Defining Architecture Components of the Big Data Ecosystem

Demchenko, Y.; de Laat, C.; Membrey, P.

DOI
10.1109/CTS.2014.6867550

Publication date
2014

Document Version
Author accepted manuscript

Published in
Proceedings of the 2014 International Conference on Collaboration Technologies and Systems

Defining Architecture Components of the Big Data Ecosystem

Yuri Demchenko, Cees de Laat
System and Network Engineering Group
University of Amsterdam
Amsterdam, The Netherlands
e-mail: {y.demchenko, C.T.A.M.deLaat}@uva.nl

Peter Membrey
Hong Kong Polytechnic University
Hong Kong SAR, China
e-mail: cspmembrey@comp.polyu.edu.hk

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.**

Keywords- Big Data Technology, Big Data Ecosystem, Big Data Architecture Framework (BDAF), Big Data Infrastructure (BDI), Big Data Lifecycle Management (BDLM), Cloud based Big Data Infrastructure Services.

I. 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 component technology (e.g. Data Analytics or Machine Learning) 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 VII refers to other works related to defining Big Data architecture and its components. The paper concludes with the summary and suggestions for further research.

II. Big Data Definition and Analysis

A. *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. We refer to the discussion in [5] presented by the authors at NBDWG about relations between these two dimensions to indicate their dependence. 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

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.

B. *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 in blogosphere where active contributors have generally converged on the most important features and incentives of the Big Data [6, 7, 8].

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.

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

During the Big Data lifecycle, each stage of the data transformation or processing changes the dataset content, state and consequently may change/enrich 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 may have a definite state, besides commonly defined as data in move, in rest, or being processed. Supporting these data properly will require scalable provenance models and tools incorporating also data integrity and confidentiality.

C. From 5V to 5 Parts Big Data Definition

It is obvious that current Big Data definition addresses only three basic Big Data properties Volume, Velocity, Variety (so-called 3V) and related technology components. To improve and extend the Big Data definition as a new technology, we need to find a way to reflect its all important features and provide a guidance/basis for further technology development. We can refer to one of the best example of the Cloud Computing definition [18] that has been given by NIST 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 infrastructure 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/Evolution

(3) New Analytics

- Real-time/streaming analytics, interactive and machine learning analytics

(4) New Infrastructure and Tools

- Cloud based infrastructure, storage, network, high performance computing
- 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 that are important aspect sometimes defining data types and data structures, e.g. raw data, data streams, correlated data

- 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 the major Big Data features and ecosystem components, 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 allow 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."

D. 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 data related processes, source, target, and outcome.

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

III. Paradigm change in Big Data and Data Intesive 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 or technologists, including wide public access to scientific or production data.
- Existence of necessary infrastructure components and management tools that allow fast infrastructure 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 both general and security paradigms change in Big Data ecosystem:

- 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 practice that assumes data access, use by different user groups and in general processing on the third party facilities/datacenters, 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 which is an important component of the future BDI [3, 16, 17].

A. *From Big Data to All-Data Methaphor*

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 are 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 and all this hype is not for them.

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

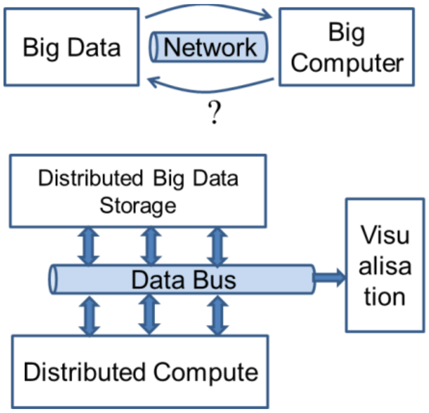

Figure 1. From Big Data to All-Data Metaphor.

B. *Moving to Data-Centric Models and Technologies*

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 that use server/host based PKI certificates and security protocols. 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, in particular to support data provenance,
- Data location, search, access
- Data integrity and identifiability, referral integrity
- Data security and data centric access control, encryption enforced and attribute based access
- Data ownership, personally identified data, privacy, opacity
- Trusted virtualisation platform, data centric trust bootstrapping

IV. Proposed Big Data Architecture Framework

Discussion above motivates 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 industry best practices and our experience in defining architectures for new technologies, in particular, NIST Cloud Computing Reference Architecture (CCRA) [18], Intercloud Architecture Framework (ICAF) by authors [19], recent documents by the NIST Big Data Working Group [4], in particular initial Big Data Reference Architecture [20] or Big Data technology Roadmap [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].

