Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure

Yuri Demchenko[1], Canh Ngo[1], Cees de Laat[1], Peter Membrey[2], Daniil Gordijenko[3]

[1]University of Amsterdam, The Netherlands
[2]Hong Kong Polytechnic University, Hong Kong
[3]Inoitech S.a.r.l., Luxembourg
{y.demchenko, c.t.ngo, C.T.A.M.deLaat}@uva.nl
cspmembrey@comp.polyu.edu.hk
dgordijenko@inoitech.eu

Abstract. Big Data technologies are changing the traditional technology domains and their successful use will require new security models and new security design approaches to address emerging security challenges. This paper intends to provide initial analysis of the security issues and challenges in Big Data and map new challenges and problems to the traditional security domains and technologies. The paper starts with the Big Data definition and discusses the features that impact the most the Big Data security, such as Veracity, Volume, Variety, and dynamicity. The paper analyses the paradigm change and new challenges to Big Data security. The paper refers to the generic Scientific Data Infrastructure (SDI) model and discusses security services related to the proposed Federated Access and Delivery Infrastructure (FADI) that serves as an integration layer for potentially multi-provider multi-domain federated project oriented services infrastructure. The paper provides suggestions for practical implementation of such important security infrastructure components as federated access control and identity management, fine-grained data-centric access control policies, and the Dynamic Infrastructure Trust Bootstrap Protocol (DITBP) that allows deploying trusted remote virtualised data processing environment. The paper refers to the past and ongoing project experience by authors and discusses how this experience can be consolidated to address new Big Data security challenges identified in this paper.

Keywords: Big Data Security, Federated Access and Delivery Infrastructure (FADI), Trusted Virtualised Environment, Cloud Infrastructure Services.

1 Introduction

Big Data and 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. Consequently, emerging data intensive technologies impose new challenges to traditional security technologies that may require re-thinking and re-factoring currently used security models and tools.

In e-Science and industry, the scientific data and technological data are complex multifaceted objects with the complex internal relations and typically distributed between different systems and locations. They are becoming an infrastructure of their own and need to be supported by corresponding physical or logical infrastructures to store, access, process and manage these data. We refer to such infrastructure as Scientific Data Infrastructure (SDI) or Big Data Infrastructure (BDI) in general. We argue that both SDI and BDI should provide capabilities to support collaborative groups of researchers or technologists due to complex character of the research projects or production processes.

The goal of this paper is to understand the 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 existing security models and tools and evaluate their potentiality to be used with Big Data technologies.

There is no well-established terminology in the area of Big Data. Expectedly this problem will be solved by the recently established NIST Big Data Working Group [4]. In this paper we primarily focus on the security issues related to Big Data and in many case use terms Big Data technologies, Data Intensive Technologies and Big Data Science as interchangeable depending on the context.

The authors made an initial attempt in their recent papers [5, 6] to summarise related Big Data discussions and provide a definition of the 5V of Big Data: Volume, Velocity, Variety, Value, and Veracity, as the main properties of the Big Data that create a challenge to modern technologies. In this paper we continue with the Big Data definition and primarily focus on the security related aspects.

The paper is organised as follows. Section 2 looks into Big Data definition and Big Data nature in science, industry, and business, analyses factors that impact security. Section 3 gives a short overview of related research and developments. Section IV discusses security challenges to Big Data infrastructure and Big Data challenges to traditional security models. Section 4 discusses paradigm shift in Big Data security and new challenges to be addressed. Section 5 briefly discussed data management and proposes the Scientific Data Lifecycle Management model, identifies security and trust related issues in handling data, and summarises the general requirements and design suggestions for cloud based Big Data security infrastructure. Section 6 discusses the main components of the consistent cloud based security infrastructure for Big Data: Federated Access and Delivery Infrastructure, fine granular data centric policy definition, and Virtual Infrastructure Trust Bootstrapping protocol. Section 7 provides suggestions for the future research and developments.

2 Big Data Definition and Security Properties

2.1 Big Data Nature in e-Science, Industry and Business

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. To show the specifics of Big Data properties and use, we can distinguish the following Big Data domains: e-Science/research, industry, and business, leaving analysis of other domains for future research.

