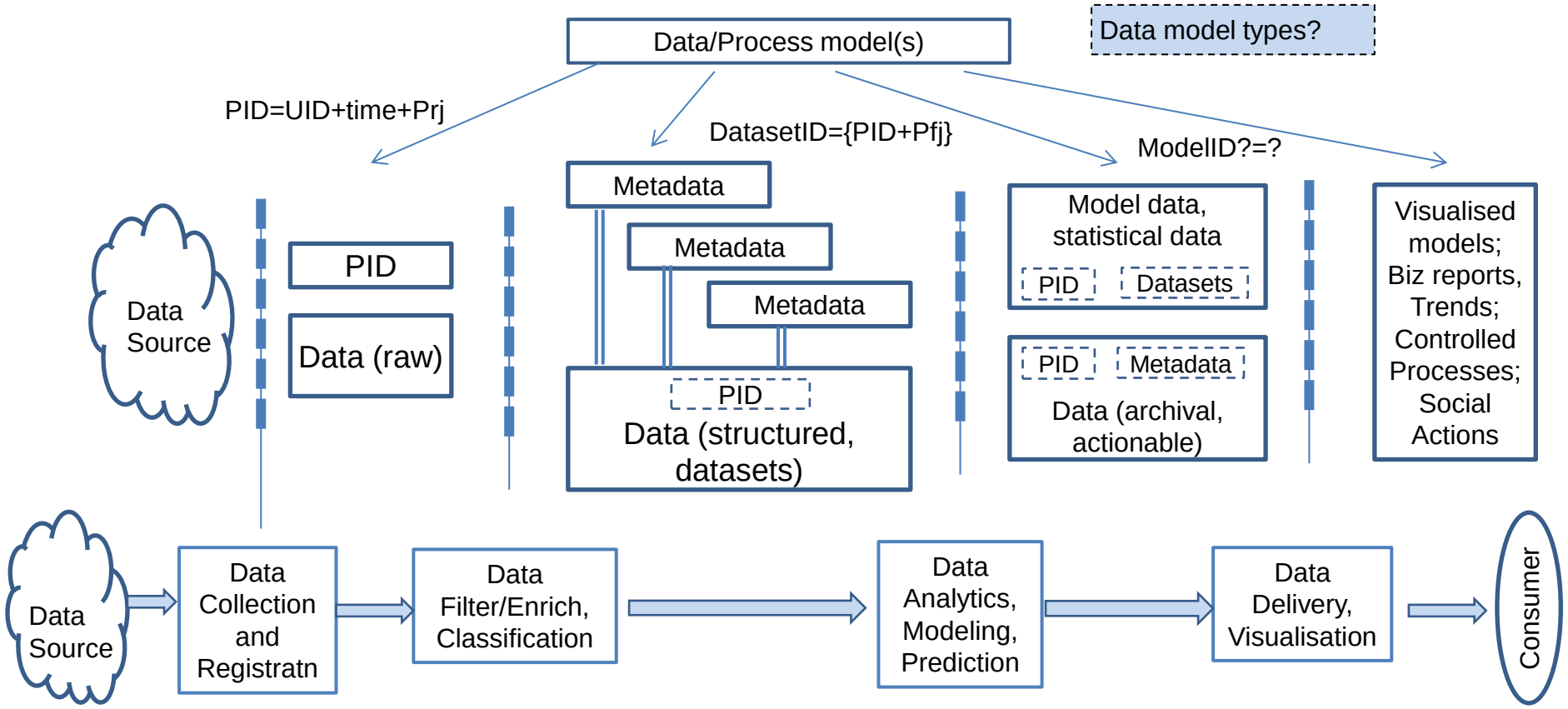
- Scientific Data and Scientific Data Lifecycle Management (SDLM) model
  - Preservation is an important issue
- General Big Data Lifecycle model
  - Actionable Data
  - Not necessary preservation is a key issue
  - Process control, actions, etc.

Data Model: Data and Information



- Data: The lowest layer of abstraction (?) from which information can be derived
- Information: A combination of contextualised data that can provide meaningful value or usage/action (scientific, business)
  - Actionable data
- Presentation (?)
- Where is knowledge (as a target of learning)?

Data Transformation Model



Security issues

- CIA and Access control
- Referral integrity
- Traceability
- Opacity

Big Data Architecture Framework (BDAF) – Target and Context for the discussion

- Data Models and Structures
  - Data types
- Big Data Lifecycle (Management) Model
  - Big Data transformation/staging
- Big Data Infrastructure (BDI)
  - Storage, Compute, (High Performance Computing,) Network
  - Sensor network, target/actionable devices
- Big Data Analytics/Tools
- Big Data Applications
  - Target use, actionable data, presentation, visualisation
- Big Data Management/Operation
  - Provenance, Curation, Archiving, Operational support
- Big Data Security
  - Data Security in-rest, in-move, trusted processing environments

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

Col: Used By Row: Requires This	Data Models Structrs	Data Lifecycle	BigData Infrastr	BigData Analytics	BigData Aplicatn	BigData Mangnt Operatn	BigData Security
Data Models		+++	++	+++	+++	+++	+++
Data Lifecycle	+++		+++	++	++	+++	+++
BigData Infastruct	+++	+++		++	++	+++	+++
BigData Analytics	+++	+	++		+++	+	++
BigData Applicatn	++	+	+++	++		++	++
BigData Mangnt	+++	+++	+++	+	++		+++
BigData Security	+++	+++	+++	+	+	++	

Big Data Architecture Framework (BDAF) – Aggregated (1)

- (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)

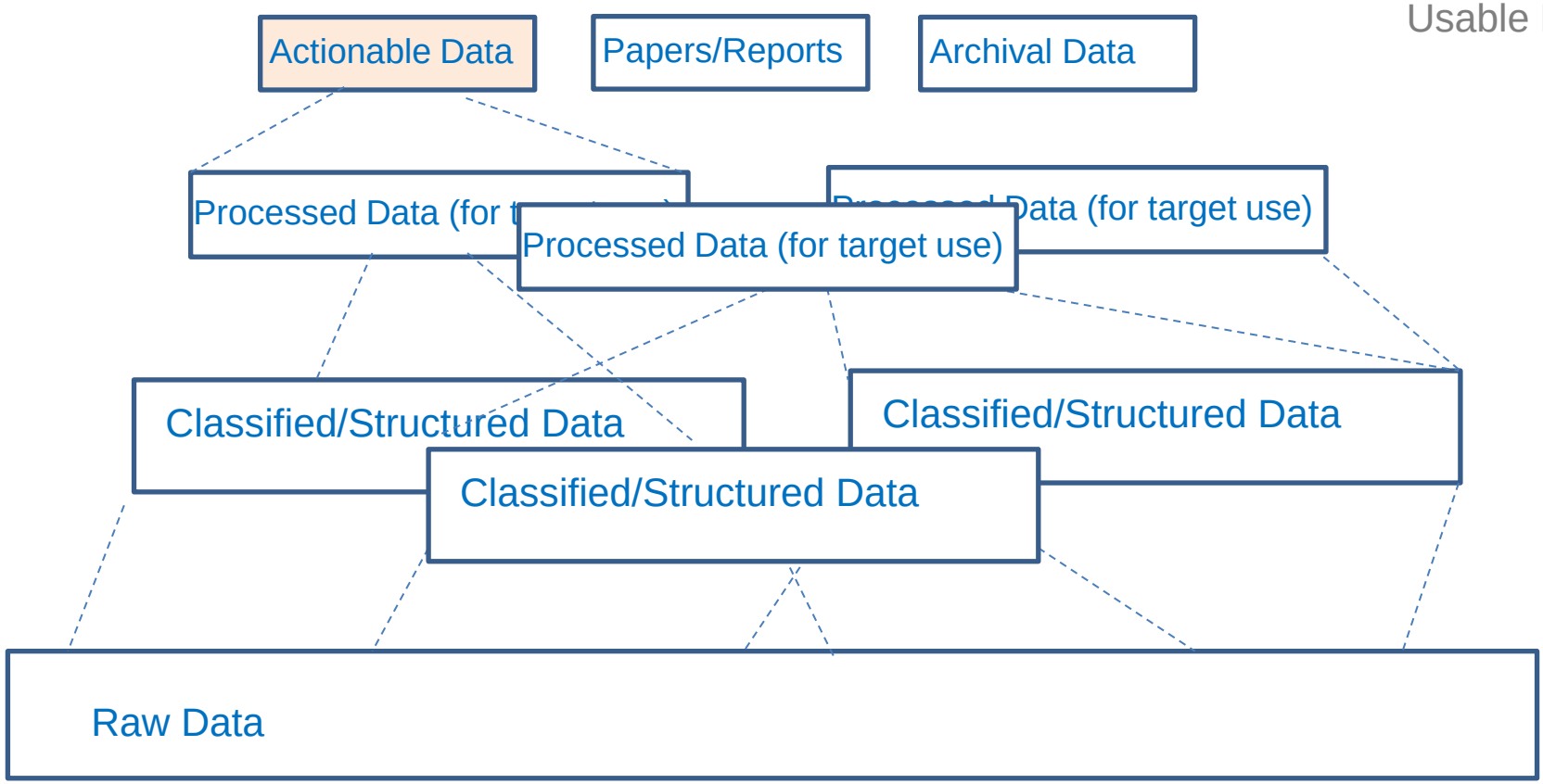
Col: Used By Row: Requires This	Data Models Structrs	Data Managmnt & Lifecycle	BigData Infrastr & Operations	BigData Analytics & Applicatn	BigData Security
Data Models & Structures		+	++	+	++
Data Managmnt & Lifecycle	++		++	++	++
BigData Infrastruct & Operations	+++	+++		++	+++
BigData Analytics & Applications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

Data Models, Structure, Types

- Data structures
  - Structured data
  - Unstructured data
- Data types [ref]
  - (a) data described via a formal data model
  - (b) data described via a formalized grammar
  - (c) data described via a standard format
  - (d) arbitrary textual or binary data
- Data models
  - Depend on target/goal, or process/object?
  - Evolve or chain/stack?