The proposed definition of the Big Data Architecture Framework summarises majority of the known to us research and discussions in this area. The proposed BDAF comprises of the following 5 components that address different aspects of the Big Data Ecosystem and Big Data definition aspects which we consider to some extent orthogonal and complementary:

(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 2 that shows what architecture component is used or required by another component.

TABLE 3. INTERRELATION BETWEEN BDAF COMPONENTS

Coln: 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 reflecting the technology oriented stakeholders. The further research on the BDAF definition should analyse the interests and messages related to different stakeholder groups in Big Data, in particular we will be looking for contribution from the data archives providers and libraries who are expected to play a renewed role in the BDE [26].

A. *Data Models and Structures*

Different stages of the Big Data transformation will require different data structures, models and formats, including also a possibility to process both structured and unstructured data [27].

The following data types can be defined according to current NBDWG discussions [28]:

(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 in many cases it is important to keep linkage between data.

We can look closer at the scientific data types, their transformation and related requirements where we have long time experience. 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.

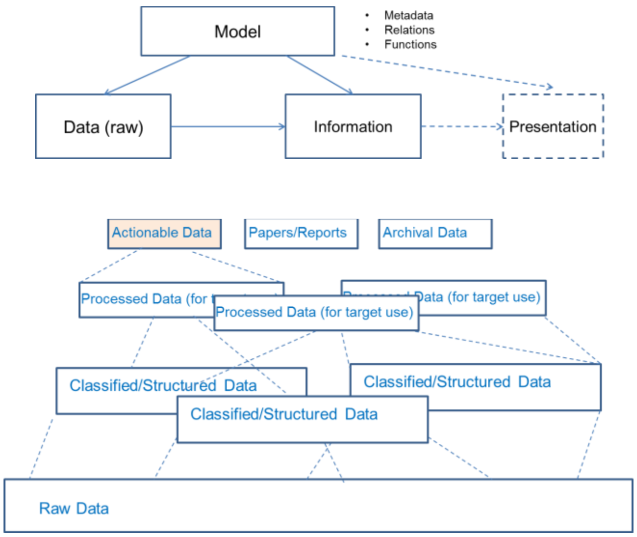
Model
Metadata
Relations
Functions
Data (raw)
Information
Presentation
Actionable Data
Papers/Reports
Archival Data
Processed Data (for target use)
Processed Data (for target use)
Processed Data (for target use)
Classified/Structured Data
Classified/Structured Data
Classified/Structured Data
Raw Data

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

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 [29].

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 that 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].

B. *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. Numerous blog articles [6, 33] and industry papers [34, 35] suggest 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.

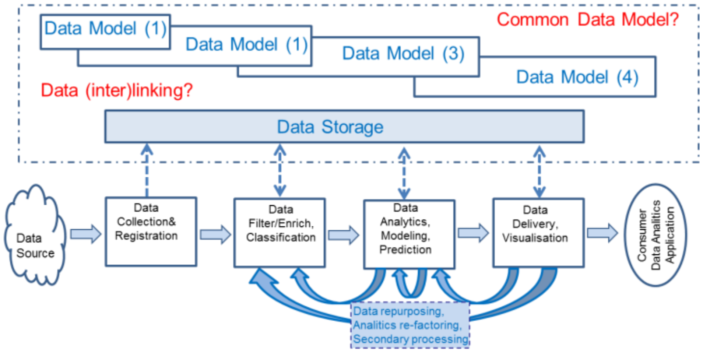
Common Data Model?
Data Model (1)
Data Model (1)
Data Model (3)
Data Model (4)
Data (inter)linking?
Data Storage
Data Source
Data Collection& Registration
Data Filter/Enrich, Classification
Data Analytics, Modeling, Prediction
Data Delivery, Visualisation
Consumer Data Analitics Application
Data repurposing, Analitics re-factoring, Secondary processing

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 [36] 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) proposed as a result of analysis of the existing practices in different scientific communities and industry technology domains.