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 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.

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.

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.

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, 7, 8] 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.

2.2 5 Vs of Big Data and Data Veracity

Despite the "Big Data" became a new buzz-word, there is no consistent definition for Big Data, nor detailed analysis of this new emerging technology. Most discussions are

going now in blogosphere where active contributors have generally converged on the most important features and incentives of the Big Data [2, 8, 9, 10]. In our recent paper [6] we summarised existing definitions and proposed a consolidated view on the generic Big Data features that was used to define the general requirements to Scientific Data Infrastructure. In this paper we provide a short summary and discuss the main Big Data properties that impose new security challenges.

For the completeness of the discussion, we quote here 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" [10]. It can be complemented more simple definition from [11]: "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 [12].

We refer to the Big Data definition proposed in our recent paper [6] as having the following 5V properties: Volume, Velocity, Variety, Value, and Veracity, as illustrated in Figure 1. We also highlight the security related properties Veracity, Variety and Volume (by the density of the property representing circles).

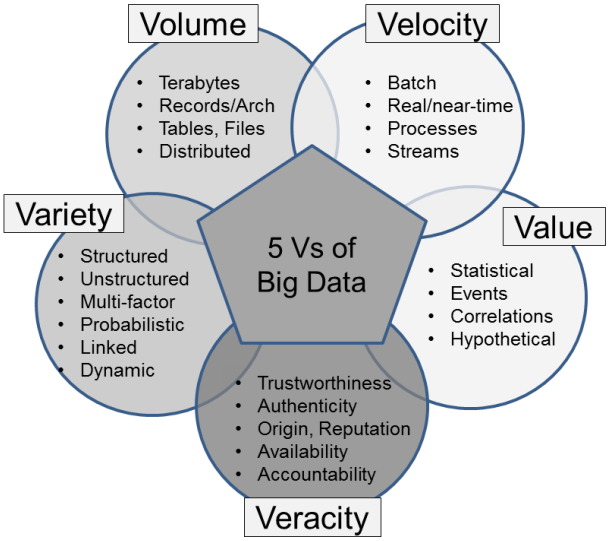

Figure 1. 5 Vs of Big Data and security related properties of Veracity, Variety, and Volume.

1) Veracity

Veracity property of Big Data is directly related to the Big Data security and includes two aspects: data consistency (or certainty) what can be defined by their statistical reliability; and data trustworthiness that is defined by a number of factors including data origin, collection and processing methods, including trusted infrastructure and facility.

Big Data veracity ensures that the data used are trusted, authentic and protected from unauthorised access and modification. The data must be secured during the whole their lifecycle from collection from trusted sources to processing on trusted compute facilities and storage on protected and trusted storage facilities.

The following aspects define and need to be addressed to ensure data veracity:

- Integrity of data and linked data (e.g., for complex hierarchical data, distributed data with linked metadata)
- Data authenticity and (trusted) origin
- Identification of both data and source
- Computer and storage platform trustworthiness
- Availability and timeliness
- Accountability and Reputation

Data veracity relies entirely on the security infrastructure deployed and available from the Big Data infrastructure. Data provenance is an important mechanism to ensure data Veracity.

2) Other impact factors: Volume, Variety and Dynamicity

Security and privacy issues are magnified by volume, variety, and Big Data dynamicity (or variability). The latter is originated from the fact that data change their structure, model, content, and may migrate between datacenters and clouds during their lifecycle.

Volume as the main generic feature of the Big Data provides also challenges to current security technologies that need to scale the size of Big Data, also taking into account their distributed character.

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 Related Research and Developments

There is not much academic works on Big Data security. The research community currently is in the process of identifying the potential research areas. However, many new research works that attempt to review the very basic security concepts and models can be potentially extended to the Big Data related challenges and problems.

First serious attempts of tackling this problem have been undertaken by the NIST by organising the two workshops in 2012 and 2013 related to Big Data [13] and establishing the Big Data Working Group [4] in July 2013. The Cloud Security Alliance (CSA) has established in 2012 the Big Data Security Working Group [14].