[ref] NIST Big Data WG discussion http://bigdatawg.nist.gov/home.php

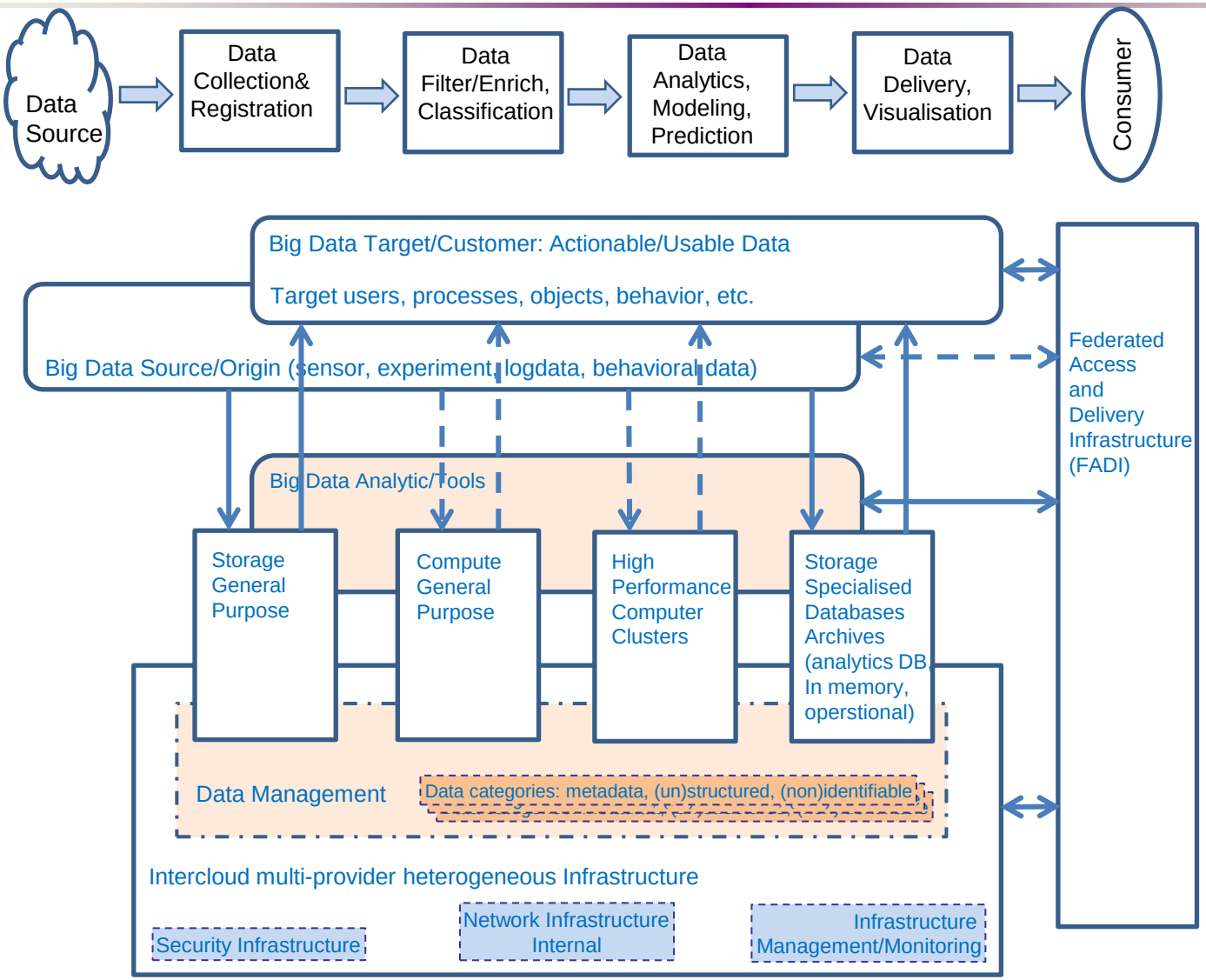
Evolutional/Hierarchical Data Model



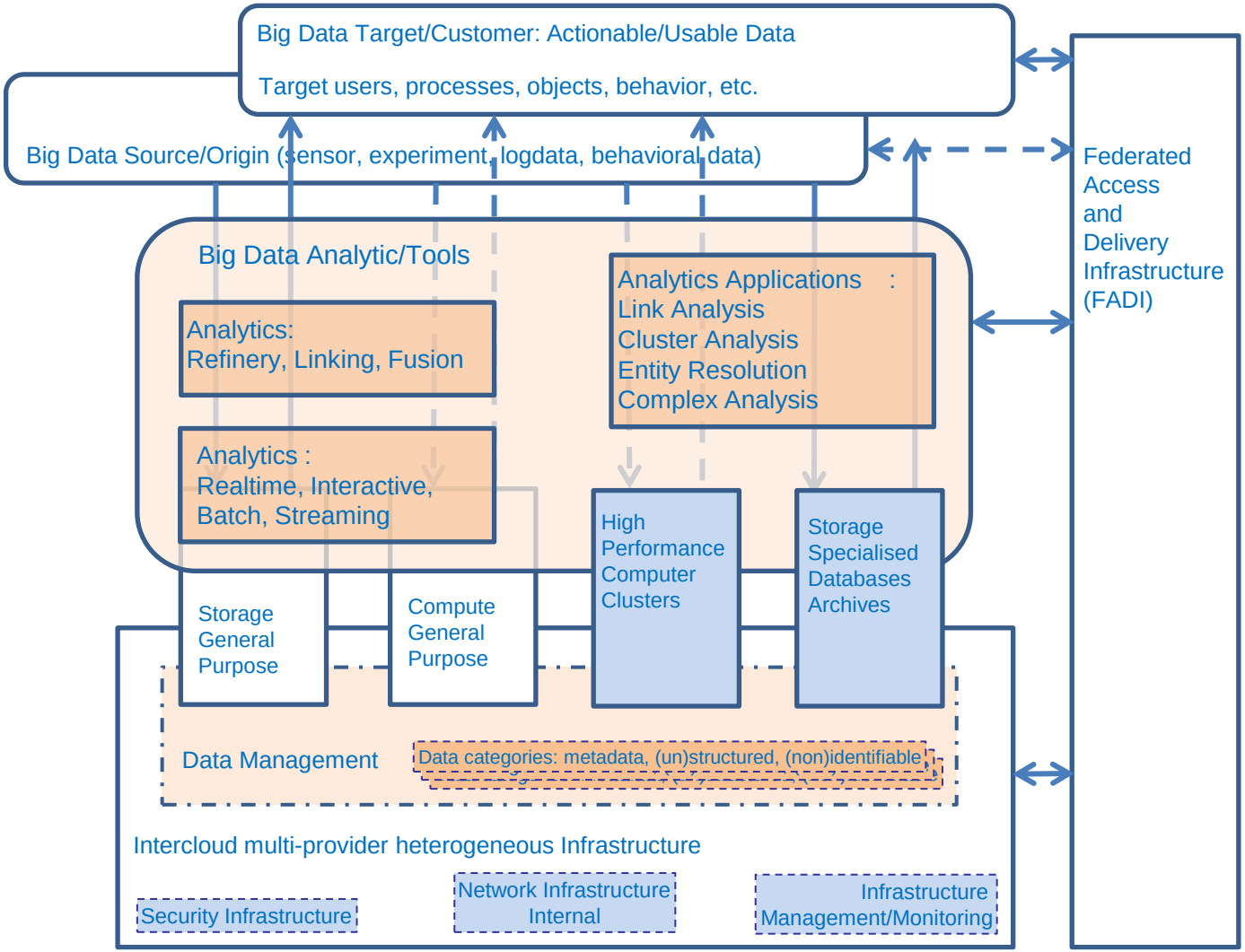
- Common Data Model?
- Data interlinking?
- Fits to Graph data type?
- Metadata

- Referrals
- Control information
- Policy
- Data patterns

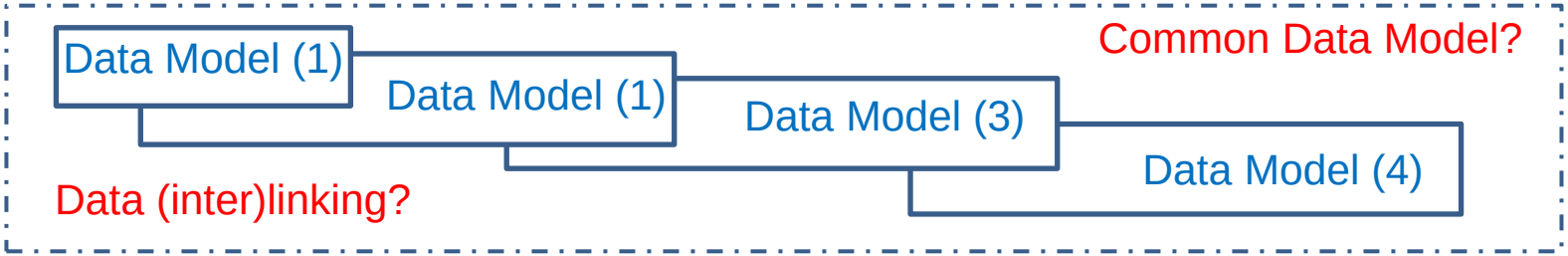
Big Data Ecosystem: Data, Lifecycle, Infrastructure