New BDLM requires data storage and preservation at all stages what should allow data re-use/re-purposing and secondary research/analytics 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 lifecycle. Data curation is an important component of the discussed BDLM and must also be done in a secure and trustworthy way.

V. 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, which in their own turn will require high-performance low-latency network.

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)

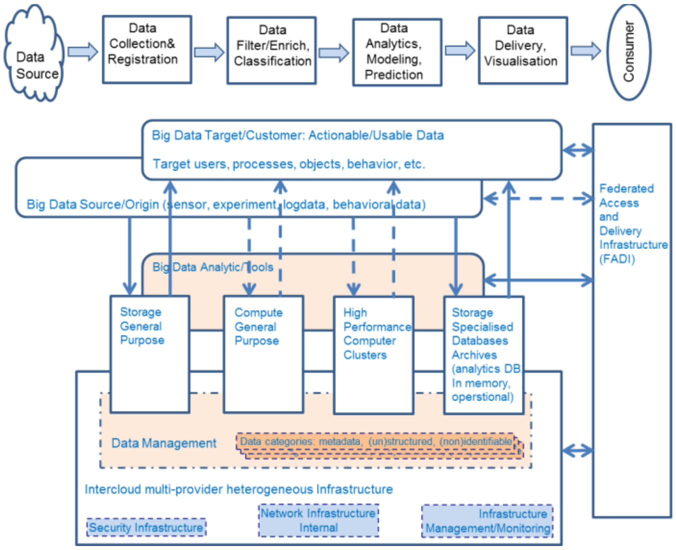
Data Source
Data Collection& Registration
Data Filter/Enrich, Classification
Data Analytics, Modeling, Prediction
Data Delivery, Visualisation
Consumer
Big Data Target/Customer: Actionable/Usable Data
Target users, processes, objects, behavior, etc.
Big Data Source/Origin (sensor, experiment, logdata, behavioral data)
Federated Access and Delivery Infrastructure (FADI)
Big Data Analytic/Tools
Storage General Purpose
Compute General Purpose
High Performance Computer Clusters
Storage Specialised Databases Archives (analytics DB In memory, operstional)
Data Management
Data categories: metadata, (un)structured, (non)identifiable
Intercloud multi-provider heterogeneous Infrastructure
Security Infrastructure
Network Infrastructure Internal
Infrastructure Management/Monitoring

Figure 4. General Big Data Infrastructure functional components

We define Federated Access and Delivery Infrastructure (FADI) as an important component of the general BDI that interconnects different components of the cloud/Intercloud based infrastructure combining dedicated network connectivity provisioning and federated access control [19, 37].

A. 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 [23, 24, 38] which we will commonly refer to as Big Data Analytics Infrastructure (BDAI):

- 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

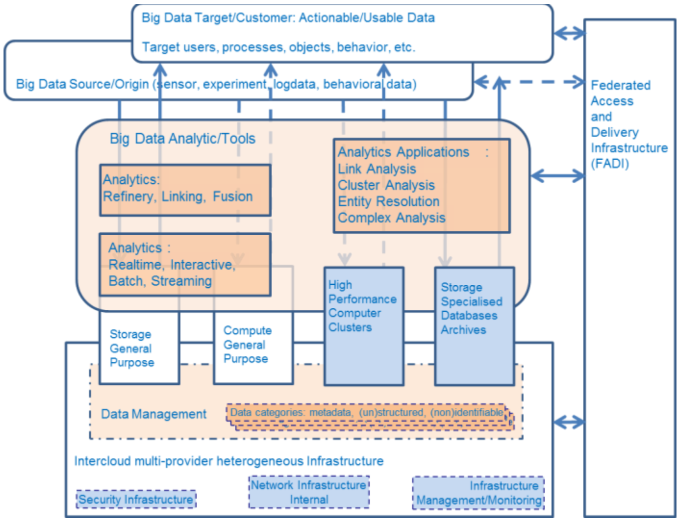
Big Data Target/Customer: Actionable/Usable Data
Target users, processes, objects, behavior, etc.
Big Data Source/Origin (sensor, experiment, logdata, behavioral data)
Federated Access and Delivery Infrastructure (FADI)
Big Data Analytic/Tools
Analytics: Refinery, Linking, Fusion
Analytics : Realtime, Interactive, Batch, Streaming
Analytics Applications : Link Analysis Cluster Analysis Entity Resolution Complex Analysis
Storage General Purpose
Compute General Purpose
High Performance Computer Clusters
Storage Specialised Databases Archives
Data Management
Data categories: metadata, (un)structured, (non)identifiable
Intercloud multi-provider heterogeneous Infrastructure
Security Infrastructure
Network Infrastructure Internal
Infrastructure Management/Monitoring