3.1 CSA Top Ten

Recently the CSA Big Data Security WG has published its first deliverable “Top Ten Big Data Security and Privacy Challenges” [15]. The document provides a good insight and initial identification of such challenges but they are clearly defined from the point of view of the Information Security and Service Management and don’t touch

security design issues. In our research and in this paper, we approach the Big Data Security problem from the Security Engineering point of view, providing also analysis of existing security technologies and their applicability and required modification to support Big Data infrastructure and processes.

We find useful to provide a short summary of the CSA Top Ten (refer to the original document [15] for details). We group them into few clusters:

A. Infrastructure security
   1) Secure computations in distributed programming frameworks
   2) Security best practices for non-relational data stores
   3) Secure data storage and transactions logs
   4) End-point input validation/filtering

B. Access control and policy
   5) Granular access control and data centric access policies
   6) Cryptographically enforced access control and secure communication

C. Data Management
   7) Real-time security/compliance monitoring
   8) Granular audits
   9) Data provenance

D. Privacy and Confidentiality
   10) Scalable and composable privacy-preserving data mining and analytics

In this paper, we will discuss different aspects of securing Big Data, identify new security challenges and propose generic security mechanisms to address these challenges.

3.2 Related Security Research

Most of currently used security models, services and mechanisms have been developed for host based, client/server, or service oriented models. Big Data have their specific security requirements, new business models and actors with different relations, and also global scalability character. All this will motivate changing current security services and development of new models and services. For the related research, besides specifically dealing with the Big Data security, we can look also at the recent research that attempt to respond to the changing landscape of the services and technologies with emerging global computing environment, ubiquitous connectivity and proliferation of personal devices, and growth of data centric applications, in particular in healthcare, behavioral and bio-science.

We found a number of interesting conceptual and innovative papers presented at the New Security Paradigms Workshop in the past 3 years. In particular, paper [16] looks at a new “clean slate” approach to the security problems originated from the healthcare that currently becomes increasingly computerized and data intensive. The healthcare use case can be one of reference cases to solve the whole bunch of the data protection related problems. Paper [17] analyses the VM and services virtualization based security models and evaluate their effectiveness. Paper [18] looks at the privacy as a process and attempts to provide a theoretical basis for new/future Privacy Enhancing Technologies (PET).

We can also refer to the related work presented at the SDM12 workshop. Paper [19] proposes an approach to build a trustworthy cloud platform motivated by the specific requirements from the healthcare applications to the trustworthiness of the healthcare platforms. The proposed solution is based on using federated cloud-of-cloud architecture to enforce common security and data protection policies in various cloud layers. Paper [20] discusses new provenance models for complex multi-source Web 2.0 data that similar to Big Data can evolve with time.

We find appropriate also to refer to our past works that attempted to review and re-factor different key security problems related to Grid security [21] and cloud security [22].

4 Paradigm Shift and New Challenges

4.1 Paradigm Shift to Data Centric Security

Traditional security models are OS/system based and host/service centric what means the security is either communication protocols based or ensured by the system/OS based security services. The security and administrative domains are the key concepts, around which the security 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.

Big Data will require different data centric security protocols, especially in the situation that 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. Keeping security context and semantic integrity, to support data provenance in particular, will require additional research.

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

- 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.

4.2 Trusted Virtualisation Platforms

In many cases the companies or users need to store or process their data on the provider facilities in the environment that is not under their control. In most cases they can rely on the provider's business practices but in some cases, both commercially and privacy sensitive, this is not sufficient. Virtualisation technologies enhanced with the trusted computing technologies can potentially provide a basis for developing proper solutions here.

Traditional secure virtualization models are domain and host based. Advancements in services virtualisation (e.g. using Java service container [23]) and developments of the wide scale cloud virtualization platforms [24] provide a sufficiently secure environment for runtime processes but still rely on the trusted hardware and virtualization/hypervisor platform. To address key data-centric (and ownership based) security model it needs to be empowered with the Trusted Computing Platform security mechanisms, in particular, implementing the remote platform trust bootstrapping protocol. We discuss such possible solution in section 6.