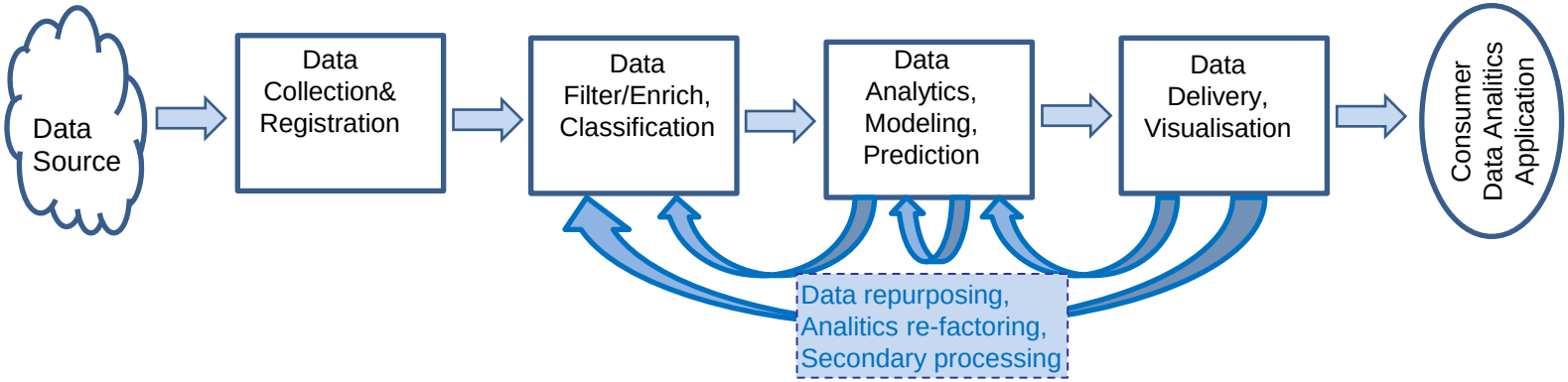


Big Data Infrastructure and Analytic Tools



Data Transformation/Lifecycle Model





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

Data Stored on the Big Data Infrastructure

- Plain, distributed, hierarchical, relational, graph data
  - Streaming data (?)
- Protected data
  - Encrypted data
  - Masked data (scrambled, padded, manipulated, etc.)
  - Anonymised and privacy enhanced
  - Identifiable and non-identifiable
  - Policy attached/enforced
- Tiered/auto-tired

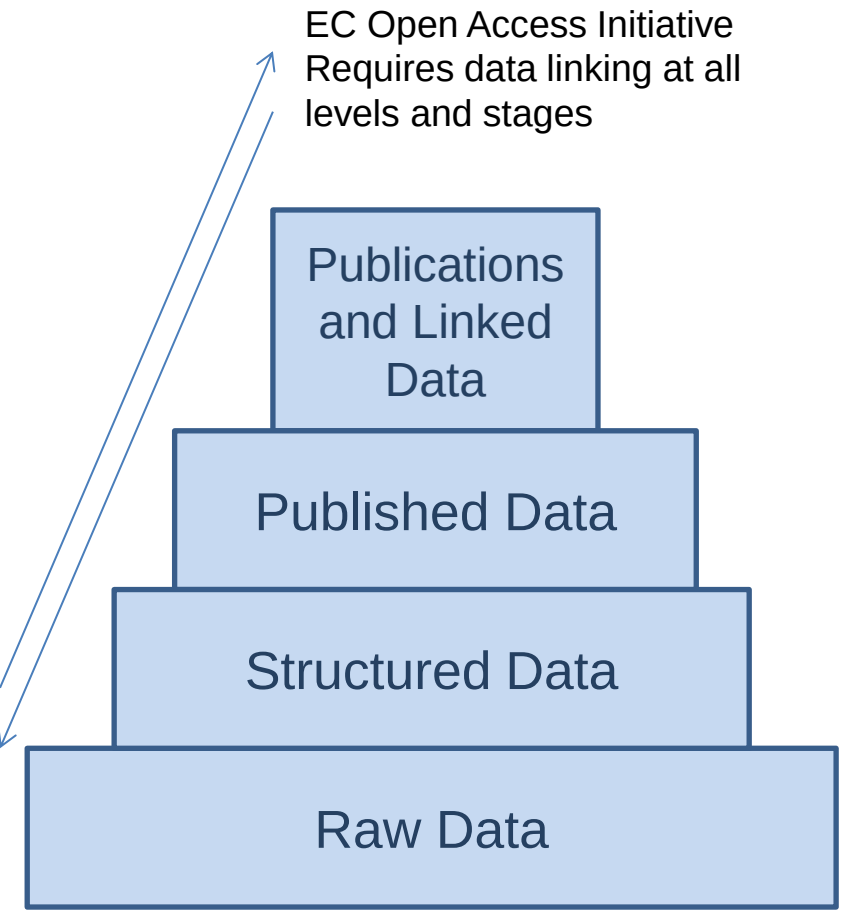
Gap Analysis and Requirements to Big Data Technologies

- Based on the collection of use cases analysis
- To validate the Big Data definition and Big Data Architecture Framework definition
- To be defined in a technology agnostic way
  - Done for the required capabilities, not selected technologies

Big Data and Data Intensive Science

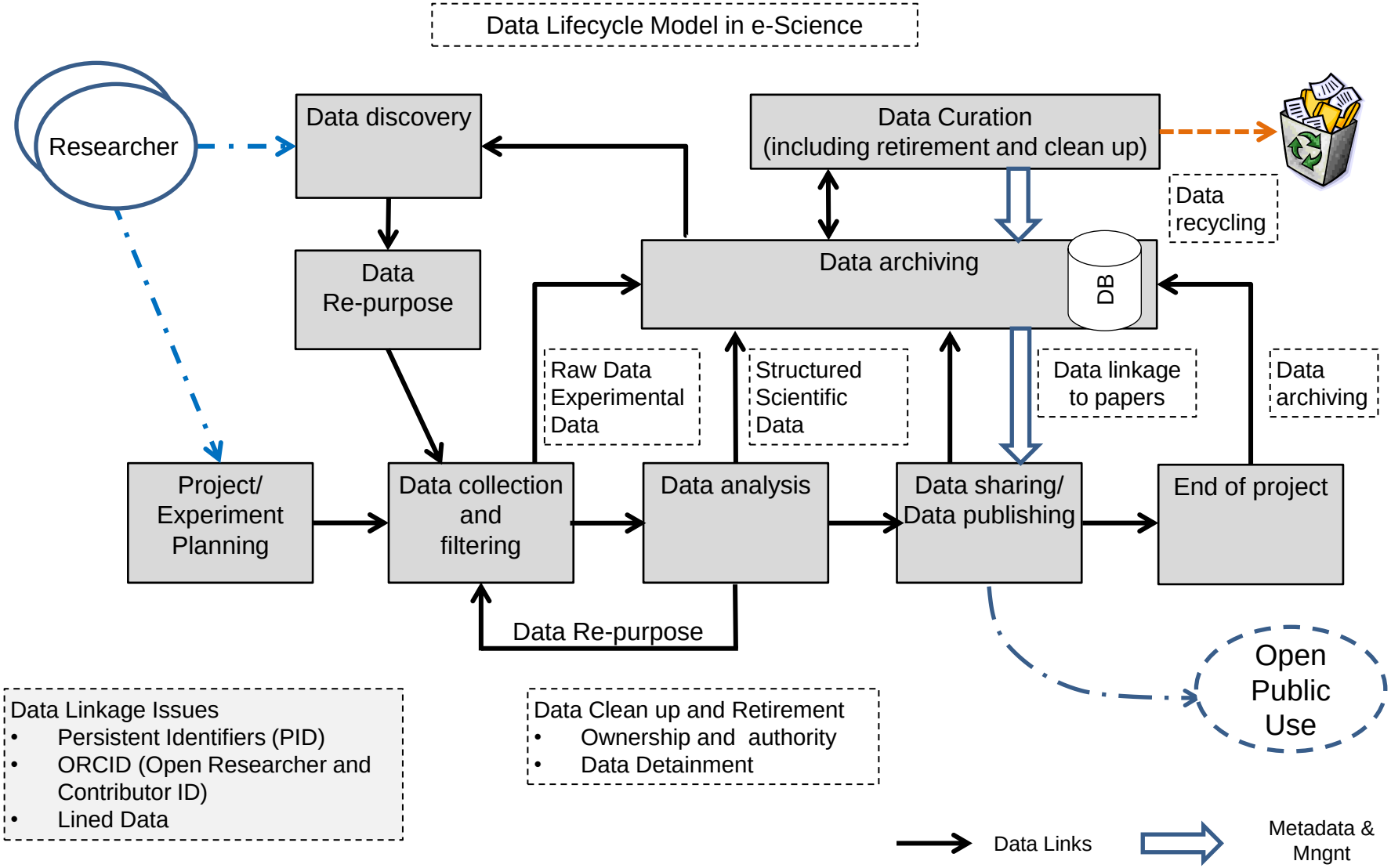
- Scientific Data types
- Scientific Data Lifecycle Management (SDLM)
- Scientific Data Infrastructure (SDI)

Scientific Data Types



- **Raw data** collected from observation and from experiment (according to an initial research model)
- **Structured data** and datasets that went through data filtering and processing (supporting some particular formal model)
- **Published data** that supports one or another scientific hypothesis, research result or statement
- **Data linked to publications** to support the wide research consolidation, integration, and openness.