Figure 5. Big Data Analytics Infrastructure components

Big Data analytics tools are currently offered by the major cloud services providers such as: Amazon Elastic MapReduce and Dynamo [39], Microsoft Azure HDInsight [40], IBM Big Data Analytics [41]. Scalable Hadoop and data analytics tools services are offered by few companies that position themselves as Big Data companies such as Cloudera, [42] and few others [43].

VI. 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 6 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.

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, 37].

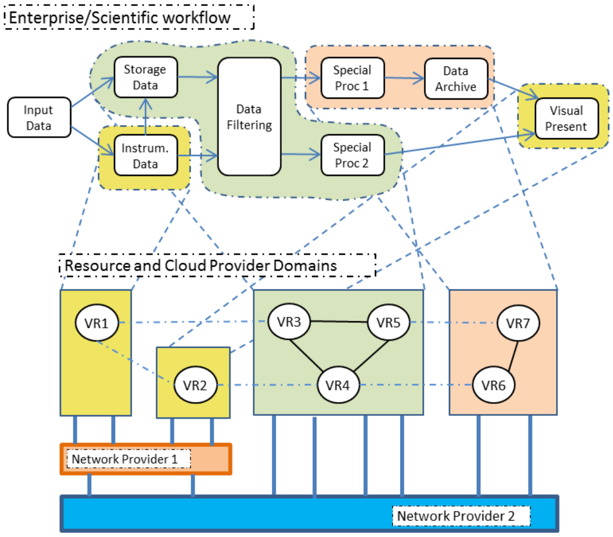

Figure 6. From scientific workflow to cloud based infrastructure.

VII. Related work

There are not many academic papers related to the definition of the Big Data Architecture or its components. Due to the specifics of this paper that intends to explore a new emerging technology domain, we have widely researched both currently existing publications related to the Big Data technology and research papers and best practices documents from other domains that could contribute to the definition of the proposed Big Data Architecture Framework. A number of publications, standards, and industry best practices have been mentioned and cited in this paper. Here we just mention these works that we consider as a foundation for our work. The authors actively contribute to the NIST Big Data Working Group that provides a good forum for discussion but have plans to produce initial set of the draft documents only by the end of September 2013. The following publications contribute to the research on the Big Data Architecture: 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].

VIII. Future Research and Development

The future research and development will include further Big Data definition initially presented in this paper. At this stage we attempted 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, also analyzing different stakeholder groups.

The authors will extend their research into defining the Big Data Security Framework with the specific focus on data centric security that should allow secure data storage, transfer and processing in distributed data storage and processing infrastructure.

The authors are also looking into defining data structures for high performance streaming applications and developing new types of disk based stream oriented data bases, continuing the work started from the authors work on CakeDB database [44].

The authors will continue contributing to the NIST Big Data WG targeting both goals to propose own approach and to validate it against the 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) [45].

The authors believe that the presented paper will contribute toward the definition of the Big Data Architecture Framework and provide a basis for wider discussion to define a new research and technology domain.

Acknowledgements

This work is supported by the FP7 EU funded projects GN3plus and EUBrazil Cloud Connect. The authors also express acknowledgement to the members of the Big Data Interest Group at the University of Amsterdam for contribution to the discussion on Big Data Architecture framework and provided valuable advices regarding use cases, suggested technology use, and basic operational models.