4.3 Data ownership

Data ownership will become one of the important concepts in data management and policy definition. Data ownership concept is widely discussed in the context of data governance and personal data protection [25], but there is no well-defined mechanisms to enforce data ownership related policies in the distributed data processing environment. Data centric ownership model is a cross-domain and needs to span the whole data lifecycle. In this respect it is different from the current facility ownership concept in IT, telecommunications and clouds, which is rather provider and domain based. Data ownership is linked to individual or organisational ownership and will affect many currently used security concepts such as identity centric access control and delegation (like implemented in the Auth2.0 protocol [26]), user centric federation and trust model, identity based trust model and data protection mechanisms, data verifiability/audibility.

Federated security models need to adopt the data ownership concept and allow building data centric cross-domains federations. It is also understood that data ownership will impact data provenance and lifecycle management model.

4.4 Personal Information, Privacy and Opacity

Modern services and infrastructure supporting social networks and human activity are tending to be of the scale of humanity, i.e. scaling world-wide (like Facebook) or targeting to support the knowledge base of the whole humanity (like Wikipedia). Their notion of Big Data actually means "ALL (relevant) data". Such systems are unavoidably dealing with the personal identifiable information, despite using existing techniques for information de-identification and anonymisation.

Lot of information can be collected about individuals and privacy protection concerns are known in this area. Big Data will motivate developments of the new privacy protection models in this area. Acknowledging general requirement to protect privacy

and personal data, we still think that existing privacy concepts and PET models will change with the Big Data technologies development and proliferation.

Healthcare system, governmental systems, defense and law enforcement systems will increasingly collect more and more information about individuals. In many cases such information is vitally important for health, life and security. On the other hand, business and service industry will also increasingly collect more information than it is needed to conduct their main business. With modern analytics tool, additional not intended personal information can be extracted from such datasets/collections by linking different datasets and/or applying behavioral analysis.

There is another aspect of the confidentiality or privacy when providing shared datasets services which we define as opacity. The researchers who are in many cases doing competitive research on the shared datasets and/or facilities,like in case of the genome research or LHC experiment, need to trust that their activity (in particular data accessed or applications used) is not tracked and cannot be seen by other competitors. The computing facilities need to make the individual activity opaque although retaining the possibility for data provenance and audit.

5 Security Infrastructure for Big Data

5.1 Scientific Data Lifecycle Management (SDLM)

In Big Data, security needs to be provided consistently during the whole data lifecycle. The generic data lifecycle includes at least the following stages: data acquisition/collection; filtering and classification; processing and analytics; visualization and delivery.

The scientific data lifecycle is more complex and 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 handling in e-Science is reflected in the proposed by the authors the Scientific Data Lifecycle Management (SDLM) model [6, 27], (see Figure 2). The SDLM incorporates both the existing practices researched in [28] and current trends in the Data Intensive Science.

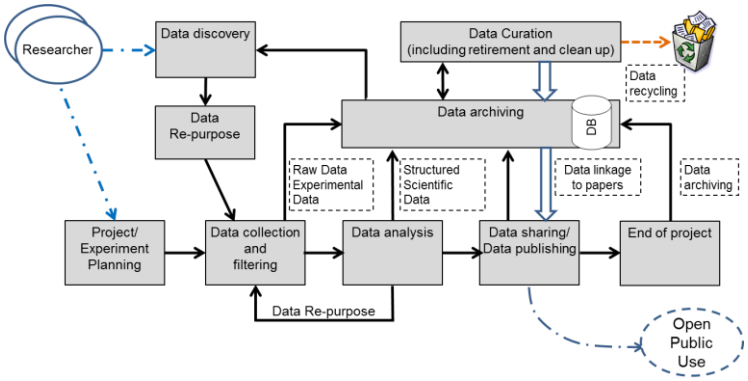

Figure 2. Scientific Data Lifecycle Management in e-Science

The new SDLM 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 SDI.

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 SDI providers [29].

Another factor that will define the SDLM and SDI requirements is the European Commission's initiative to support Open Access [30] to scientific data from publicly funded projects suggests introduction of the following mechanisms to allow linking publications and data: persistent data ID (PDI) [31], and Open Researcher and Contributor Identifier (ORCID) [32].

Data integrity, access control and accountability must be supported during the whole data during lifecycle. Data curation is an important component of the discussed SDLM and must also be done in a secure and trustworthy way.