Scientific Data Lifecycle Management (SDLM) Model



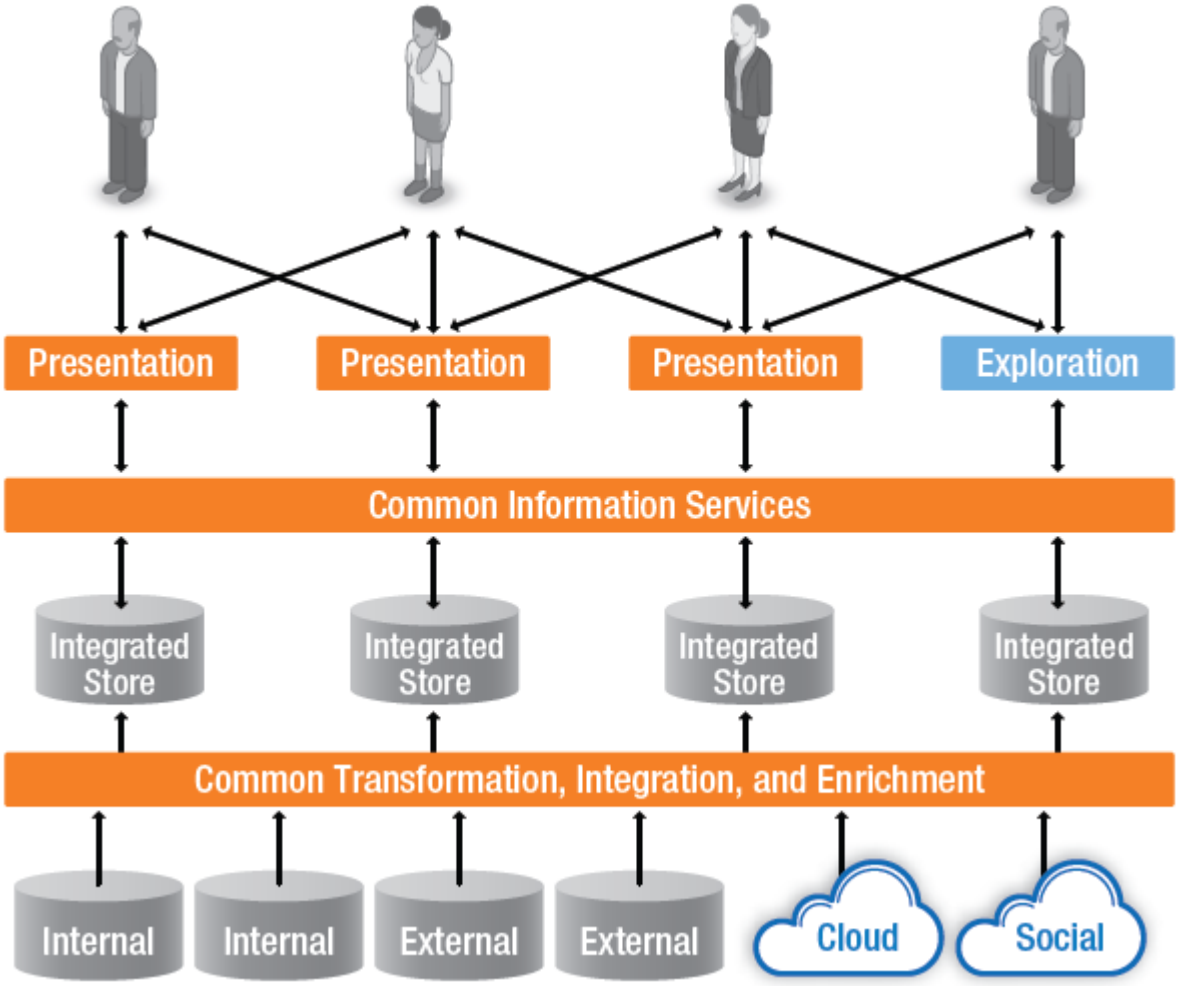
Additional Information

- Existing proposed Big Data architectures
- e-Science and Scientific Data Infrastructure (SDI)
- Cloud computing as a platform for SDI

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
- NIST Big Data Working Group (NBD-WG)
  - Range of activities

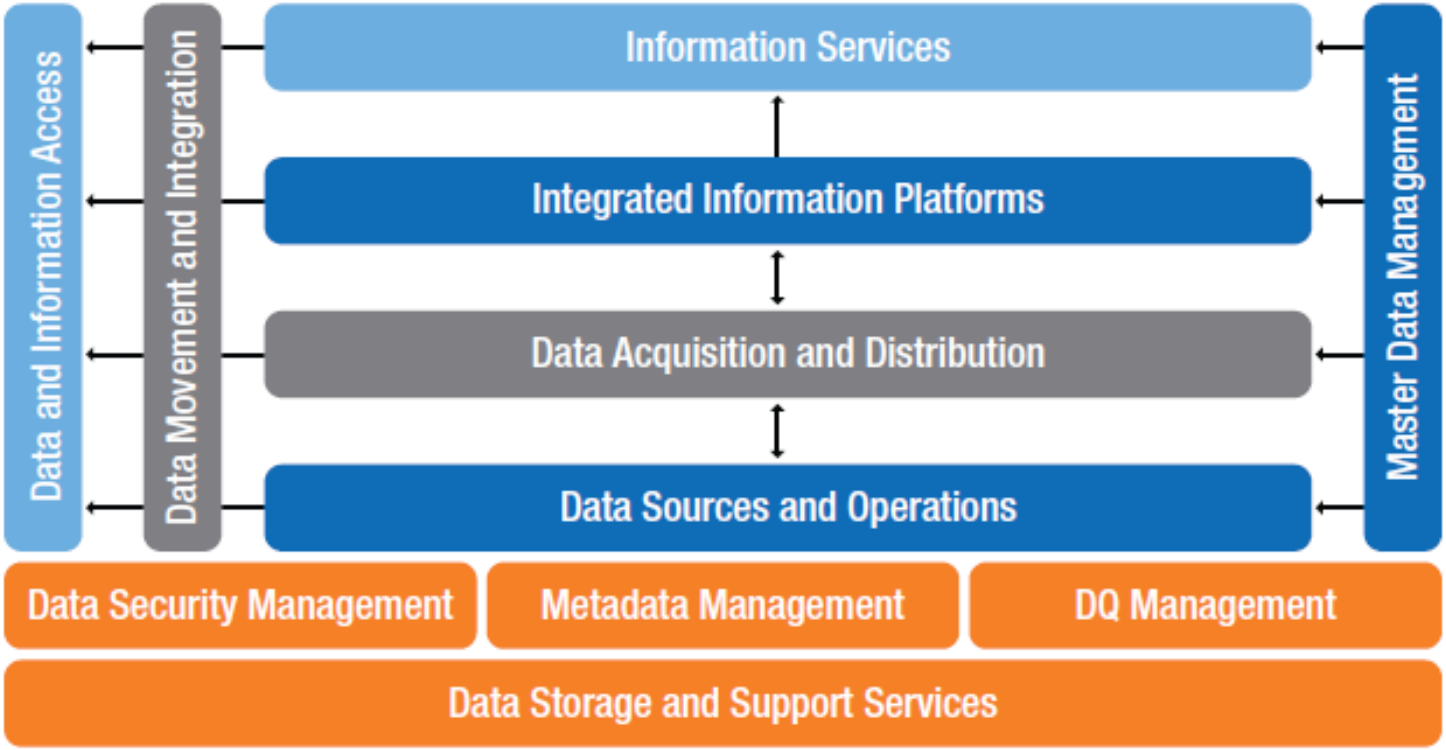
ODCA INFOaaS – Information as a Service



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

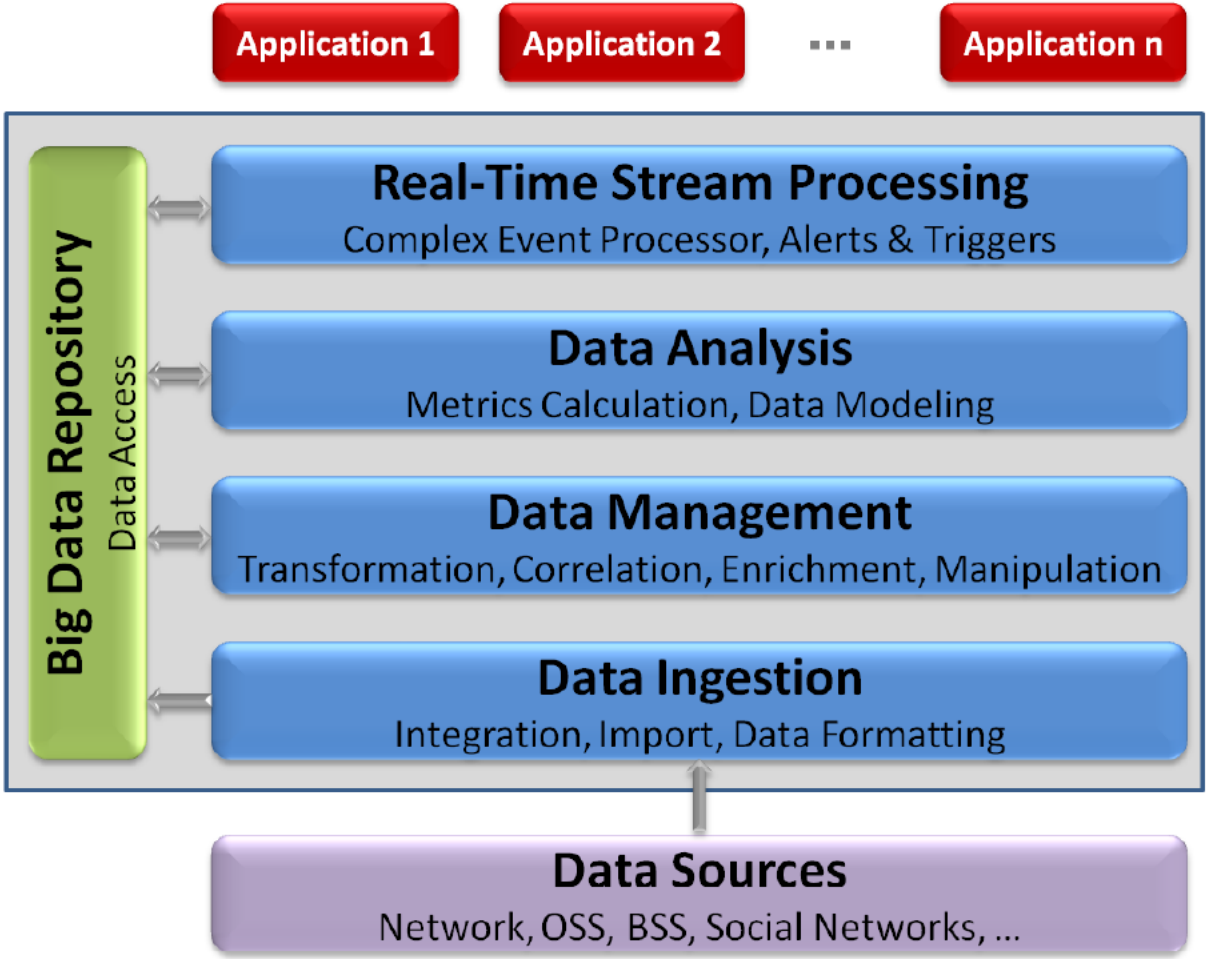
[ref] Open Data Center Alliance Master Usage model: Information as a Service, Rev 1.0. http://www.opendatacenteralliance.org/docs/Information_as_a_Service_Master_Usage_Model_Rev1.0.pdf

