System and Network Engineering Group, UvA

UNIVERSITEIT VAN AMSTERDAM

SNE

System and Network Engineering

Architecture Framework and Components for the Big Data Ecosystem

Draft Version 0.2

Yuri Demchenko, Canh Ngo, Peter Membrey

12 September 2013

http://www.uazone.org/demch/worksinprogress/sne-2013-02-techreport-bdaf-draft02.pdf

Abstract

Big Data are becoming a new technology focus both in science and in industry and motivate technology shift to data centric architecture and operational models. There is a vital need to define the basic information/semantic models, architecture components and operational models that together comprise a so-called Big Data Ecosystem. This paper discusses a nature of Big Data that may originate from different scientific, industry and social activity domains and proposes improved Big Data definition that includes the following parts: Big Data properties (also called Big Data 5V: Volume, Velocity, Variety, Value and Veracity), data models and structures, data analytics, infrastructure and security. The paper discusses paradigm change from traditional host or service based to data centric architecture and operational models in Big Data. The Big Data Architecture Framework (BDAF) is proposed to address all aspects of the Big Data Ecosystem and includes the following components: Big Data Infrastructure, Big Data Analytics, Data structures and models, Big Data Lifecycle Management, Big Data Security. The paper analyses requirements to and provides suggestions how the mentioned above components can address the main Big Data challenges. The presented work intends to provide a consolidated view of the Big Data phenomena and related challenges to modern technologies, and initiate wide discussion.

Table of Contents

1 Introduction

Big Data, also referred to as Data Intensive Technologies, are becoming a new technology trend in science, industry and business [1, 2, 3]. Big Data are becoming related to almost all aspects of human activity from just recording events to research, design, production and digital services or products delivery to the final consumer. Current technologies such as Cloud Computing and ubiquitous network connectivity provide a platform for automation of all processes in data collection, storing, processing and visualization.

The goal of our research at current stage is to understand the nature of Big Data, their main features, trends and new possibilities in Big Data technologies development, identify the security issues and problems related to the specific Big Data properties, and based on this to review architecture models and propose a consistent approach to defining the Big Data architecture/solutions to resolve existing challenges and known issues/problems.

In this paper we continue with the Big Data definition and enhance the definition given in [3] that includes the 5V Big Data properties: Volume, Variety, Velocity, Value, Veracity, and suggest other dimensions for Big Data analysis and taxonomy, in particular comparing and contrasting Big Data technologies in e-Science, industry, business, social media, healthcare. With a long tradition of working with constantly increasing volume of data, modern e-Science can offer industry the scientific analysis methods, while industry can bring advanced and fast developing Big Data technologies and tools to science and wider public.

In Big Data, data are rather a "fuel" that "powers" the whole complex of technical facilities and infrastructure components built around a specific data origin and their target use. We will call it a Big Data Ecosystem (BDE). By defining BDE we contrast its data centric character to traditional definition of the architecture that is more applicable for facility or service centric technologies. We discuss the major (architecture) components that together constitute the Big Data Ecosystem: 5V Big Data properties, Data Models and Structures, Big Data Infrastructure, Big Data lifecycle management (or data transformation flow), Big Data Security Infrastructure.

There are not many academic papers related to Big Data; in most cases they are focused on some particular technology or solution that reflect only a small part of the whole problem area. The same relates to the Big Data definition that would provide a conceptual basis for the further technology development. There is no well-established terminology in this area. Currently this problem is targeted by the recently established NIST Big Data Working Group (NBD-WG) [4] that meets at weekly basis in subgroups focused on Big Data definition, Big Data Reference Architecture, Big Data Requirements, Big Data Security. The authors are actively contributing to the NBD-WG and have presented the approach and ideas proposed/discussed in this paper at one of NBD-WG virtual meetings [5]. We will refer to the NBD-WG discussions and documents in many places along this paper to support our ideas or illustrate alternative approach.

The paper is organised as follows. Section II investigates different Big Data origin domains and target use and based on this proposes a new extended/improved Big Data definition as the main component of the Big Data Ecosystem. Section III analyses the paradigm change in Big Data and Data Intensive technologies. Section IV proposes the Big Data Architecture Framework that combines all the major components of the Big Data Ecosystem. The section also briefly discusses Big Data Management issues and required Big Data structures. Section V provides suggestions about building Big Data Infrastructure and specifically Big Data Analytics components. Section VI discusses Big Data Security Infrastructure issues and its major challenges. Section VII provides short overview /refers to other works related to defining Big Data architecture and its components. The paper concludes with the summary and suggestions for further research.

2 Big Data Nature and Application domains

We observe that Big Data "revolution" is happening in different human activity domains empowered by significant growth of the computer power, ubiquitous availability of computing and storage resources,

increase of digital content production, mobility. This creates a variety of the Big Data origin and usage domains.

Table 1 lists the main Big Data origin domains and targeted use or application, which are not exhausting and are presented to illustrate a need for detailed analysis of these aspects. Table 2 illustrates possible relations between these two dimensions and indicates their relevance. We can assume high relevance of Big Data to business; this actually explains the current strong interest to Big Data from business which is actually becoming the main driving force in this technology domain.

Table 1. Big Data origin and target use domains

Big Data Origin	Big Data Target Use
1. Science 2. Telecom 3. Industry 4. Business 5. Living Environment, Cities 6. Social media and networks 7. Healthcare	(a) Scientific discovery (b) New technologies (c) Manufacturing, process control, transport (d) Personal services, campaigns (e) Living environment support (f) Healthcare support

Table 2. Interrelation between Big Data origin and target use.

	Scient discov	New Techn	Manufact Control	Pers service	Living support	Health care
Science	++++	++++	+	-	++	+++
Telecom	+	++++	++	+	++++	+
Indust	++	++++	+++++	-	-	++
Biz	+	+++	++	-	+	++
Cities Liv Env	++	++	++	++	+++++	+
Social n/w	+	++	-	++++	++	-
Health care	+++	++	-	-	++	++++

Science has been traditionally dealing with challenges to handle large volume of data in complex scientific research experiments, involving also wide cooperation among distributed groups of individual scientists and research organizations. Scientific research typically includes collection of data in passive observation or active experiments which aim to verify one or another scientific hypothesis. Scientific research and discovery methods are typically based on the initial hypothesis and a model which can be refined based on the collected data. The refined model may lead to a new more advanced and precise experiment and/or the previous data re-evaluation. The future Scientific Data and Big Data Infrastructure (SDI/BDI) needs to support all data handling operations and processes providing also access to data and to facilities to collaborating researchers. Besides traditional access control and data security issues, security services need to ensure secure and trusted environment for researcher to conduct their research.

In business, private companies will not typically share data or expertise. When dealing with data, companies will intend always to keep control over their information assets. They may use shared third party facilities, like clouds or specialists instruments, but special measures need to be taken to ensure workspace safety and data protection, including input/output data sanitization.

Big Data in industry are related to controlling complex technological processes and objects or facilities. Modern computer-aided manufacturing produces huge amount of data which are in general need to be

stored or retained to allow effective quality control or diagnostics in case of failure or crash. Similarly to e-Science, in many industrial applications/scenarios there is a need for collaboration or interaction of many workers and technologists.

Big Data rise is tightly connected to social data revolution that both provided initial motivation for developing large scale services, global infrastructure and high performance analytical tools, and produces huge amount of data on their own. Social network are widely used for collecting personal information and providing better profiled personal services staring from personal search advice to targeted advertisements and precisely targeted campaigns.

We accept the proposed analysis is not exhaustive and can be extended and detailed but we use it to illustrate a need for a more detailed research in this area.

3 Big Data Definition

3.1 5V of Big Data

Despite the "Big Data" became a new buzz-word, there is no consistent definition of Big Data, nor detailed analysis of this new emerging technology. Most discussions until now have been going now in blogosphere where active contributors have generally converged on the most important features and incentives of the Big Data [6, 7, 8].

For the completeness of the discussion, we quote here few definitions by leading experts and consulting companies. We start with the IDC definition of Big Data (rather strict and conservative): "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" [9].

It can be complemented with more simple definition by Jason Bloomberg [8]: "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." This is also in accordance with the definition given by Jim Gray in his seminal book [10].

We concur with the Gartner definition of Big Data that is termed as 3 parts definition: "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." [11, 12]

We refer to our recent paper [3] where we summarized the existing at that time discussions and proposed the Big Data definition as having the following 5V properties: Volume, Velocity, Variety that constitute native/original Big Data properties, and Value and Veracity as acquired as a result of data initial classification and processing in the context of a specific process or model.

Further analysis of the Big Data use cases, in particular those discussed by NBD-WG [4] reveals other aspects and Big Data features.

During Big Data lifecycle, each stage of the data transformation or processing changes the data content, state and consequently may change the data model. In many cases there is a need to link original data and processed data, keeping referral integrity (see more discussion about this in the following sections).

This motivates other Big Data features: Dynamicity or Variability and Linkage or referral integrity. Dynamicity/Variability reflects the fact that data are in constant change and have a definite state besides typically/commonly defined as data in move, in rest, or being processed.

Dynamicity and data linkage are the two other factors that reflect changing or evolving character of data and need to keep their linkage during the whole their lifecycle. This will require scalable provenance models and tools incorporating also data integrity and confidentiality.

3.2 From 5V to 5 Parts Big Data Definition

To make a such new technology definition as Big Data, we need to find a way to reflect its all important features and provide a guidance/basis for further technology development (e.g., following one of the best example with the Cloud Computing definition that has been given in 2008 and actually shaped the current cloud industry).