References

[1] Global Research Data Infrastructures: Towards a 10-year vision for global research data infrastructures. Final Roadmap, March 2012. [Online]. Available: 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] Y.Demchenko, P.Membrey, P.Grosso, C. de Laat, "Addressing Big Data Issues in Scientific Data Infrastructure," in *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]. Available: http://bigdatawg.nist.gov/home.php

[5] Defining Big Data Architetcure Framework: Outcome of the Brainstorming Session at the University of Amsterdam, 17 July 2013. Presented at NBD-WG, 24 July 2013 [Online]. Available: 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] Available 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]. Available: http://strata.oreilly.com/2012/01/what-is-big-data.html

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

[9] J.Gantz and David Reinsel, Extracting Value from Chaos, *IDC IVIEW, June 2011*. [Online]. Available: 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]. Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/

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

[12] S.Sicular, "Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s", Gartner, Inc. 27 March 2013. [Online]. Available: 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] J.Layton, "The Top of the Big Data Stack: Database Applications", July 27, 2012. [Online]. Available: http://www.enterprisestorageforum.com/storage-management/the-top-of-the-big-data-stack-database-applications.html

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

[15] A.Bloom, 7 Myths on Big Data: Avoiding Bad Hadoop and Cloud Analytics Decisions, April 22, 2013. [Online]. Available: 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: Available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/aaa-study-final-report.pdf

[17] Y.Demchenko 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]. Available: http://collaborate.nist.gov/twiki-cloud-computing/pub/CloudComputing/ReferenceArchitectureTaxonomy/NIST_SP_500-292_-_090611.pdf

[19] Y.Demchenko, 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] NIST Big Data Reference Architecture. NBD-WG, NIST [Online]. Available: http://bigdatawg.nist.gov/_uploadfiles/M0226_v10_1554566513.docx

[21] NIST Big Data Technology Roadmap. NBD-WG [Online]. Available: http://bigdatawg.nist.gov/_uploadfiles/M0087_v8_1456721868.docx

[22] Open Data Center Alliance Master Usage model: Information as a Service, Rev 1.0. [Online]. Available: 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). IBM, 2013. [Online]. Available: 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] A.M.Middleton, HPCC Systems: Introduction to HPCC (High Performance Computer Cluster), LexisNexis Risk Solutions, LexiNexis, May 24, 2011

[26] Bierauge, M., Keeping Up With... Big Data. American Library Association, 2013 [Online]. Available: http://www.ala.org/acrl/publications/keeping_up_with/big_data

[27] Unstructured Data Management, Hitachi Data System, 2013. [online] http://www.hds.com/solutions/it-strategies/unstructured-data-management.html

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

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

[30] Open Access: Opportunities and Challenges. European Commission for UNESCO. [Online]. Available: 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]. Available: http://www.openaire.eu/

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

[33] Roundup of Big Data Pundits' Predictions for 2013. Blog post by David Pittman. January 18, 2013. [Online]. Available:http://www.ibmbigdatahub.com/blog/roundup-big-data-pundits-predictions-2013

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

[35] Big data: The next frontier for innovation, competition, and productivity, May 2011. McKinsey Global Institute. [Online]. Available: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

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

[37] M.Makkes, C.Ngo, Y.Demchenko, R.Strijkers, R.Meijer, C. de Laat, "Defining Intercloud Federation Framework for Multi-provider Cloud Services Integration", in *The Fourth International Conference on Cloud Computing, GRIDs, and Virtualization (CLOUD COMPUTING 2013)*, May 27 - June 1, 2013,Valencia, Spain.

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

[39] Amazon Big Data. [Online]. Available: http://aws.amazon.com/big-data/

[40] Microsoft Azure Big Data. Microsoft, 2013 [Online]. Available: http://www.windowsazure.com/en-us/home/scenarios/big-data/

[41] IBM Big Data Analytics. IBM, 2o13 [Online]. Available: http://www-01.ibm.com/software/data/infosphere/bigdata-analytics.html

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

[43] 10 hot big data startups to watch in 2013, 10 January 2013 [Online]. Available: http://beautifuldata.net/2013/01/10-hot-big-data-startups-to-watch-in-2013/

[44] P.Membrey, K.C.C. Chan, Y.Demchenko, A Disk Based Stream Oriented Approach For Storing Big Data. In *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.

[45] Research Data Alliance (RDA). [Online]. Available: http://rd-alliance.org/