ODCA Example INFOaaS Architecture



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

TMF Big Data Analytics Architecture

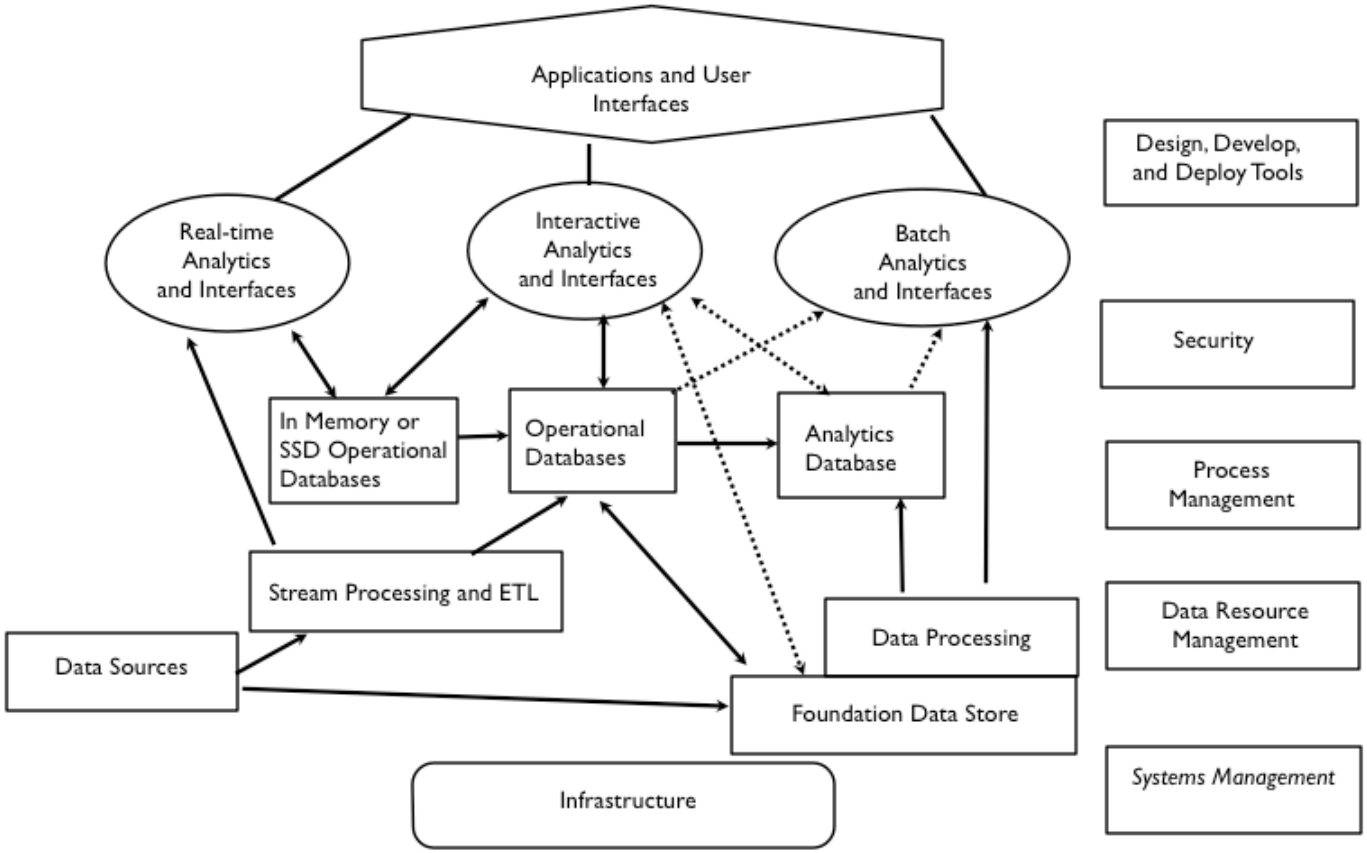


[ref] TR202 Big Data Analytics Reference Model. Version 1.9, April 2013.

NIST Big Data Working Group (NBD-WG)

- Deliverables target – September 2013
- Activities: Conference calls every day 17-19:00 (CET) by subgroup - http://bigdatawg.nist.gov/home.php
  - Big Data Definition and Taxonomies
  - Requirements (chair: Jeffrey Fox)
  - Big Data Security
  - Reference Architecture
  - Technology Roadmap
- BigdataWG mailing list and useful documents
  - Input documents http://bigdatawg.nist.gov/show_InputDoc2.php
  - Brainstorming summary and Lessons learnt (from brainstorming) http://bigdatawg.nist.gov/_uploadfiles/M0010_v1_6762570643.pdf
  - Big Data Ecosystem Reference Architecture (Microsoft) http://bigdatawg.nist.gov/_uploadfiles/M0015_v1_1596737703.docx