5.2 Security and Trust in Cloud based Infrastructure

Ensuring data veracity in Big Data infrastructure and applications requires deeper analysis of all factors affecting data security and trustworthiness during their whole lifecycle. Figure 3 illustrates the main actors and their relations when processing data on remote system. User/customer and service provider are the two actors concerned with their own data/content security and each other system/platform trustworthiness: users want to be sure that their data are secure when processed or stored on the remote system.

Figure 3 illustrates the complexity of trust and security relations even in a simple usecase of the direct user/provider interaction. In clouds data security and trust model needs to be extended to distributed, multi-domain and multi-provider environment.

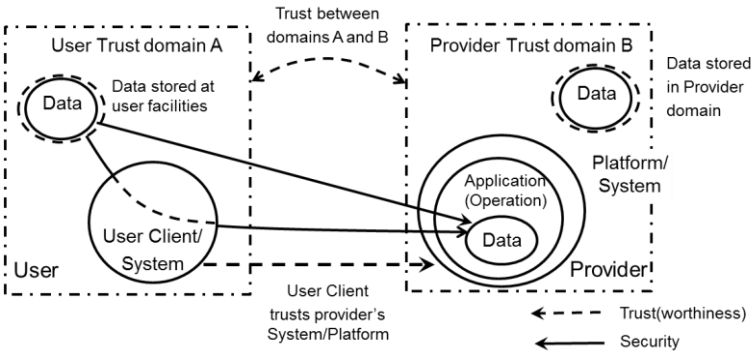

Figure 3. Security and Trust in Data Services and Infrastructure.

5.3 General Requirements to Security Infrastructure

To support secure data processing, the future SDI/BDI should be supported by an corresponding security infrastructure that would ensure normal infrastructure operation, assets and information protection, and allow user identification/authentication and policy enforcement in distributed multi-organisational environment.

Moving to Open Access [30] may require partial change of business practices of currently existing scientific information repositories and libraries, and consequently the future Access Control and Accounting Infrastructure (ACAI) should allow such transition and fine grained access control and flexible policy definition and control.

Taking into account that future SDI/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, the future ACAI should also support all stages of the data lifecycle, including policy attachment to data to ensure persistency of the data policy enforcement during continuous online and offline processes [33].

The required ACAI should support the following features:

- Empower researchers (and make them trust) to do their data processing on shared facilities of large datacentres with guaranteed data and information security
- Motivate/ensure researchers to share/open their research environment to other researchers by providing tools for instantiation of customised pre-configured infrastructures to allow other researchers to work with existing or own data sets.
- Protect data policy, ownership, linkage (with other data sets and newly produced scientific/research data), when providing (long term) data archiving. (Data preservation technologies should themselves ensure data readability and accessibility with the changing technologies).

6 SDI/BDI Security Infrastructure Components

6.1 Federated Access and Delivery Infrastructure (FADI)

In [6] we proposed the generic SDI Architecture model for e-Science (e-SDI) that contains the following layers:

Layer D6: User side and campus based services that may include user portals, identity management services and also visualization facilities.

Layer D5: Federated Access and Delivery Infrastructure (FADI) that interconnects Federation and Policy layer that includes federation infrastructure components, including policy and collaborative user groups support functionality.

Layer D4: (Shared) Scientific platforms and instruments (including potentially distributed/global sensor network) specific for different research areas that also include high performance clusters for Big Data analytics and shared datasets.

Layer D3: Infrastructure virtualisation layer that is represented by the Cloud/Grid infrastructure services and middleware supporting specialised scientific platforms deployment and operation

Layer D2: Datacenters and computing resources/facilities

Layer D1: Network infrastructure layer represented by the general purpose Internet infrastructure and dedicated network infrastructure

Note: "D" prefix denotes relation to data infrastructure.

The proposed SDI reflects the main components required to process, consume and manage data and can easily adopted to the general Big Data Infrastructure.

Modern cloud technologies provide a proper basis for implementing SDI/BDI, in particular for Layer D3 and Layer D4 that correspondingly provide the general infrastructure virtualization platform and shared scientific platform and instruments that typically provide services on-demand for dynamically created virtual groups of users, also called Virtual Organisations. The main efforts to create and operate infrastructure for specific scientific projects will be put into the Layer D5 Federated Access and Delivery Infrastructure (FADI).