We propose a Big Data definition as having five parts that group the main Big Data features and related components:

(1) Big Data Properties: 5V

- Volume, Variety, Velocity, Value, Veracity
- Additionally: Data Dynamicity (Variability) and Linkage.

(2) New Data Models

- Data linking, provenance and referral integrity
- Data Lifecycle and Variability

(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

To reflect all components of the Big Data features, we can summarise them in a form of the improved Gartner definition:
"Big Data (Data Intensive) Technologies are targeting to process 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."

3.3 Big Data Ecosystem

Big Data is not just a database or Hadoop problem, although they constitute the core technologies and components for large scale data processing and data analytics [13, 14, 15]. It is the whole complex of components to store, process, visualize and deliver results to target applications. Actually Big Data is "a fuel" of all these processes, source, target, and outcome.

All this complex interrelated can be defined as the Big Data Ecosystem (BDE) that deals with the evolving data, models and required infrastructure during the whole Big Data lifecycle. In the following we will provide more details about our vision of the BDE.

4 Paradigm change in Big Data and Data Intensive Science and Technologies

The recent advancements in the general ICT, Cloud Computing and Big Data technologies facilitate the paradigm change in modern e-Science and industry that is characterized by the following features[3, 16]:

- 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.
- Automation of all data production, consumption and management processes including data collection, storing, classification, indexing and other components of the general data curation and provenance.
- Possibility to re-use and repurpose the initial data sets for new and secondary data analysis based on the model improvement
- 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 allow fast infrastructures 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 and production infrastructures and allow creating trusted secure environment for cooperating groups of researchers and technology specialists.

The following are additional factors that will create new challenges and motivate security paradigms change in Big Data ecosystem/technology:

- Virtualization: can improve security of data processing environment but cannot solve data security "in rest".
- Mobility of the different components of the typical data infrastructure: sensors or data source, data consumer, and data themselves (original data and staged/evolutional data). This in its own cause the following problems
    - On-demand infrastructure services provisioning
    - Inter-domain context communication
- Big Data aggregation that may involve data from different administrative/logical domains and evolutionally changing data structures (also semantically different).
- Policy granularity: Big Data may have complex structure and require different and high-granular policies for their access control and handling.

The future Big Data Infrastructure (BDI) should support the whole data lifecycle and explore the benefit of the data storage/preservation, aggregation and provenance in a large scale and during long/unlimited period of time. Important is that this infrastructure must ensure data security (integrity, confidentiality, availability, and accountability), and data ownership protection. With current needs to process big data that require powerful computation, there should be a possibility to enforce data/dataset policy that they can be processed on trusted systems and/or complying other requirements. Customers must trust the BDI to process their data on BDI facilities and be ensured that their stored research data are protected from non-authorised access. Privacy issues are also arising from distributed remote character of BDI that can span multiple countries with different local policies. This should be provided by the Access Control and Accounting Infrastructure (ACAI) which is an important component of SDI [16, 17].

4.1 From Big Data to All-Data Metaphor

One of difficulties in defining Big Data and setting a common language/vocabulary for Big Data is the different view of the potential stakeholders. , For example, big business and big science re arguing how big are big data: is Petabyte a big data? Is Exabyte a big data? While smaller businesses and "long-tale" science [8] (i.e., that doesn't generate huge amount of data) may conclude that they will never become Big Data players.

In this respect, it is important to look at the current Big Data related trends in general and investigate/analyse what are the components of the Big Data ecosystem and how they impact the present ICT infrastructure in first place, and how these changes will affect other IT domains and applications.

Following the trend in some Big Data analytics domain to collect and analyse all available data (all data that can be collected), we can extend it to the following metaphor: "From Big Data to All-Data". It is depicted in Figure 1 that illustrates that there that the traditional dilemma "move data to computing or computing to data" is not valid in this case, and we really need to look at the future Big Data/All-Data processing model and infrastructure differently.

All-Data infrastructure will need tom adopt generically/naturally distributed storage and computing, a complex of functionalities which we depicted as Data Bus will provide all complex functionality to exchange data, distributed and synchronise processes, and many other functions that should cope with the continuous data production, processing and consumption.

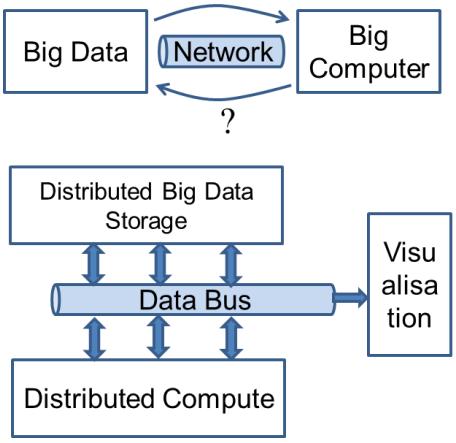


Figure 1. From Big Data to All-Data Metaphor.

4.2 Moving to Data-Centric Models and Technologies

Traditional/current IT and communication technologies are OS/system based and host/service centric what means that all communication or processing are bound to host/computer that runs application software. This is especially related to security services. The administrative and security domains are the key concepts, around which the services and protocols are built. A domain provides a context for establishing security context and trust relation. This creates a number of problems when data (payload or session context) are moved from one system to another or between domains, or operated in a distributed manner.

Big Data will require different data centric operational models and protocols, what is especially important in situation when the object or event related data will go through a number of transformations and become even more distributed, between traditional security domains. The same relates to the current federated access control model that is based on the cross administrative and security domains identities and policy management.

When moving to generically distributed data centric models additional research are needed to address the following issues:

- Maintaining semantic and referral integrity, to support data provenance in particular,
- Data location, search, access
- Data integrity and identifiability, referral integrity
- Data security and data centric access control
- Data ownership, personally identified data, privacy, opacity
- Trusted virtualisation platform, data centric trust bootstrapping

5 Proposed Big Data Architecture Framdework

Discussion above motivated a need for a new approach to the definition of the Big Data Ecosystem that would address the major challenges related to the Big Data properties and component technologies.

In this section we propose the Big Data Architecture Framework (BDAF) that would support the extended Big Data definition given in section II.C and support the main components and processes in the Big Data Ecosystem (BDE). We base our BDAF definition on our and communities experience and industry best practices in defining architectures for new technologies, in particular, NIST Cloud Computing Reference Architecture (CCRA) [18], Intercloud Architecture Framework (ICAF) by authors [19], recent discussions by the NIST Big Data Working Group [4] and in particular initial Big Data Ecosystem Architecture definition by Microsoft [20], Big Data technology analysis by G.Mazzaferro [21]. We also refer to other related architecture definitions: Information as a Service by Open Data Center Alliance [22], TMF Big Data Analytics Architecture [23], IBM Business Analytics and Optimisation Reference Architecture [24], LexisNexis HPCC Systems [25].

We propose our definition of the Big Data Architecture Framework that summarises all known to us research and discussions in this area. The proposed BDAF comprises of the following 5 components that address different Big Data Ecosystem and Big Data definition aspects:

(1) Data Models, Structures, Types

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

(2) Big Data Management

- Big Data Lifecycle (Management)
- 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

To simply validate the consistency of the proposed definition we can look how the proposed components are related to each other. This is illustrated in Table 3 that shows what architecture component is used or required by another component.

Table 3. Interrelation between BDAF components

Col: Used By Row: Reqs This	Data Models	Data Mngnt& Lifecycle	BD Infra & Operat	BD Analytics	Big Data Security
Data Models		+	++	+	++
Data Mngnt& Lifecycle	++		++	++	++
BD Infrastr & Operation	+++	+++		++	+++
BD Analytics	++	+	++		++
Big Data Security	+++	+++	+++	+	

The proposed BDAF definition is rather technical and infrastructure focused and actually reflects the technology oriented stakeholders. The further BDAF improvement should also consider other stakeholder groups such as data archives providers and libraries who will play renewed role in the BDE [26].

5.1 Data Models and Structures

Emergence of computer aided research methods is transforming the way research is done and scientific data are used. The following types of scientific data are defined [16]:

- 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.

Once the data is published, it is essential to allow other scientists to be able to validate and reproduce the data that they are interested in, and possibly contribute with new results. Capturing information about the processes involved in transformation from raw data up until the generation of published data becomes an important aspect of scientific data management. Scientific data provenance becomes an issue that also needs to be taken into consideration by Big Data providers [27].

Different stages of the Big Data transformation will require and use different data structures, models and formats, including also a possibility to process both structured and unstructured data [28].

The following data types can be defined [29]

(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

Figure 2 illustrates the Big Data structures, models and their linkage at different processing stages. We can admit that data structures and correspondingly models may be different at different data processing stages, however it is required/important to keep linkage between data.