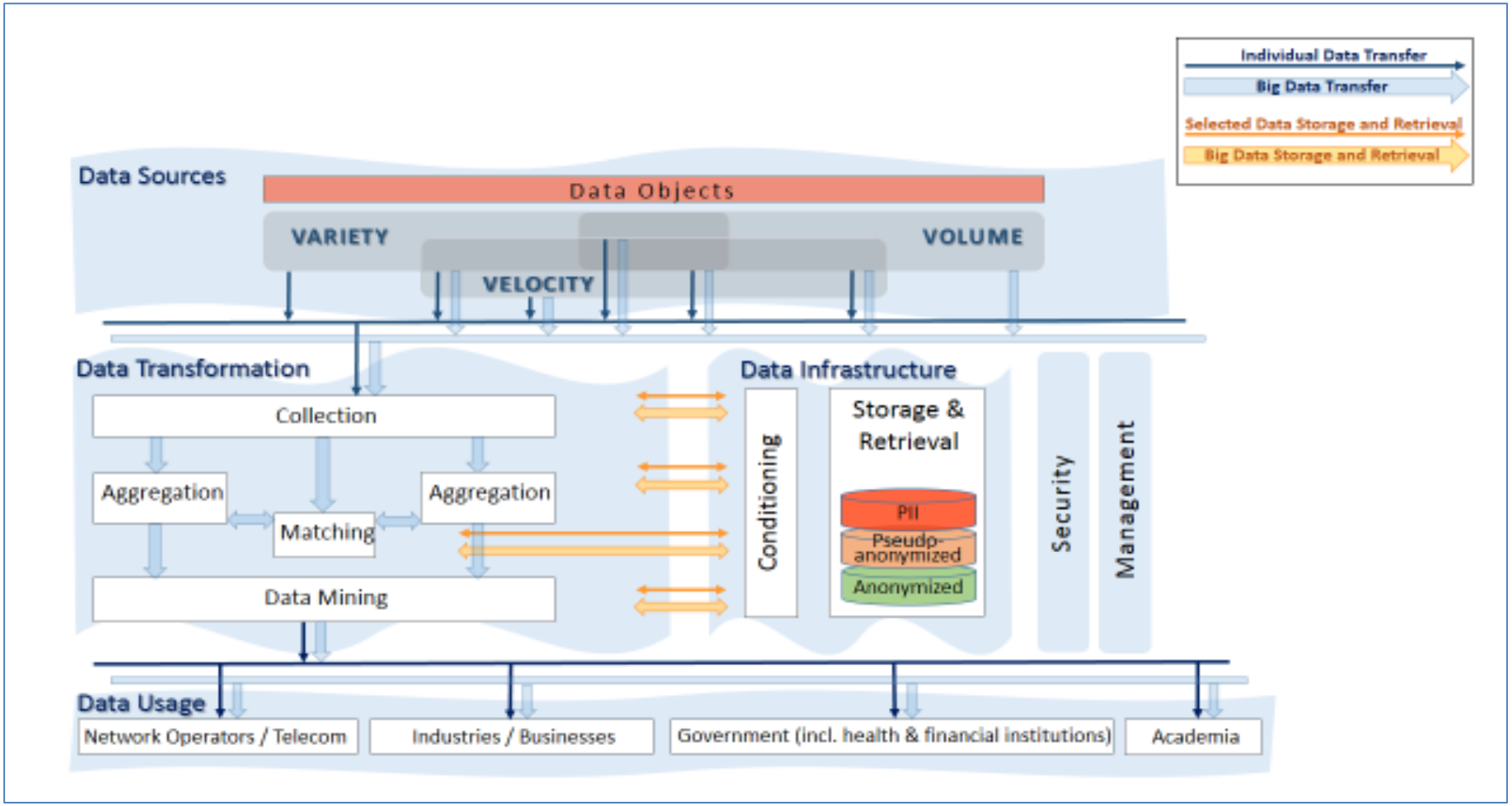
NIST Proposed Reference Architecture



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

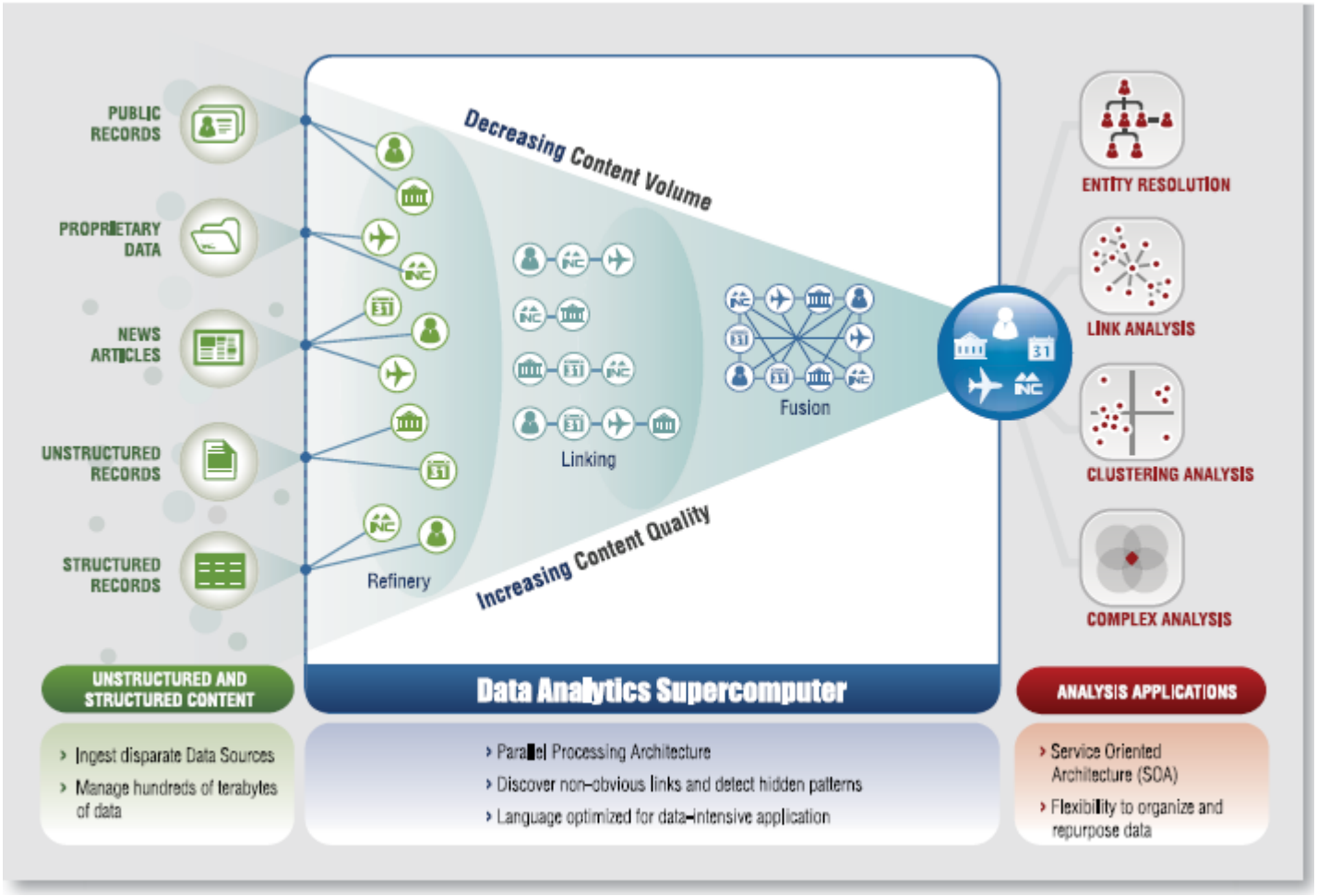
[ref] NIST Big Data WG mailing list discussion
http://bigdatawg.nist.gov/_uploadfiles/M0010_v1_6762570643.pdf

Big Data Ecosystem Reference Architecture (By Microsoft) [ref]



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

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



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

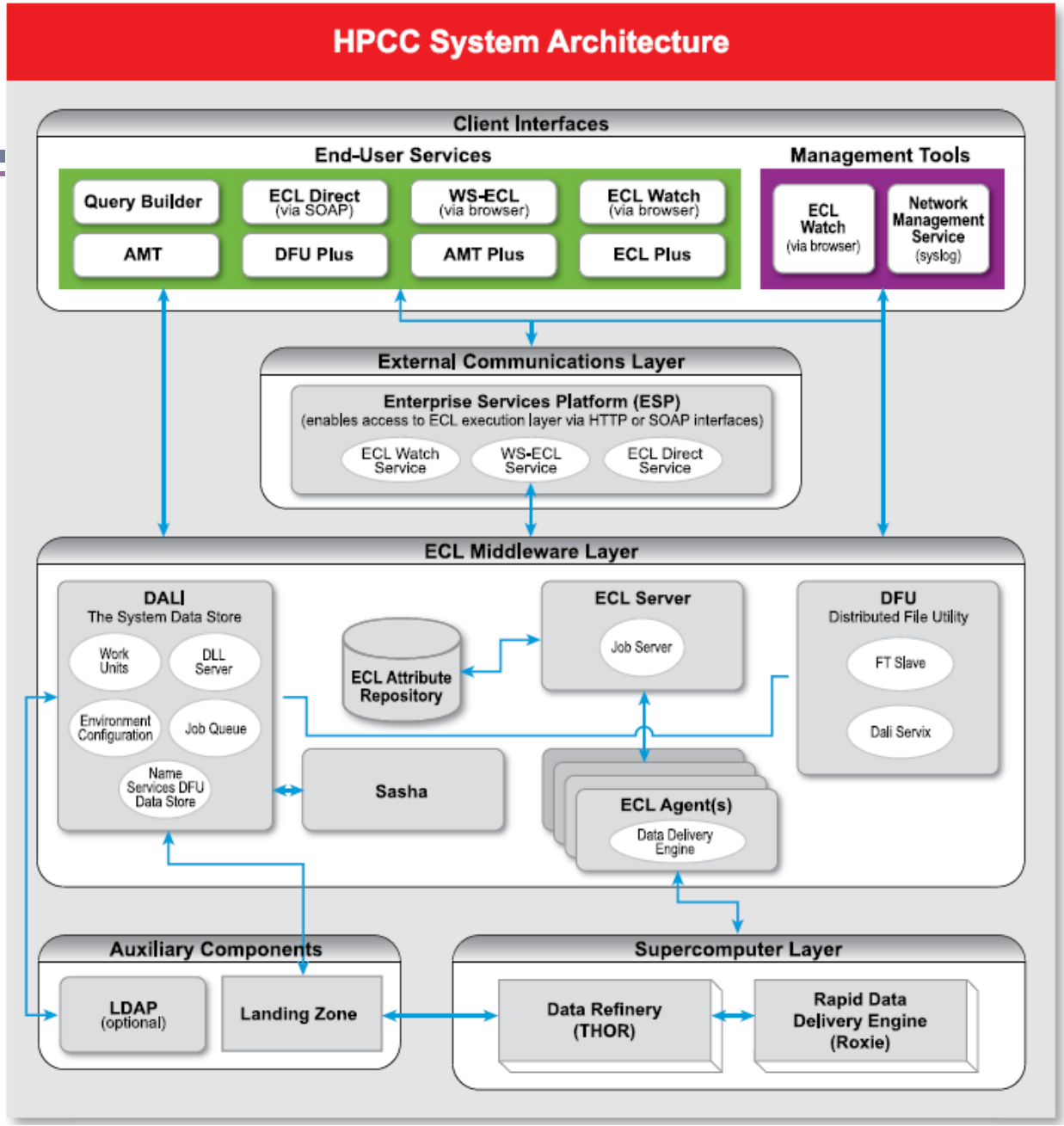
LexisNexis HPCC System Architecture

ECL – Enterprise Data Control Language

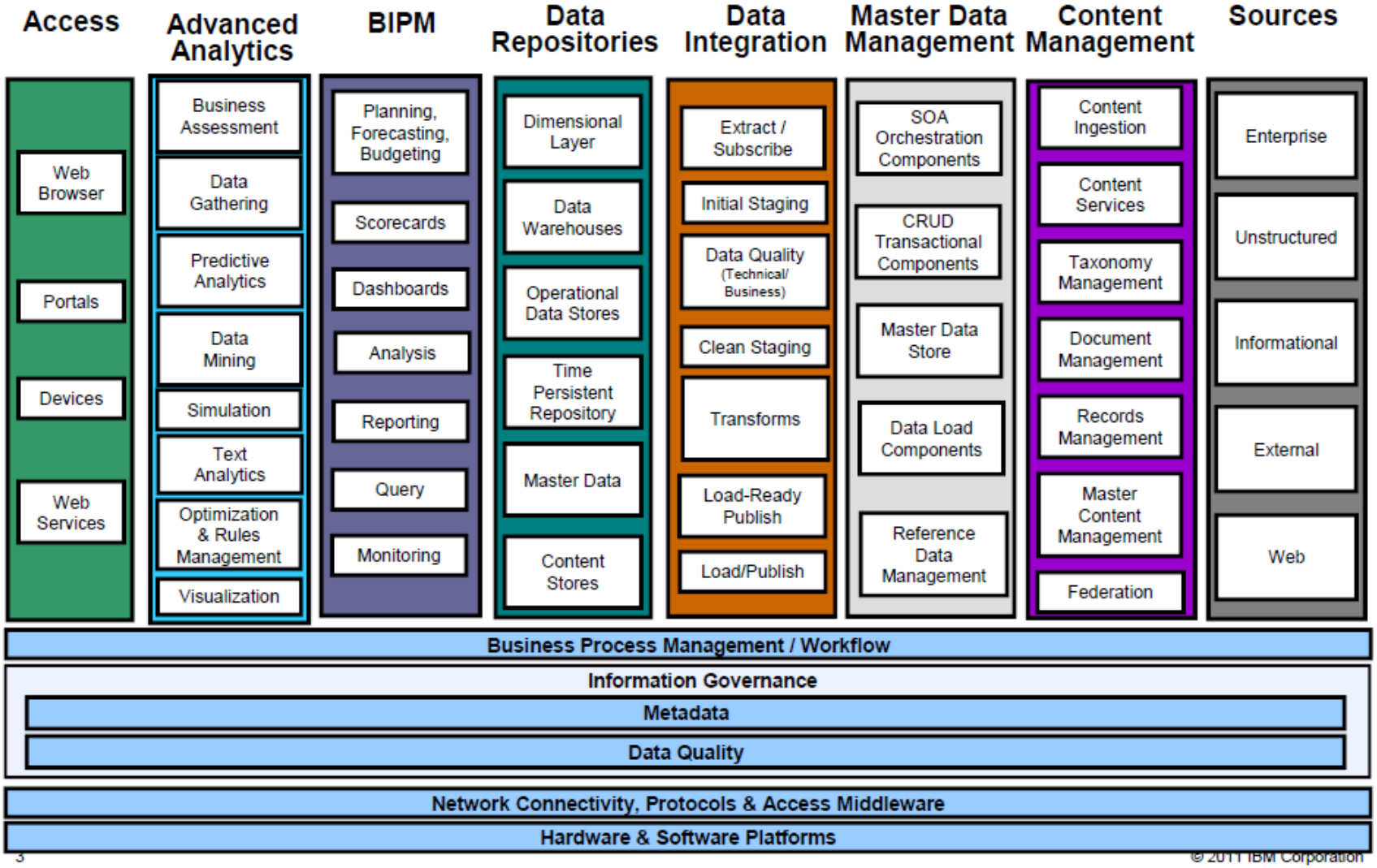
THOR Processing Cluster (Data Refinery)