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 [34]. 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 service gateway that typically communicates with other domains.

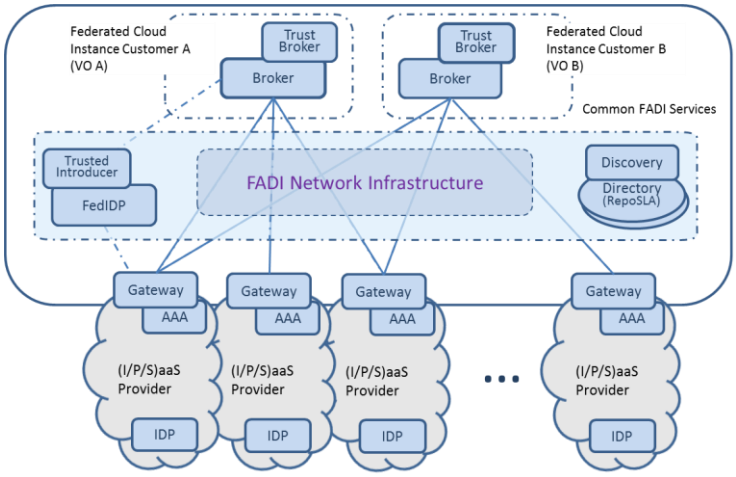

Figure 4. Federated Access and Delivery Infrastructure (FADI)

FADI incorporates related federated infrastructure management and access technologies [34, 35, 36]. Using federation model for integrating multi-provider heterogeneous services and resources reflects current practice in building and managing com-

plex infrastructures (SDI and enterprise infrastructures) and allows for inter-organisational resource sharing.

6.2 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.

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 [22, 37]. 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 [38, 39, 40]. The proposed in [41] the Multi-data-types Interval Decision Diagrams (MIDD) policy decision request evaluation method allows for significant performance gain for massively large policy sets.

2) Access control in NoSQL databases

The popular NoSQL databases for structured data storage MongoDB [42], Cassandra [43], Accumulo [44] 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 [44] 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.

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 [45, 46] 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.

6.3 Trusted Infrastructure Bootstrapping Protocol

To address the issues with creating trusted remote/distributed environment for processing sensitive data, in our earlier papers [47, 48] 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) [49, 50]. The TPM is used to provide a root of trust that extends from the physical hardware itself, and to generate a key pair in hardware where the private key is never revealed (i.e. non-migratable).

There are four functional components to support the bootstrapping process:

Domain Authentication Server (DAS) provides a trusted root for the third party's domain.

Bootstrap Initiator (BI) is the application that is transferred to the remote machine in order to confirm the machine's status before any infrastructure or software is deployed.

Bootstrap Requester (BREQ) is a client application that runs on the machine responsible for provisioning remote infrastructure. It communicates with its counterpart on the remote machine and handles the first/initial stage of the bootstrapping process.

Bootstrap Responder (BRES) is the counterpart server application. It is responsible for authenticating the machine to a remote client and verifying that the client is authorized to bootstrap the machine. Once each end point has been authenticated, the BRES will receive, decrypt and decompress the payload sent by the client.

The bootstrapping process includes the following 4 steps:

1) Initially the BRES on the target machine, registers and authenticates itself with the DAS. This done over a TCP connection. Hardware based keys from the TPM are used to authenticate the instance and complete the handshake. Key data is then signed and stored on the DAS.

2) When the BREQ needs to authenticate a target machine, it connects to the DAS and authenticates itself. This authentication could be simple user and password based authentication, or could also involve security tokens or pre-shared certificates and keys.

3) After authentication, the DAS provides the BREQ with the certificates and keys for the target machine. The BREQ then constructs a bootstrapping request with an encrypted payload containing the Bootstrap Initiator (BI), secured using the provided credentials. This requests is then sent to the DAS over the same authenticated TCP

channel. The DAS then signs and forwards the request with the encrypted payload to the BRES.