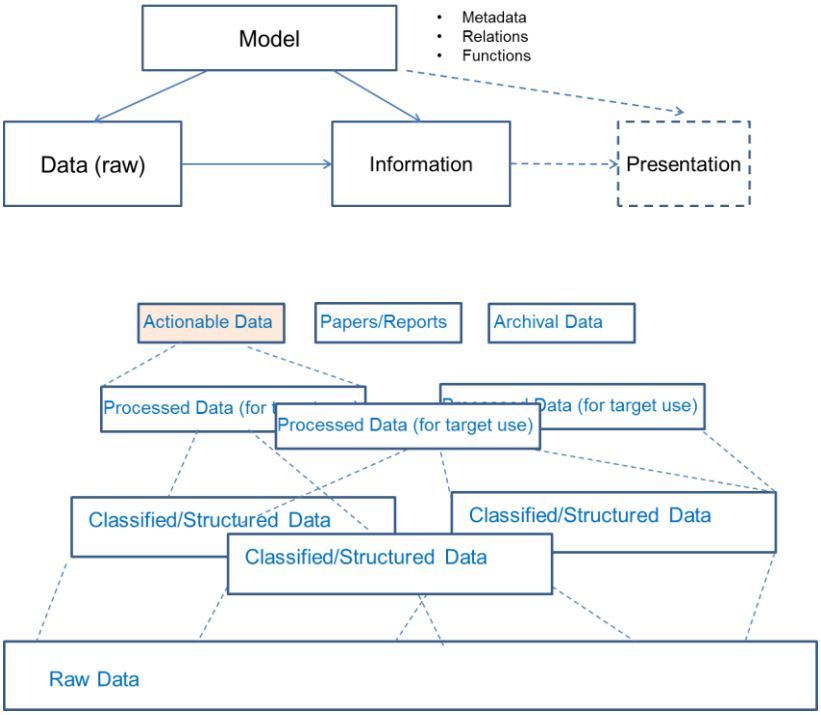


Figure 2. Big Data structures, models and their linkage at different processing stages.

Another aspect to take into consideration is to guarantee reusability of published data within the scientific community. Understanding semantics of the published data becomes an important issue to allow for reusability, and this had been traditionally been done manually. However, as we anticipate unprecedented scale of published data that will be generated in Big Data Science, attaching clear data semantic becomes a necessary condition for efficient reuse of published data. Learning from best practices in semantic web community on how to provide a reusable published data, will be one of consideration that will be addressed by BDI/SDI.

Big data are typically distributed both on the collection side and on the processing/access side: data need to be collected (sometimes in a time sensitive way or with other environmental attributes), distributed and/or replicated. Linking distributed data is one of the problems to be addressed by Big Data structures and underlying infrastructure.

We can mention as the main motivation The European Commission's initiative to support Open Access to scientific data from publicly funded projects suggests introduction of the following mechanisms to allow linking publications and data [30, 31]:

- PID - persistent data ID
- ORCID – Open Researcher and Contributor Identifier [32].

5.2 Data Management and Big Data Lifecycle

With the digital technologies proliferation into all aspects of business activities, the industry and business are entering a new playground where they need to use scientific methods to benefit from the new opportunities to collect and mine data for desirable information, such as market prediction, customer behavior predictions, social groups activity predictions, etc. Refer to numerous blog articles [3, 33] suggesting that the Big Data technologies need to adopt scientific discovery methods that include iterative model improvement and collection of improved data, re-use of collected data with improved model.

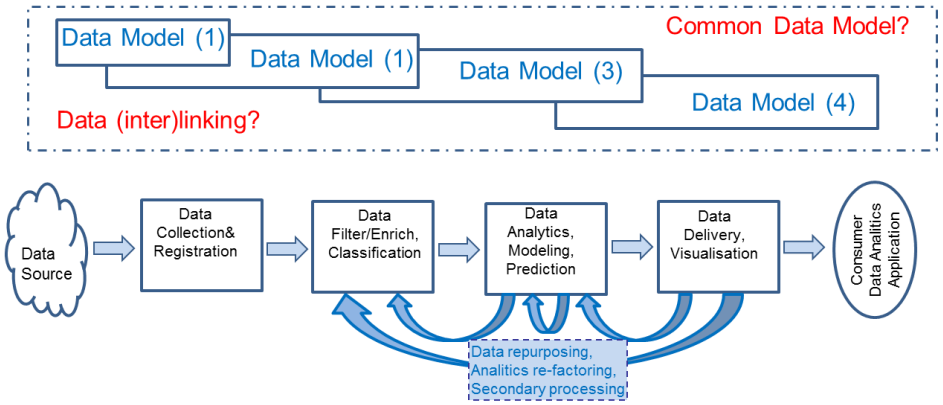


Figure 3. Big Data Lifecycle in Big Data Ecosystem.

We refer to the Scientific Data Lifecycle Management model described in our earlier paper [3, 16] and was a subject for detailed research in another work [34] that reflects complex and iterative process of the scientific research that includes a number of consequent stages: research project or experiment planning; data collection; data processing; publishing research results; discussion, feedback; archiving (or discarding)

The required new approach to data management and processing in Big Data industry is reflected in the Big Data Lifecycle Management (BDLM) model (see Figure 3) we as a result of analysis of the existing practices in different scientific communities.

New BDLM requires data storage and preservation at all stages what should allow data re-use/re-purposing and secondary research on the processed data and published results. However, this is possible only if the full data identification, cross-reference and linkage are implemented in BDI. Data integrity, access control and accountability must be supported during the whole data during lifecycle. Data curation is an important component of the discussed BDLM and must also be done in a secure and trustworthy way.

6 Big Data Infrastructure (BDI)

Figure 4 provides a general view on the Big Data infrastructure that includes the general infrastructure for general data management, typically cloud based, and Big Data Analytics part that will require high-performance computing clusters.

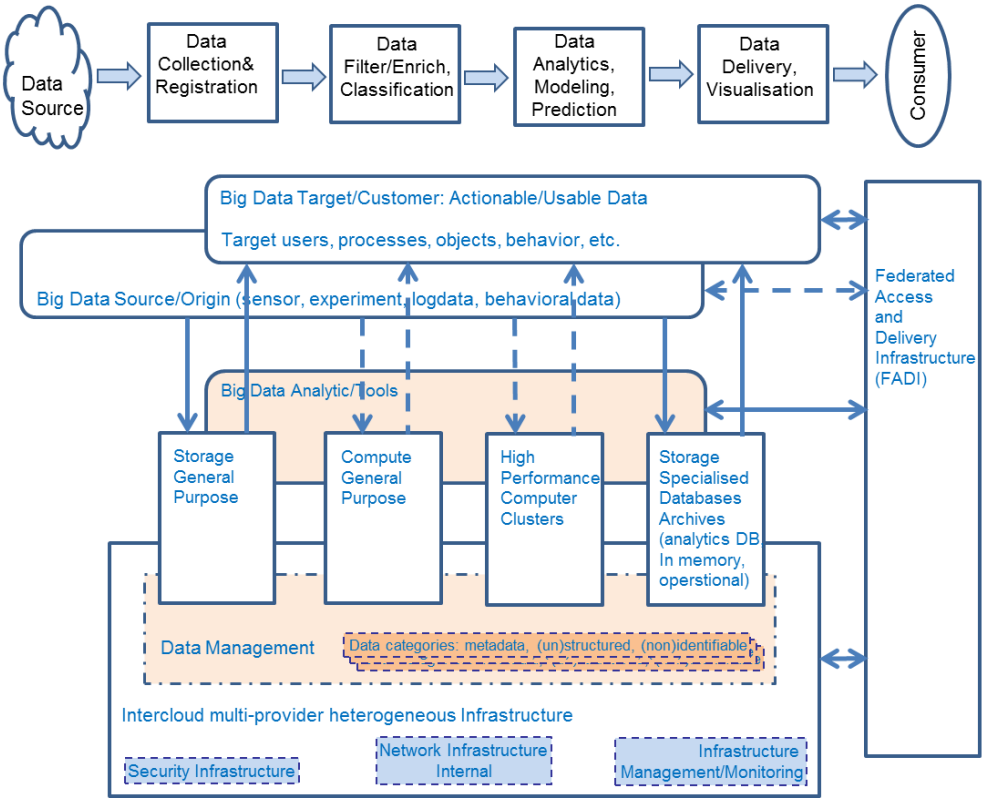


Figure 4. General Big Data Infrastructure functional components

General BDI services and components include

- Big Data Management tools
- Registries, indexing/search, semantics, namespaces
- Security infrastructure (access control, policy enforcement, confidentiality, trust, availability, privacy)
- Collaborative environment (groups management)

6.1 Big Data Analytics Infrastructure

Besides the general cloud base infrastructure services (storage, compute, infrastructure/VM management) the following specific applications and services will be required to support Big Data and other data centric applications [35]:

- Cluster services
- Hadoop related services and tools
- Specialist data analytics tools (logs, events, data mining, etc.)
- Databases/Servers SQL, NoSQL
- MPP (Massively Parallel Processing) databases

Big Data analytics tools are currently offered by the major cloud services providers such as: Amazon Elastic MapReduce and Dynamo [36], Microsoft Azure HDInsight [37], IBM Big Data Analytics [38]. Scalable Hadoop and data analytics tools services are offered by few companies that position themselves as Big Data companies such as Cloudera, [39] and few others [40].