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



The IBM Business Analytics and Optimization Reference Architecture Overview

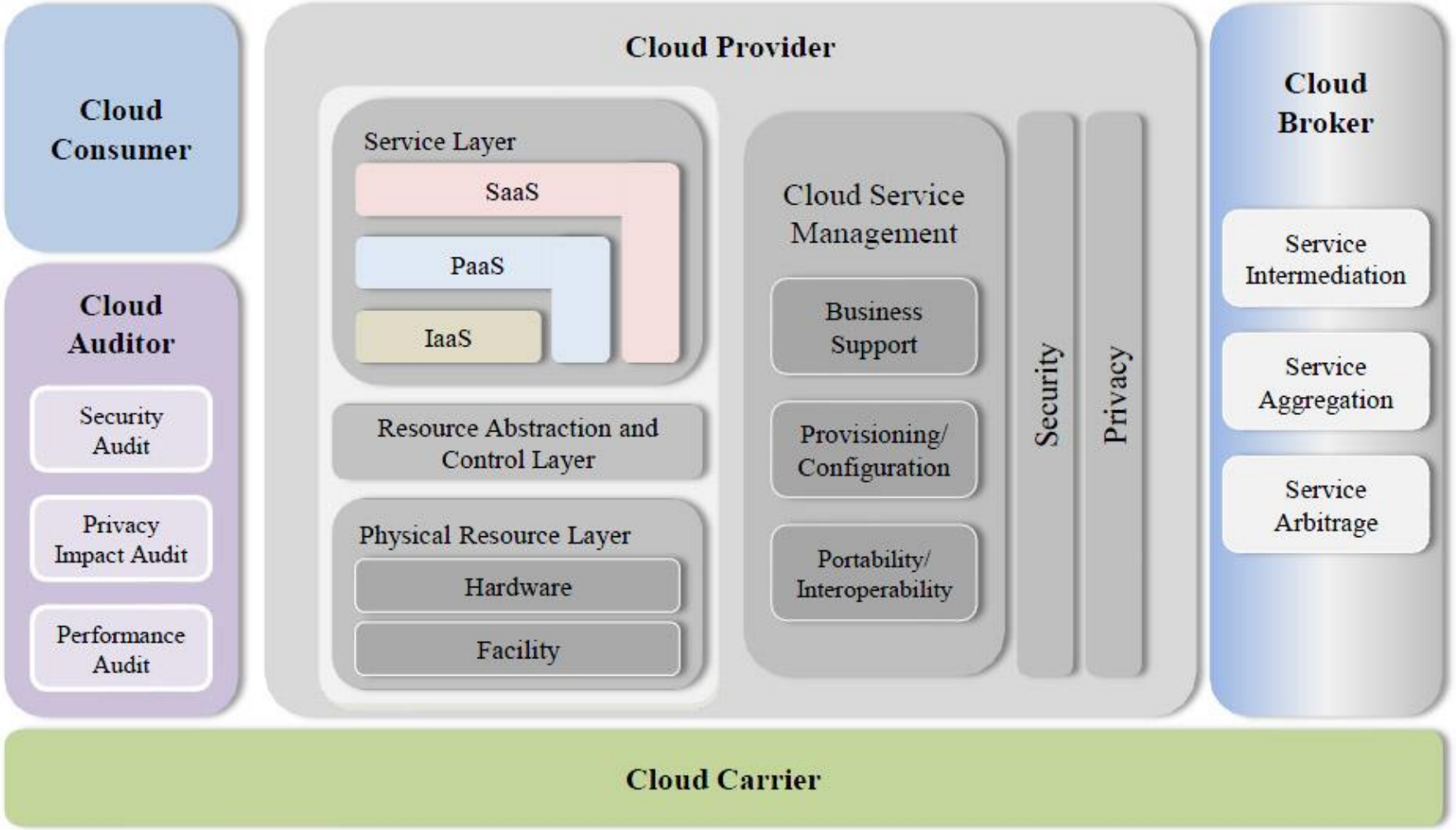


IBM GBS Business Analytics and Optimisation (2011).
https://www.ibm.com/developerworks/mydeveloperworks/files/basic/anonymous/api/library/48d92427-47d3-4e75-b54c-b6acfbd608c0/document/aa78f77c-0d57-4f41-a923-50e5c6374b6d/media&ei=yrknUbjMNM_liwKQhoCQBQ&usg=AFQjCNF_Xu6aifcAhlF4266xXNhKfKaTLw&sig2=j8JiFV_md5DnzfQl0spVrg&bvm=bv.42768644,d.cGE

BCP in Cloud/Intercloud Architecture Definition

- NIST Cloud Computing Reference Architecture (CCRA)
  - Service oriented and IT/Cloud Service Management focused
  - NIST SP 500-292, *Cloud Computing Reference Architecture, v1.0*. [Online] http://www.nist.gov/customcf/get_pdf.cfm?pub_id=909505
- Intercloud Architecture Framework (ICAF) by University of Amsterdam
  - Leverages modern Internet (IETF, ITU-T, TMF) and SOA best practices
  - *Intercloud Architecture for Interoperability and Integration*. By Demchenko, Y., C.Ngo, M.Makkes, R.Strijkers, C. de Laat. In Proc. The 4th IEEE Conf. on Cloud Computing Technologies and Science (CloudCom2012), 3 - 6 December 2012, Taipei, Taiwan.
  - *Cloud Reference Framework*, IEFT Draft, 3 July 2013. http://tools.ietf.org/html/draft-khasnabish-cloud-reference-framework-05

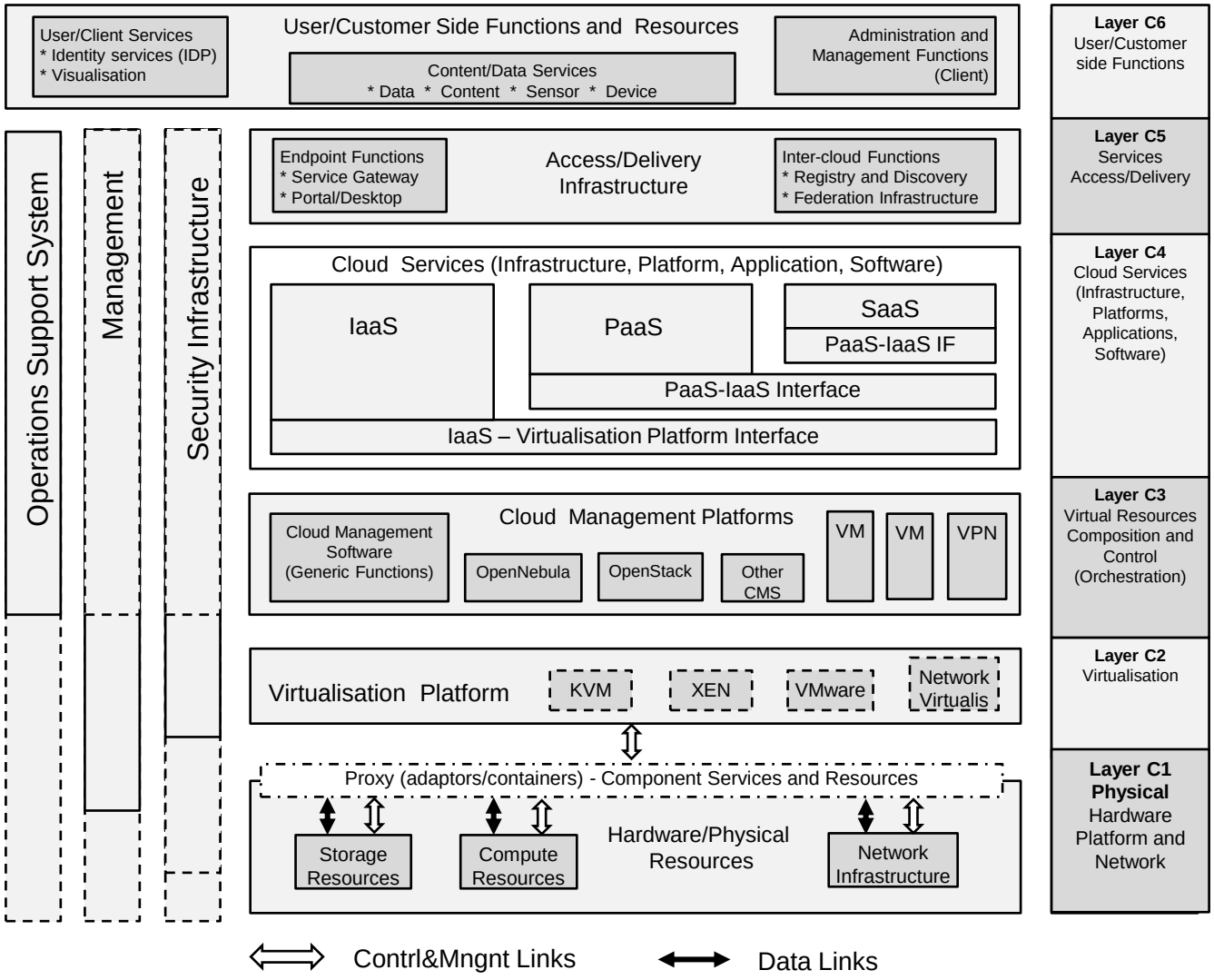
NIST Cloud Computing Reference Architecture (CCRA) 2.0 – Consolidated View