4) As the payload is encrypted with the target machines public key / certificate which is tied to the TPM (non-migratable keypair), only the target machine is able to decrypt the payload. Once decrypted, the BRES executes the BI and hands control over to it. The BI can effectively execute any code on the machine and thus can verify that the machine and the platform are as expected and as required. Once complete, the BI can then download the infrastructure payload (this would be implementation specific) and can then execute it and hand over control to the framework.

A prototype implementation of the BREQ and BRES is called Yin and Yang and described in [48]. The NodeJS and SocketIO libraries, provide a two-way message framework that allows the focus to remain on the message content and their structure. NodeJS has bindings for NaCl which provide a range of cryptographic functions. At present there is no native binding for TPM functionality, however initially software generated keys and certificates can be exchanged for developing and verifying the protocol.

7 Future Research and Development

The authors will continue their research to understand the new challenges and required solutions for Big Data infrastructure and applications. The future research and development will include further enhancement of the Big Data definition. This should provide a better basis for proposing a consistent Big Data security model and architecture addressing identified security challenges presented in this paper. At this stage we tried to review existing security technologies, own experience and consolidate them around the main security problems in Big Data such as providing trusted virtualized environment for data processing and storing, fine granular access control, and general infrastructure security for scientific and general Big Data applications.

The authors will also continue working on the data centric and user centric security models that should also incorporate new Big Data properties such as data ownership. A number of technical security problems will arise with the implementation of persistent data and researcher identifiers (PID and ORCID), as required by the new EC initiative, and related privacy and provenance issues.

As a part of the general infrastructure research we will continue research on the infrastructure issues in Big Data targeting more detailed and technology oriented definition of SDI and related security infrastructure definition. Special attention will be given to defining the whole cycle of the provisioning SDI services on-demand, specifically tailored to support instant scientific workflows using cloud IaaS and PaaS platforms. This research will be also supported by development of the corresponding Cloud and InterCloud architecture framework to support the Big Data e-Science processes and infrastructure operation.

The authors will look also at the possibility to contribute to the standardisation activity at the Research Data Alliance (RDA) [51] and recently established NIST Big Data Working Group [4].

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. 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
3. Roundup of Big Data Pundits' Predictions for 2013. Blog post by David Pittman. January 18, 2013. [online] http://www.ibmbigdatahub.com/blog/roundup-big-data-pundits-predictions-2013
4. NIST Big Data Working Group (NBD-WG) [online] http://bigdatawg.nist.gov/home.php/
5. Demchenko, Y., Z.Zhao, P.Grosso, A.Wibisono, C. de Laat, Addressing Big Data Challenges for Scientific Data Infrastructure. The 4th IEEE Conf. on Cloud Computing Technologies and Science (CloudCom2012), 3 - 6 December 2012, Taipei, Taiwan.
6. 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). Proc. The 2013 International Conference on Collaboration Technologies and Systems (CTS 2013), May 20-24, 2013, San Diego, California, USA.
7. 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
8. 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
9. The 3Vs that define Big Data. Posted by Diya Soubra on July 5, 2012 [online] http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data
10. IDG IDC's Latest Digital Data Study: A Deep Dive, Blogpost by Mary Ludloff. [online] http://blog.patternbuilders.com/2011/07/08/idcs-latest-digital-data-study-deep-dive/
11. The Big Data Long Tail. Blog post by Jason Bloomberg on Jan 17, 2013. [online] http://www.devx.com/blog/the-big-data-long-tail.html
12. 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/
13. NIST Big Data Workshop, 13-14 June 2012. [online] http://www.nist.gov/itl/ssd/is/big-data.cfm
14. CSA Big Data Working Group. [online] https://cloudsecurityalliance.org/research/big-data/
15. Expanded Top Ten Big Data Security and Privacy Challenges. CSA Report, 16 June 2013. [online] https://downloads.cloudsecurityalliance.org/initiatives/bdwg/ Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf
16. Peisert, S., E.Talbot, M.Bishop, Turtles all the way down: a clean-slate, ground-up, first-principles approach to secure systems, Proceeding NSPW '12 Proceedings of the 2012 workshop on New security paradigms, ACM New York, NY, USA 2012.
17. Bratus, S., M.Locasto, A.Ramaswamy, S.Smith, VM-based Security Overkill: A Lament for Applied Systems Security Research, Proceeding NSPW '10 Proceedings of the 2010 workshop on New security paradigms, ACM New York, NY, USA 2010.