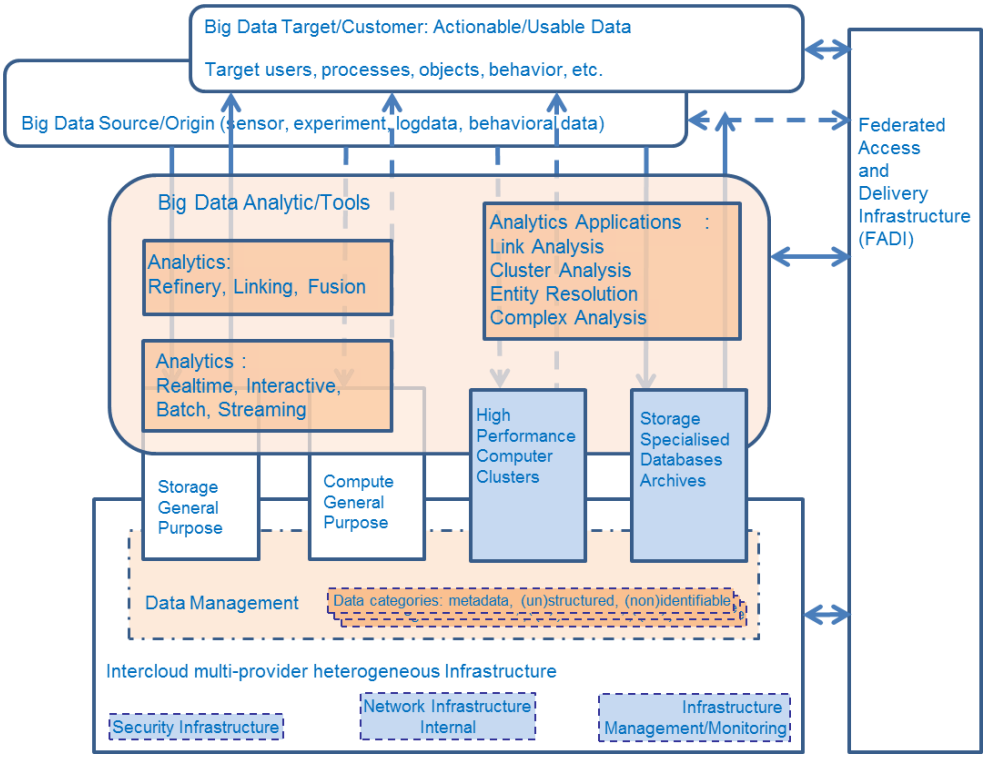


Figure 5. Big Data Analytics infrastructure components

7 Cloud Based Infrastructure Services for BDI

Figure 6 illustrates the typical e-Science or enterprise collaborative infrastructure that is created on demand and includes enterprise proprietary and cloud based computing and storage resources, instruments, control and monitoring system, visualization system, and users represented by user clients and typically residing in real or virtual campuses.

The main goal of the enterprise or scientific infrastructure is to support the enterprise or scientific workflow and operational procedures related to processes monitoring and data processing. Cloud technologies simplify the building of such infrastructure and provision it on-demand. Figure 3 illustrates how an example enterprise or scientific workflow can be mapped to cloud based services and later on deployed and operated as an instant inter-cloud infrastructure. It contains cloud infrastructure segments IaaS (VR3-VR5) and PaaS (VR6, VR7), separate virtualised resources or services (VR1, VR2), two interacting campuses A and B, and interconnecting them network infrastructure that in many cases may need to use dedicated network links for guaranteed performance.

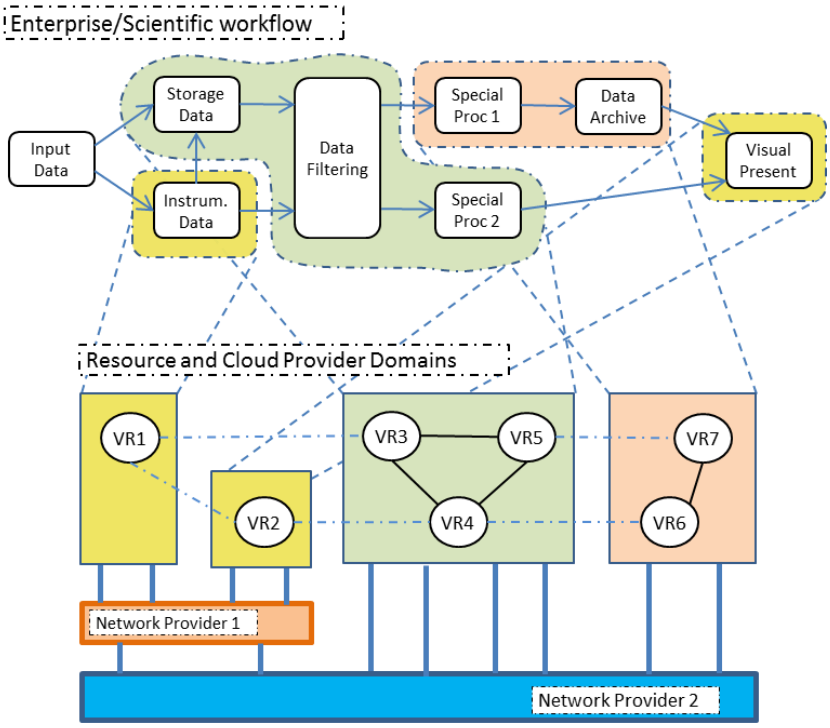


Figure 6. From scientific workflow to cloud based infrastructure.

Efficient operation of such infrastructure will require both overall infrastructure management and individual services and infrastructure segments to interact between themselves. This task is typically out of scope of the existing cloud service provider models but will be required to support perceived benefits of the future e-SDI. These topics are a subject of another research we did on the InterCloud Architecture Framework [19, 41].

8 Big Data Security Framework Components

This section discusses the Big Data Security Framework that supports a new paradigm of the data centric security. The following components are included:

- Security lifecycle
- Fine grained access control
- Encryption enforced access control
- Trusted environment
- FADI for cooperation and services integration

8.1 Federated Access and Delivery Infrastructure (FADI)

Federated Access and Delivery Infrastructure (FADI) is defined as Layer 5 in the generic SDI Architecture model for e-Science (e-SDI). It includes federation infrastructure components, including policy and collaborative user groups support functionality.

When implemented in clouds, the FADI and SDI in general may involve multiple providers and both cloud and non-cloud based infrastructure components. Our vision and intention is to use for this purpose the general Intercloud Architecture Framework (ICAF) proposed in our works [19]. ICAF provides a common basis for building adaptive and on-demand provisioned multi-provider cloud based services.

Figure 4 illustrates the general architecture and the main components of the FADI (that corresponds to the ICAF Access and Delivery Layer C5) that includes infrastructure components to support inter-cloud federations services such as Cloud Service Brokers, Trust Brokers, and Federated Identity Provider. Each service/cloud domain contains an Identity Provider IDP, Authentication, Authorisation, Accounting (AAA) service and typically communicate with other domains via service gateway.

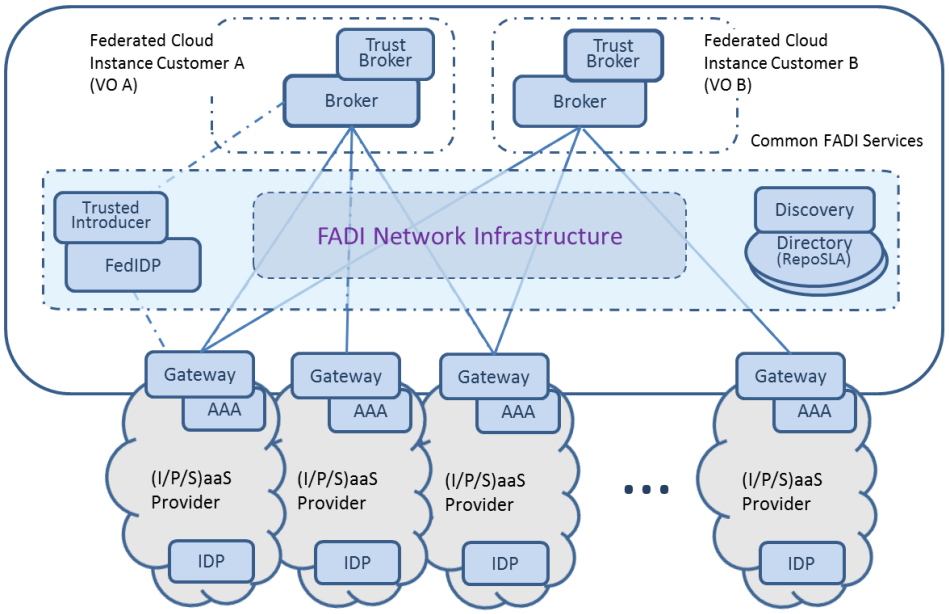


Figure 6. Federated Access and Delivery Infrastructure (FADI)

FADI incorporates related federated infrastructure management and access technologies [16, 41, 42, 43]. Using federation model for integrating multi-provider heterogeneous services and resources reflects current practice in building and managing complex infrastructures (both SDI and enterprise infrastructures) and allows for inter-organisational resource sharing.

8.2 Data Centric Access Control

SDI/BDI will incorporate standards and if needed advance access control services and mechanisms at the level of FADI and users/services level. However consistent data centric security and access control will require solving the following problems:

- Fine-granular access control policies.
- Encryption enforced attribute based access control

Depending on the data type and format, the two basic access control and policy models can be defined: resource and/or document based access control, including intra document; and cell or record based access control for data stored in databases. We identify XACML policy language as appropriate for document/intra-document access control. For databases we need to combine their native access control mechanisms and general document based access control.

8.2.1 XACML policies for fine granular access control

The policies for data centric access control model should provide the fine-grained authorization features, based not only on the request context attributes such as subjects/users, data identifiers, actions or lifetimes, but also on the structured data content. A prospective direction is to design and apply attribute based access control mechanisms with policies incorporate along with data granularity. Such policies may contain complex logic expressions of attributes. Based on input attribute values from users, their queries could return either authorized data or errors. In this respect, managing SDI/BDI big data using attribute-based policy languages like XACML is applicable. However, for large documents or complex data structures XACML policies evaluation may create a significant performance overhead.