InterCloud Architecture Framework (ICAF) Components

- Multi-layer Cloud Services Model (CSM)
  - Combines IaaS, PaaS, SaaS into multi-layer model with inter-layer interfaces
  - Including interfaces definition between cloud service layers and virtualisation platform
- InterCloud Control and Management Plane (ICCMP)
  - Allows signaling, monitoring, dynamic configuration and synchronisation of the distributed heterogeneous clouds
  - Including management interface from applications to network infrastructure and virtualisation platform
- InterCloud Federation Framework (ICFF)
  - Defines set of protocols and mechanisms to ensure heterogeneous clouds integration at service and business level
  - Addresses Identity Federation, federated network access, etc.
- InterCloud Operations Framework (ICOF)
  - RORA model: Resource, Ownership, Role, Action
    - RORA model provides basis for business processes definition, SLA and access control
  - Broker and federation operation
- InterCloud Security Framework (ICSF)

Multilayer Cloud Services Model (CSM)



CSM layers

(C6) User/Customer side Functions
(C5) Intrecloud Services Access and Delivery
(C4) Cloud Services (Infrastructure, Platform, Applications)
(C3) Virtual Resources Composition and Orchestration
(C2) Virtualisation Layer
(C1) Hardware platform and dedicated network infrastructure

Control/ Mngnt Links

Data Links

E-Science Features

- ***Automation*** of all e-Science processes including data collection, storing, classification, indexing and other components of the general data curation and provenance
- ***Transformation*** of all processes, events and products ***into digital form*** by means of multi-dimensional multi-faceted measurements, monitoring and control; digitising existing artifacts and other content
- Possibility to ***re-use*** the initial and published research ***data*** with possible data re-purposing for secondary research
- ***Global data availability*** and access over the network for cooperative group of researchers, including wide public access to scientific data
- Existence of necessary infrastructure components and management tools that allows fast i***nfrastructures and services composition, adaptation and provisioning on demand*** for specific research projects and tasks
- ***Advanced security and access control*** technologies that ensure secure operation of the complex research infrastructures and scientific instruments and allow creating ***trusted secure environment*** for cooperating groups and individual researchers.

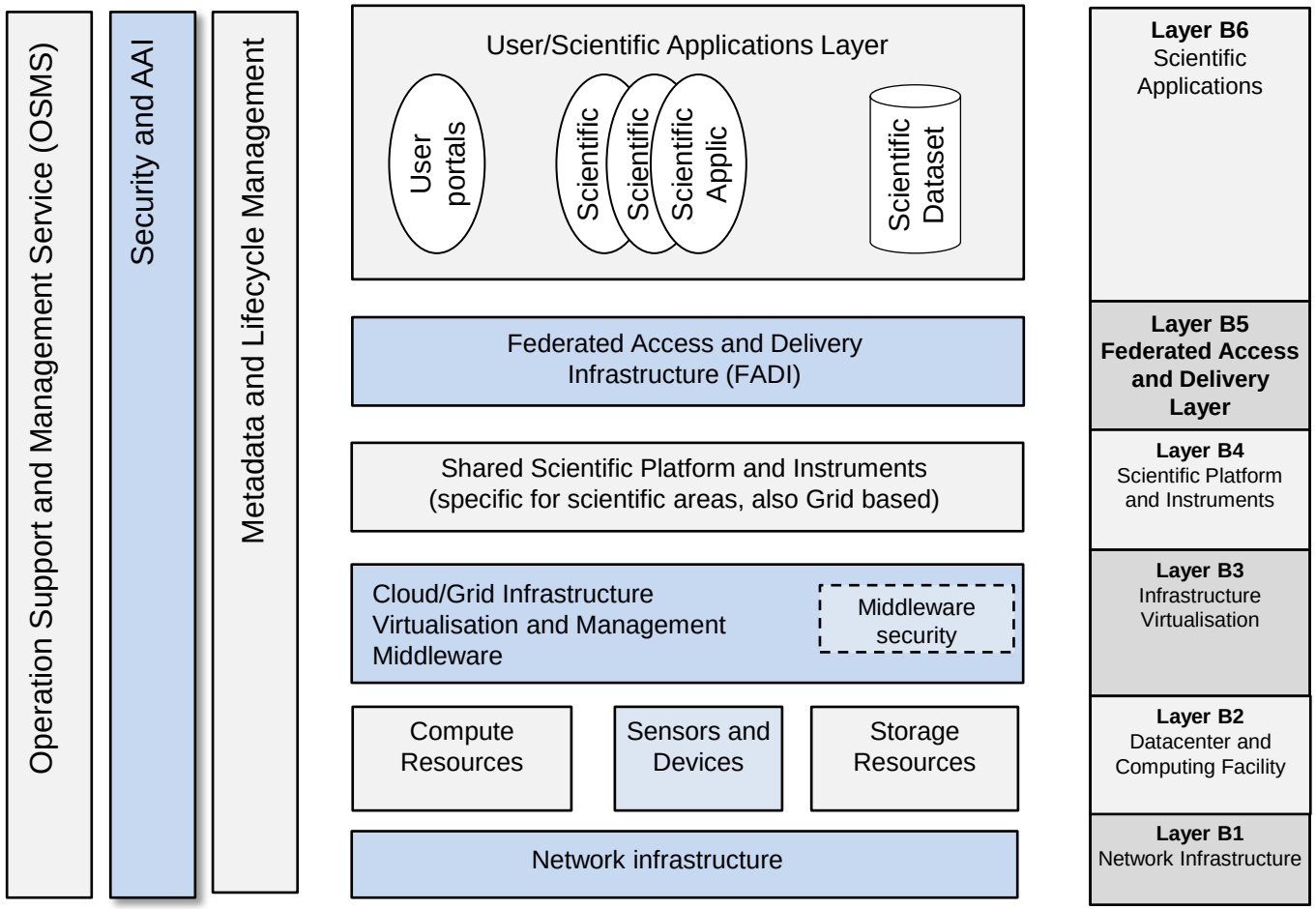
General requirements to SDI for emerging Big Data Science

- Support for *long running experiments and large data volumes* generated at high speed
- *Multi-tier inter-linked data distribution and replication*
- *On-demand infrastructure provisioning* to support data sets and scientific workflows, mobility of data-centric scientific applications
- Support of *virtual scientists communities*, addressing dynamic user groups creation and management, federated identity management
- Support for the *whole data lifecycle* including metadata and data source linkage
- *Trusted environment* for data storage and processing
  - Research need to trust SDI to put all their data on it
- Support for data integrity, confidentiality, accountability
- *Policy binding to data* to protect privacy, confidentiality and IPR

Defining Architecture framework for SDI and Security

- Scientific Data Lifecycle Management (SDLM) model
- e-SDI multi-layer architecture model
- RORA model to define relationship between resources and actors
  - RORA (Resource-Ownership-Role-Actor) model defines relationship between resources, owners, managers, users
  - Initially defined for telecom domain
  - New actors in SDI (and Big Data Infrastructure)
    - Subject of data (e.g. patient, or scientific object/paper)
    - Data Manager (doctor, seller)
- Security and Access Control and Accounting Infrastructure (ACAI)
  - Trust management infrastructure
  - Authentication, Authorisation, Accounting
    - Supported by logging service
  - Extended to support data access control and operations on data

SDI Architecture Model



SDI Architecture Layers

- **Layer D1**: Network infrastructure layer represented by the general purpose Internet infrastructure and dedicated network infrastructure
- **Layer D2**: Datacenters and computing resources/facilities, including sensor network
- **Layer D3**: Infrastructure virtualisation layer that is represented by the Cloud/Grid infrastructure services and middleware supporting specialised scientific platforms deployment and operation
- **Layer D4**: (Shared) Scientific platforms and instruments specific for different research areas
- **Layer D5**: Federated Access and Delivery Infrastructure (FADI) Layer: Federation infrastructure components, including policy and collaborative user groups support functionality
- **Layer D6**: Scientific applications and user portals/clients

SDI move to Clouds

- Cloud technologies allow for infrastructure virtualisation and its profiling for specific data structures or to support specific scientific workflows
  - Clouds provide just right technology for infrastructure virtualisation to support data sets
  - *Complex distributed data require infrastructure*
    - *Demand for inter-cloud infrastructure*
- Cloud can provide infrastructure on-demand to support project related scientific workflows
  - Similar to Grid but with benefits of the full infrastructure provisioning on-demand
- Software Defined Infrastructure Services
  - As wider than currently emerging SDN (Software Defined Networks)
- Distributed Hadoop clusters for HPC and MPP

Federated Access and Delivery Infrastructure (FADI)