18. Morton, A., A.Sasse, Privacy is a Process, not a PET: A Theory for Effective Privacy Practice, Proceeding NSPW '12 Proceedings of the 2012 workshop on New security paradigms, ACM New York, NY, USA 2012. ISBN: 978-1-4503-1794-8
19. Deng, M., M.Nalin, M.Petkovich, I.Baroni, A.Marco, Towards Trustworthy Health Platform Cloud. 9th VLDB Workshop, SDM 2012, Istanbul, Turkey, August 27, 2012. Proceedings.
20. Bienvenu, M., D.Deutch, F.Suchanek, Provenance for Web 2.0 Data, 9th VLDB Workshop, SDM 2012, Istanbul, Turkey, August 27, 2012. Proceedings.
21. Demchenko, Y., C. de Laat, O. Koeroo, D. Groep, Re-thinking Grid Security Architecture. Proceedings of IEEE Fourth eScience 2008 Conference, December 7–12, 2008, Indianapolis, USA. Pp. 79-86. IEEE Computer Society Publishing. ISBN 978-0-7695-3535-7
22. 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
23. Oracle Fusion Middleware Security Guide: Overview Java Security Models [online] http://docs.oracle.com/cd/E12839_01/core.1111/e10043/introjps.htm
24. Hypervisors, virtualization, and the cloud: Learn about hypervisors, system virtualization, and how it works in a cloud environment. By Bhanu P. Tholeti, IBM. [online] http://www.ibm.com/developerworks/cloud/library/cl-hypervisorcompare/
25. Prins, C., When personal data, behavior and virtual identities become a commodity: Would a property rights approach matter? SCRIPT-ed - A Journal of Law, Technology & Society, Volume 3, Issue 4, June 2006. [online] http://www2.law.ed.ac.uk/ahrc/script-ed/vol3-4/prins.pdf
26. RFC6749: The OAuth 2.0 Authorization Framework. http://tools.ietf.org/html/rfc6749
27. 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
28. Data Lifecycle Models and Concepts. [online] http://wgiss.ceos.org/dsig/whitepapers/Data%20Lifecycle%20Models%20and%20Concepts%20v8.docx
29. Koopa, D., 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
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. Demchenko, Y., D.R. Lopez, J.A. Garcia Espin, C. de Laat, "Security Services Lifecycle Management in On-Demand Infrastructure Services Provisioning", International Workshop on Cloud Privacy, Security, Risk and Trust (CPSRT 2010), 2nd IEEE International Conference on Cloud Computing Technology and Science (CloudCom2010), 30 November - 3 December 2010, Indianapolis, USA.
34. 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.

35. 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.
36. eduGAIN - Federated access to network services and applications. [online] http://www.edugain.org
37. 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
38. Ngo; C., Demchenko, Y.; de Laat, C., "Toward a Dynamic Trust Establishment approach for multi-provider Intercloud environment," Cloud Computing Technology and Science (CloudCom), Proc. 2012 IEEE 4th International Conference on , vol., no., pp.532,538, 3-6 Dec. 2012
39. Demchenko Y., L. Gommans, C. de Laat. "Using SAML and XACML for Complex Resource Provisioning in Grid based Applications". In Proc. IEEE Workshop on Policies for Distributed Systems and Networks (POLICY 2007), Bologna, Italy, 13-15 June 2007.
40. Demchenko, Y., C. M. Cristea, de Laat, XACML Policy profile for multidomain Network Resource Provisioning and supporting Authorisation Infrastructure, IEEE International Symposium on Policies for Distributed Systems and Networks (POLICY 2009), July 20-22, 2009, London, UK.
41. 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).
42. MongoDB [online] http://www.mongodb.org/
43. Apache Cassandra [online] http://cassandra.apache.org/
44. Apache Accumulo [online] http://accumulo.apache.org/
45. 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
46. 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
47. Demchenko Y., L.Gommans, C. 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.
48. 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
49. 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.
50. 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
51. Research Data Alliance (RDA). [online] http://rd-alliance.org/