We refer to our experience in developing Dynamically provisioned Access Control Infrastructure (DACI) for complex infrastructure services and resources [44]. It uses advanced features of the XACML based policies that allow describing access control rules for complex multi-domain resources, including domain,

session context, multi-domain identity and trust delegation [45, 46]. The proposed in [47] the Multi-data-types Interval Decision Diagrams (MIDD) policy decision request evaluation method allows for significant performance gain for massively large policy sets.

8.2.2 Access control in NoSQL databases

The popular NoSQL databases for structured data storage MongoDB [48], Cassandra [49], HBase [50], Accumulo [51] provide different levels of security and access control. Most of them have coarse-grain authorization features, both on user management and on protected data granularity like table-level or row-level security. Accumulo [51] provides the most advanced features to allow cell-level security with which accesses from keys to values are only granted when the submitted attributes satisfy predefined Boolean expressions provided as a security label of the cell key index. However, the current policy language in Accumulo is at early development stage and lacks of features for distributed, multi-domains environments.

8.2.3 Encryption enforced access control

Described above solutions are capable to address majority of the problems for data access, transfer and processing stages, however data in-rest when stored on remote facilities may remain unprotected. The solution to this problem can be found with using the encryption enhanced access control policies that in addition to the traditional access control, use also attributes based encryption [52, 53] to allow data decryption only to the targeted subject or attribute owner. We admit such approach as potentially effective and applicable to many data protection use cases in Big Data, in particular, healthcare or targeted broadcast of streaming data that make take place when using distributed sensor networks.

8.3 Trusted Infrastructure Bootstrapping Protocol

To address the issues with creating trusted remote/distributed environment for processing sensitive data, in our earlier papers [54, 55] we proposed a generic Dynamic Infrastructure Trust Bootstrapping Protocol (DITBP). This includes supporting mechanisms and infrastructure that takes advantage of the TCG Reference Architecture (TCGRA) and Trusted Platform Module (TPM) [56, 57]. The TPM is used to provide a root of trust that extends from the physical hardware itself. The TPM is used to generate a key pair in hardware where the private key is never revealed (the key pair is non-migratable). The key is only available when the machine is in a known and trusted state. The key pair is used to authenticate the machine and to decrypt the payload which is then executed to bootstrap the rest of the virtual infrastructure.

9 Related work

There are not many academic papers related to the definition of the Big Data Architecture or its components. However for our purposes we have used a number of blog posts, standards, and industry best practices that were mentioned and cited in many places in the paper. Here we just mention these works that we consider as a foundation for our work. The following publications contribute to the research on the Big Data Architecture. NIST Big Data Working Group which provide a good forum for discussion but have only plans to produce initial draft document by the end of September 2013. Other documents and works include: NIST Cloud Computing Reference Architecture (CCRA) [18], Big Data Ecosystem Architecture definition by Microsoft [20], Big Data technology analysis by G.Mazzaferro [21].

We also refer to other related architecture definitions: Information as a Service by Open Data Center Alliance [22], TMF Big Data Analytics Architecture [23], IBM Business Analytics and Optimisation Reference Architecture [24], LexisNexis HPCC Systems [25].

10 Future Research and Development

The future research and development will include further Big Data definition initially presented in this paper. At this stage we tried to summarise and re-think some widely used definitions related to Big Data,

further research will require more formal approach and taxonomy of the general Big Data use cases in different Big Data origin and target domains.

The authors will continue contributing to the NIST Big Data WG targeting both goal to propose own approach and to validated it against industry standardisation process.

Another target research direction is defining a Common Body of Knowledge (CBK) in Big Data to provide a basis for a consistent curriculum development. This work and related to the Big Data metadata, procedures and protocols definition is planned to be contributed to the Research Data Alliance (RDA) [58].

The authors believe that the proposed paper will provide a step toward the definition of the Big Data Architecture framework and Common Body of Knowledge (CBK) in Big Data and Data Intensive technologies.

Acknowledgement

The authors acknowledge a possibility of discussing the proposed ideas and received useful comments from the SNE Big Data Interest Group that is started to provide an open discussion forum for new emerging technology of Big Data and Data Intensive Technologies.

11 References

[1] Global Research Data Infrastructures: Towards a 10-year vision for global research data infrastructures. Final Roadmap, March 2012. [online] http://www.grdi2020.eu/Repository/FileScaricati/6bdc07fb-b21d-4b90-81d4-d909fdb96b87.pdf

[2] Riding the wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. October 2010. [online] Available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[3] Demchenko, Y., P.Membrey, P.Grosso, C. de Laat, Addressing Big Data Issues in Scientific Data Infrastructure. First International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2013). Part of The 2013 Int. Conf. on Collaboration Technologies and Systems (CTS 2013), May 20-24, 2013, San Diego, California, USA.

[4] NIST Big Data Working Group (NBD-WG). [online] http://bigdatawg.nist.gov/home.php

[5] Definting Big Data Architetcure Framework: Outcome of the Brainstorming Session at the University of Amsterdam, 17 July 2013. Presenttaion to NBD-WG, 24 July 2013 [online] http://bigdatawg.nist.gov/_uploadfiles/M0055_v1_7606723276.pdf

[6] Reflections on Big Data, Data Science and Related Subjects. Blog by Irving Wladawsky-Berger. [online] http://blog.irvingwb.com/blog/2013/01/reflections-on-big-data-data-science-and-related-subjects.html

[7] E.Dumbill, What is big data? An introduction to the big data landscape. [online] http://strata.oreilly.com/2012/01/what-is-big-data.html

[8] The Big Data Long Tail. Blog post by Jason Bloomberg on January 17, 2013. [online] http://www.devx.com/blog/the-big-data-long-tail.html

[9] Extracting Value from Chaos, By John Gantz and David Reinsel, IDC IVIEW, June 2011. [online] http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

[10] The Fourth Paradigm: Data-Intensive Scientific Discovery. Edited by Tony Hey, Stewart Tansley, and Kristin Tolle. Microsoft Corporation, October 2009. ISBN 978-0-9825442-0-4 [online] http://research.microsoft.com/en-us/collaboration/fourthparadigm/

[11] Big Data defintion, Gartner, Inc. [online] http://www.gartner.com/it-glossary/big-data/

[12] Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s, By Svetlana Sicular, Gartner, Inc. 27 March 2013. [online] http://www.forbes.com/sites/gartnergroup/2013/03/27/ gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/

[13] The Top of the Big Data Stack: Database Applications, By Jeffrey Layton, July 27, 2012. [online] http://www.enterprisestorageforum.com/storage-management/the-top-of-the-big-data-stack-database-applications.html

[14] Explore big data analytics and Hadoop. [online] http://www.ibm.com/developerworks/training/kp/os-kp-hadoop/

[15] 7 Myths on Big Data—Avoiding Bad Hadoop and Cloud Analytics Decisions, by Adam Bloom, April 22, 2013. [online] http://blogs.vmware.com/vfabric/2013/04/myths-about-running-hadoop-in-a-virtualized-environment.html

[16] European Union. A Study on Authentication and Authorisation Platforms For Scientific Resources in Europe. Brussels : European Commission, 2012. Final Report. Contributing author. Internal

identification SMART-Nr 2011/0056. [online] Available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/aaa-study-final-report.pdf

[17] Demchenko, Y., P.Membrey, C.Ngo, C. de Laat, D.Gordijenko., Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure, Proc. Secure Data Management (SDM'13) Workshop. Part of VLDB2013 conference, 26-30 August 213, Trento, Italy

[18] NIST SP 500-292, Cloud Computing Reference Architecture, v1.0. [Online] http://collaborate.nist.gov/twiki-cloud-computing/pub/CloudComputing/ReferenceArchitectureTaxonomy/NIST_SP_500-292_-_090611.pdf

[19] Demchenko, Y., M. Makkes, R.Strijkers, C.Ngo, C. de Laat, Intercloud Architecture Framework for Heterogeneous Multi-Provider Cloud based Infrastructure Services Provisioning, The International Journal of Next-Generation Computing (IJNGC), Volume 4, Issue 2, July 2013

[20] Big Data Ecosystem Reference Architecture (Microsoft). NBDWG Contribution, NIST http://bigdatawg.nist.gov/_uploadfiles/M0015_v1_1596737703.docx

[21] Towards A Big Data Reference Architecture 2011 (Selected Slides). Revised. NBDWG contribution. [onlie] http://bigdatawg.nist.gov/_uploadfiles/M0054_v1_8456980532.pdf

[22] 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

[23] TR202 Big Data Analytics Reference Model. TMF Document, Version 1.9, April 2013.

[24] 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

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

[26] Bierauge, M., Keeping Up With... Big Data. American Library Association. [online] http://www.ala.org/acrl/publications/keeping_up_with/big_data

[27] D.Koopa, et al, A Provenance-Based Infrastructure to Support the Life Cycle of Executable Papers, International Conference on Computational Science, ICCS 2011. [online] http://vgc.poly.edu/~juliana/pub/vistrails-executable-paper.pdf

[28] Unstructured Data Management, Hitachi Data System. [online] http://www.hds.com/solutions/it-strategies/unstructured-data-management.html

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

[30] Open Access: Opportunities and Challenges. European Commission for UNESCO. [online] http://ec.europa.eu/research/science-society/document_library/pdf_06/open-access-handbook_en.pdf

[31] OpenAIR – Open Access Infrastructure for Research in Europe. [online] http://www.openaire.eu/

[32] Open Researcher and Contributor ID. [online] http://about.orcid.org/

[33] The Forrester Wave: Big Data Predictive Analytics Solutions, Q1 2013. Mike Gualtieri, January 13, 2013. [online] http://www.forrester.com/pimages/rws/reprints/document/85601/oid/1-LTEQDI

[34] Data Lifecycle Models and Concepts. [online] http://wgiss.ceos.org/dsig/whitepapers/Data%20Lifecycle%20Models%20and%20Concepts%20v8.docx

[35] A chart of the big data ecosystem, take 2. By Matt Turk [online] http://mattturck.com/2012/10/15/a-chart-of-the-big-data-ecosystem-take-2/

[36] Amazon Big Data. [online] http://aws.amazon.com/big-data/

[37] Microsoft Azure Big Data. [online] http://www.windowsazure.com/en-us/home/scenarios/big-data/

[38] IBM Big Data Analytics. [online] http://www-01.ibm.com/software/data/infosphere/bigdata-analytics.html

[39] Cloudera Impala Big Data Platform http://www.cloudera.com/content/cloudera/en/home.html

[40] 10 hot big data startups to watch in 2013, 10 January 2013 [online] http://beautifuldata.net/2013/01/10-hot-big-data-startups-to-watch-in-2013/

[41] Makkes, Marc, Canh Ngo, Yuri Demchenko, Rudolf Strijkers, Robert Meijer, Cees de Laat, Defining Intercloud Federation Framework for Multi-provider Cloud Services Integration, The Fourth International Conference on Cloud Computing, GRIDs, and Virtualization (CLOUD COMPUTING 2013), May 27 - June 1, 2013,Valencia, Spain.

[42] EGI federated cloud task force. [online] http://www.egi.eu/infrastructure/cloud/cloudtaskforce.html

[43] eduGAIN - Federated access to network services and applications. [online] http://www.edugain.org

[44] Demchenko, Y., C.Ngo, C. de Laat, T.Wlodarczyk, C.Rong, W.Ziegler, Security Infrastructure for On-demand Provisioned Cloud Infrastructure Services, Proc. 3rd IEEE Conf. on Cloud Computing Technologies and Science (CloudCom2011), 29 November - 1 December 2011, Athens, Greece. ISBN: 978-0-7695-4622-3

[45] Ngo; C., Membrey, P.; Demchenko, Y.; De Laat, C., "Policy and Context Management in Dynamically Provisioned Access Control Service for Virtualized Cloud Infrastructures," Availability, Reliability and Security (ARES), 2012 Seventh International Conference on , vol., no., pp.343,349, 20-24 Aug. 2012

[46] Ngo; C., Demchenko, Y.; de Laat, C., "Toward a Dynamic Trust Establishment approach for multi-provider Intercloud environment," Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on , vol., no., pp.532,538, 3-6 Dec. 2012

[47] Ngo, C., M. Makkes, Y. Demchenko and C. de Laat, "Multi-data-types Interval Decision Diagrams for XACML Evaluation Engine", 11th International Conference on Privacy, Security and Trust 2013 (PST 2013), July 10-12, 2013 (to be published).

[48] MongoDB [online] http://www.mongodb.org/

[49] Apache HBase [online] http://hbase.apache.org/

[50] Apache Cassandra [online] http://cassandra.apache.org/

[51] Apache Accumulo [online] http://accumulo.apache.org/

[52] Goyal, V., O.Pandeyy A.Sahaiz, B.Waters, Attribute-Based Encryption for Fine-Grained Access Control of Encrypted Data. Proceeding CCS '06 Proceedings of the 13th ACM conference on Computer and communications security [online] http://research.microsoft.com/en-us/um/people/vipul/abe.pdf

[53] Chase, M., Multi-Authority Attribute Based Encryption. ProceedingTCC'07 Proceedings of the 4th conference on Theory of cryptography. http://cs.brown.edu/~mchase/papers/multiabe.pdf

[54] Demchenko Y., Leon Gommans, Cees de Laat, "Extending User-Controlled Security Domain with TPM/TCG in Grid-based Virtual Collaborative Environment". In Proceedings The 2007 International Symposium on Collaborative Technologies and Systems (CTS 2007), May 21-25, 2007, Orlando, FL, USA. ISBN: 0-9785699-1-1. Pp. 57-65.

[55] Membrey, P., K.C.C.Chan, C.Ngo, Y.Demchenko, C. de Laat, Trusted Virtual Infrastructure Bootstrapping for On Demand Services. The 7th International Conference on Availability, Reliability and Security (AReS 2012), 20-24 August 2012, Prague. ISBN 978-0-7695-4775-6

[56] Yahalom, R., B. Klein, and T. Beth, "Trust relationships in secure systems-a dis- tributed authentication perspective," in Research in Security and Privacy, 1993. Proceedings., 1993 IEEE Computer Society Symposium on. IEEE, 1993, pp. 150–164.

[57] Brickell, E., J.Camenisch, and L. Chen, "Direct anonymous attestation," Proc. of the 11th ACM conference on Trust and Security in Computer Systems, 2004. [Online]. Available: http://portal.acm.org/citation.cfm?id=1030083.1030103

[58] Research Data Alliance (RDA). [online] http://rd-alliance.org/

Appendix A Big Data Architecture Definition by Industry Associations

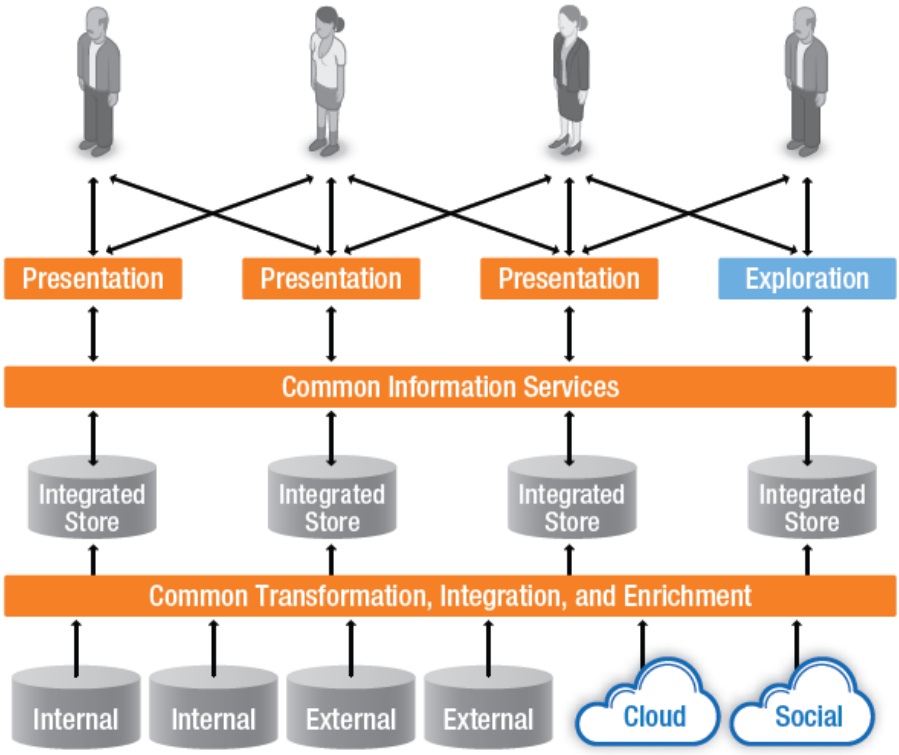
This section will provide overview and analysis of the Big Data architecture models proposed by different industry associations and leading technology companies.

- 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

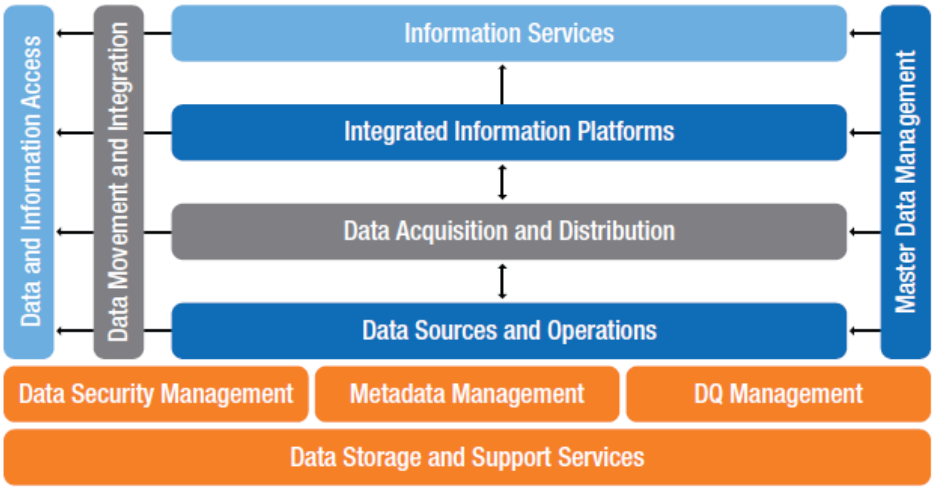
A.1 Open Data Center Alliance (ODCA) Information as a Service (INFOaaS)

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

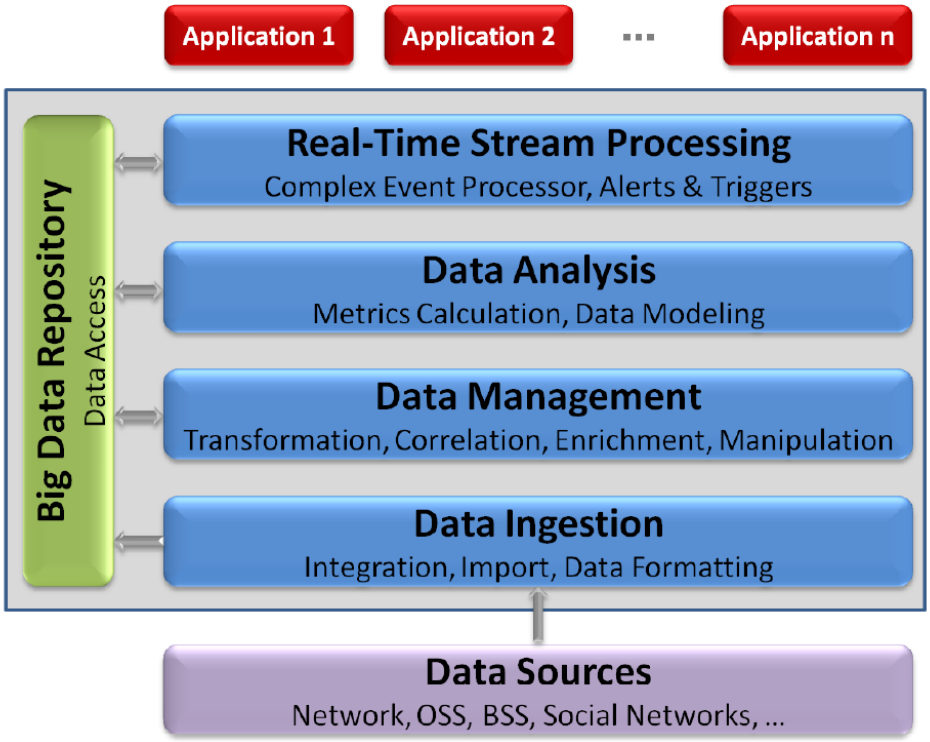
[ref] Open Data Center Alliance Master Usage model: Information as a Service, Rev 1.0. http://www.opendatacenteralliance.org/docs/Information_as_a_Service_Master_Usage_Model_Rev1.0.pdf



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

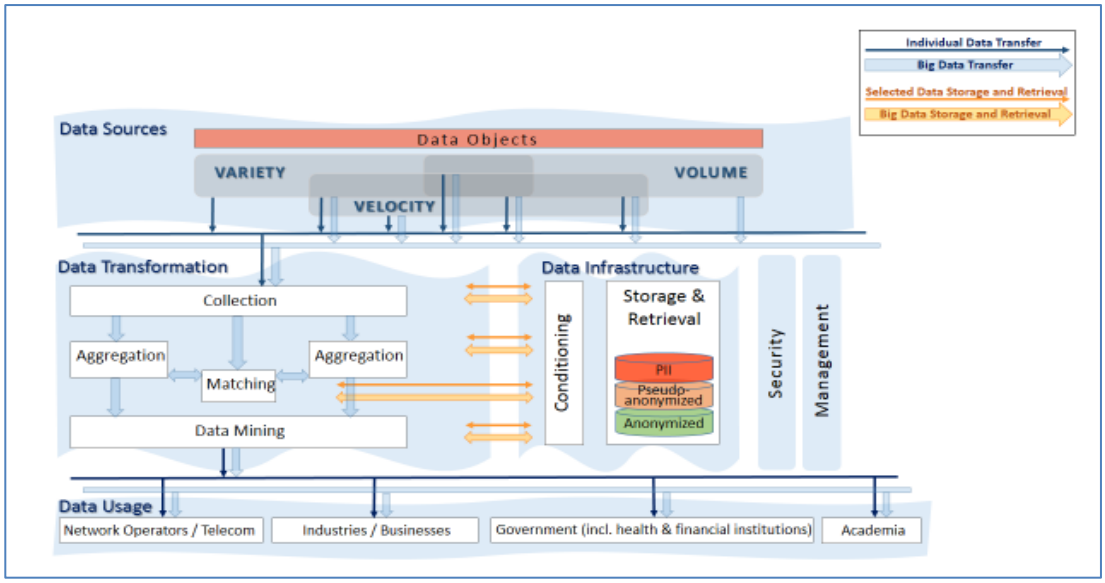


A.2 TMF Big Data Analytics Reference Architecture



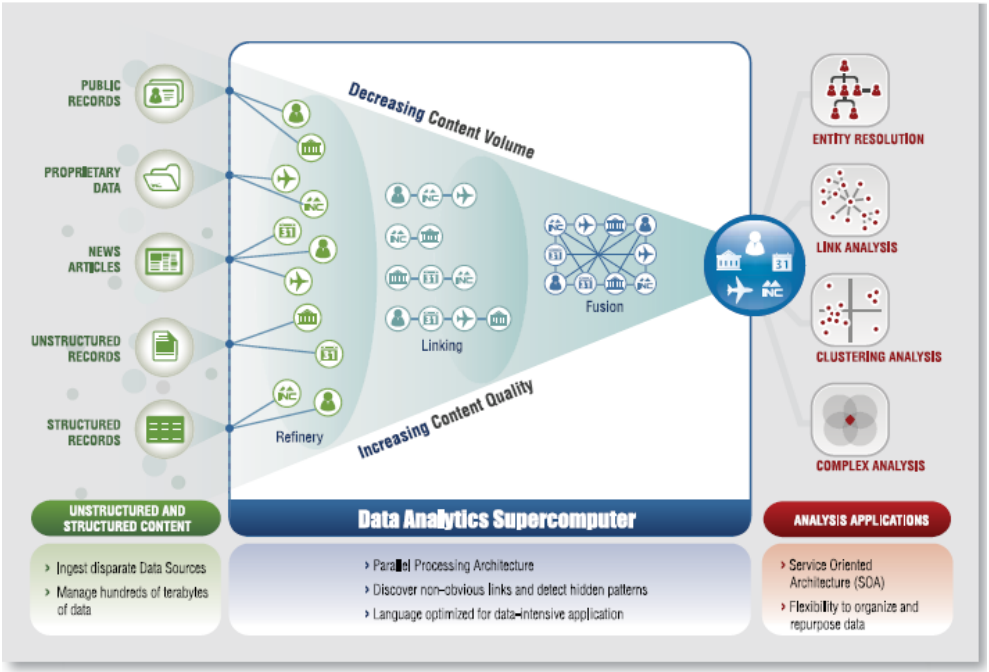
A.3 Big Data Ecosystem Reference Architecture (By Microsoft)

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



A.4 LexisNexis Vision for Data Analytics Supercomputer (DAS)

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

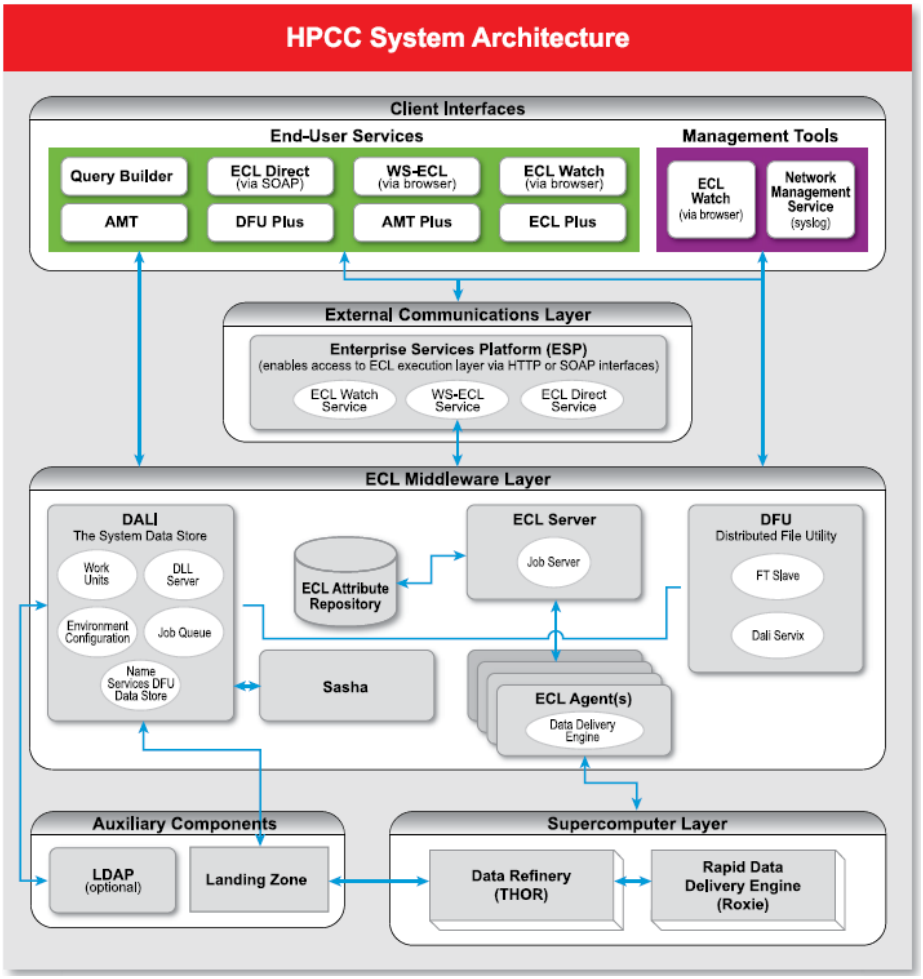


LexisNexis HPCC System Architecture

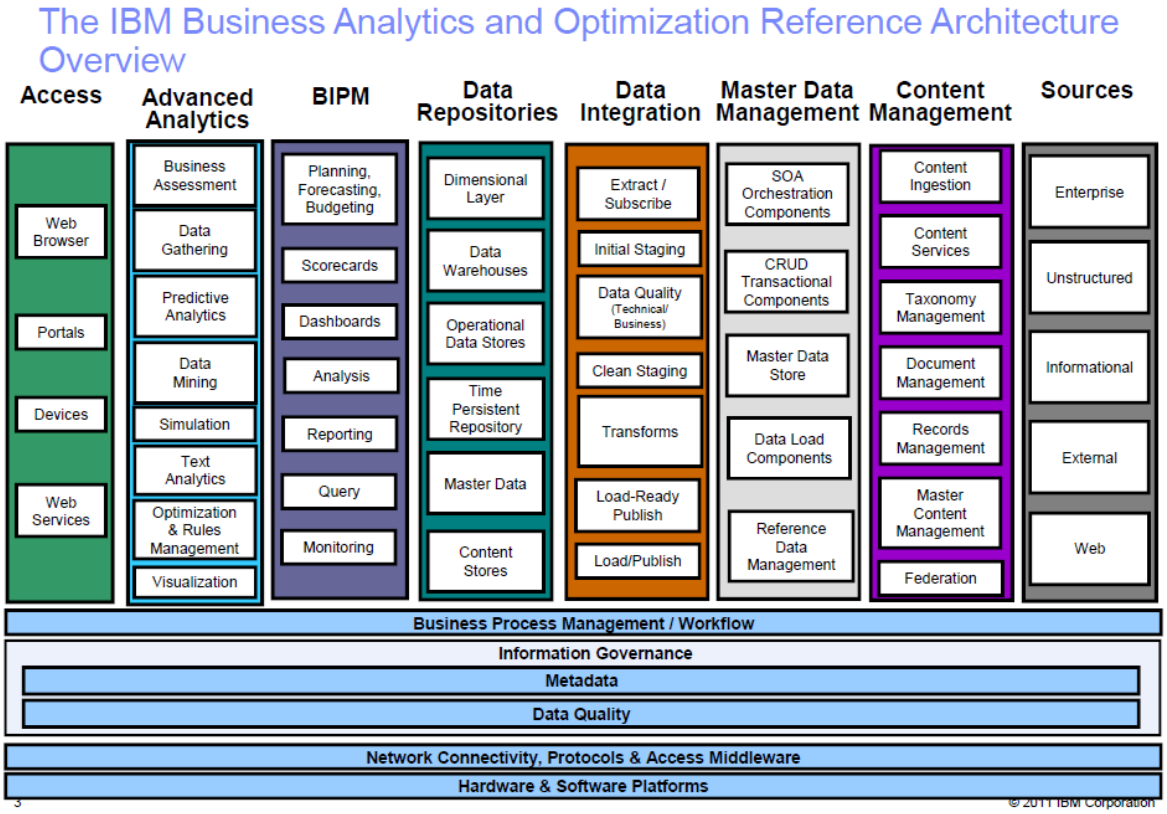
- ECL – Enterprise Data Control Language

- 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

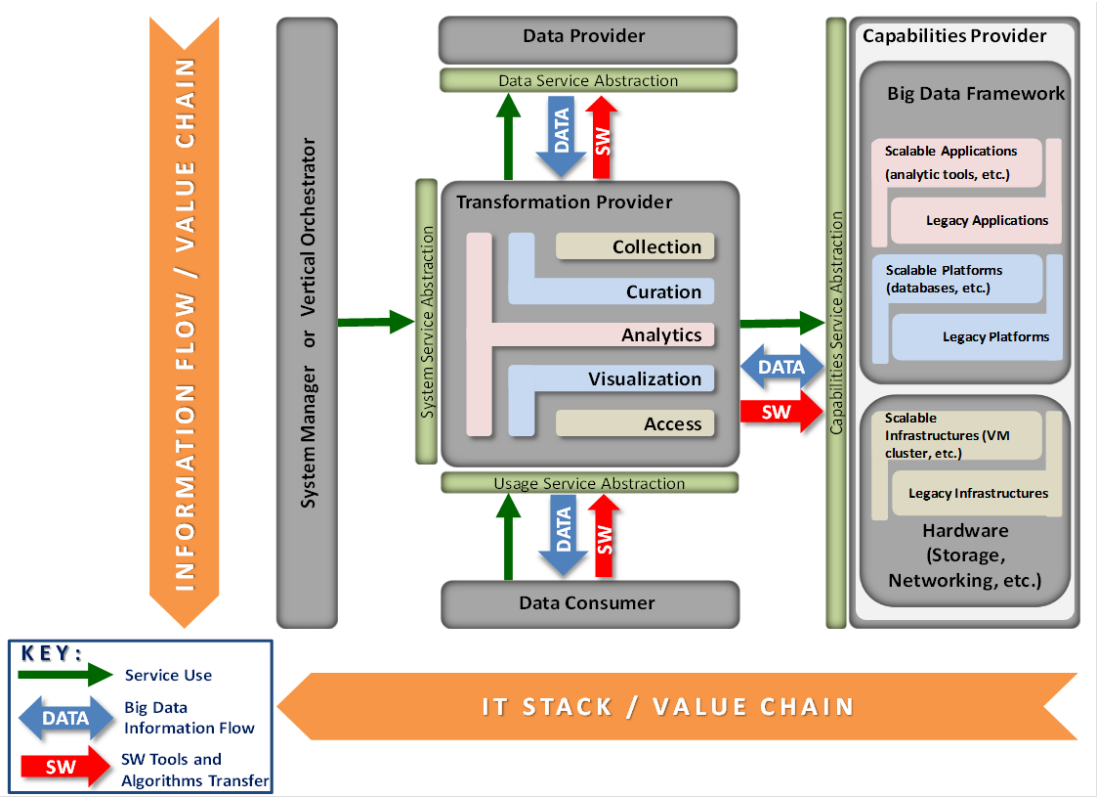


A.5 The IBM Business Analytics and Optimisation Reference Architecture



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

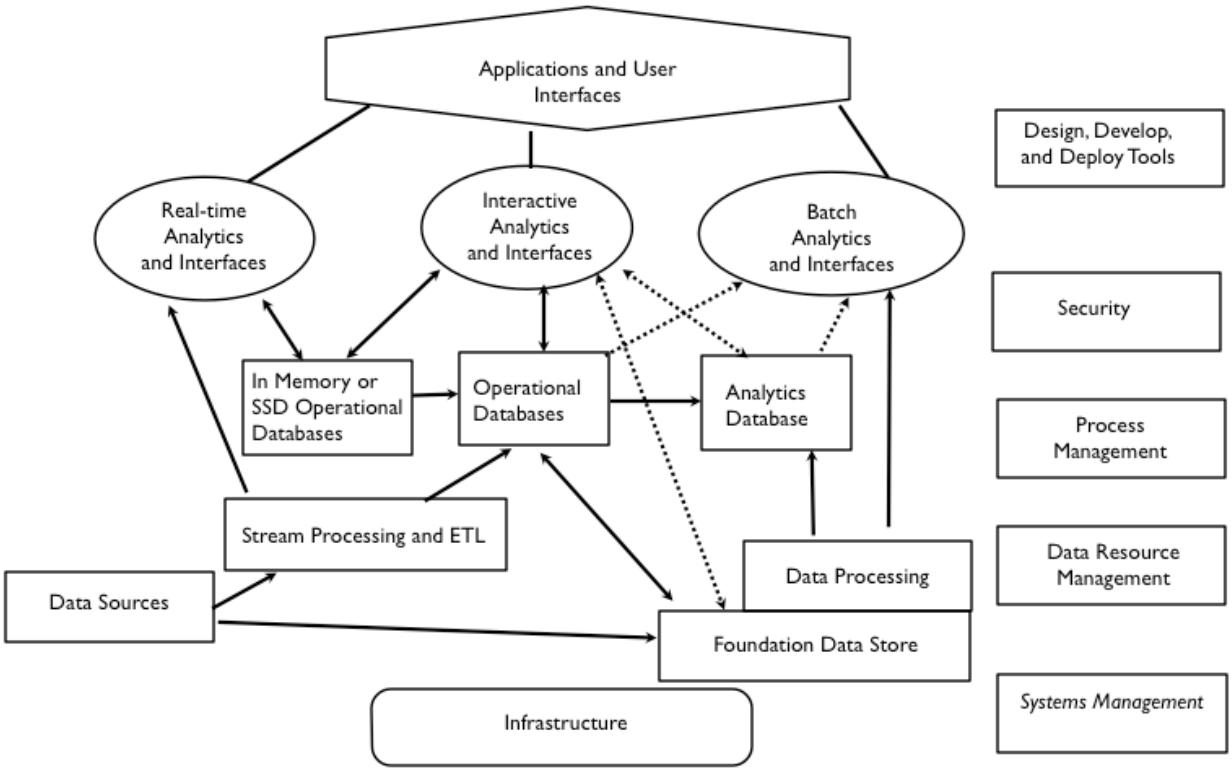
A.6 NIST Proposed Reference Architecture (Initial as of July 2013)



NIST Big Data Architecture of July 2013:

- 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


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

Appendix B NIST Big Data Working Group (NBD-WG) Activities and Documents

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

BigData WG mailing list:

- Input documents http://bigdatawg.nist.gov/show_InputDoc2.php

BigData WG useful documents

- 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