Cloud Based Big Data Infrastructure: Architectural Components and Automated Provisioning

Yuri Demchenko, Fatih Turkmen, Cees de Laat
System and Network Engineering Group
University of Amsterdam
Amsterdam, The Netherlands
{y.demchenko, f.turkmen, C.T.A.M.deLaat}@uva.nl

Christophe Blanchet
CNRS IFB
Orsay, France
christophe.blanchet@france-bioinformatique.fr

Charles Loomis
SixSq Sàrl
Geneva, Switzerland
cal@sixsq.com

***Abstract*—This paper describes the general architecture and functional components of the cloud based Big Data Infrastructure (BDI). The proposed BDI architecture is based on the analysis of the emerging Big Data and data intensive technologies and supported by the definition of the Big Data Architecture Framework (BDAF) that defines the following components of the Big Data technologies: Big Data definition, Data Management including data lifecycle and data structures, Big Data Infrastructure (generically cloud based), Data Analytics technologies and platforms, and Big Data security, compliance and privacy. The paper provides example of requirements analysis and implementation of two bioinformatics use cases on cloud and using SlipStream based cloud applications deployment and management automation platform being developed in the CYCLONE project. The paper also refers to importance of standardisation of all components of BDAF and BDI and provides short overview of the NIST Big Data Interoperability Framework (BDIF). The paper discusses importance of automation of all stages of the Big Data applications developments, deployment and management and refers to existing cloud automation tools and new developments in the SlipStream cloud automation platform that allows multi-cloud applications deployment and management.**

Keywords-Big Data Architecture Framework, Big Data Infrastructure, Multi-cloud Services Provisioning Automation, SlipStream Cloud Automation platform, Intercloud Architecture Framework (ICAF)

I. INTRODUCTION

Big Data technologies (also called Data Science, Data Intensive, Data Centric, Data Driven or Data Analytics) are becoming a current focus and a general trend both in science and in industry. Modern e-Science, empowered with advance computing, brings new possibilities for industry to benefit from advanced data processing methods; industry in its own turn offers to scientific/research community advanced computing, data storage and communication platforms

Modern research is becoming more and more data intensive and requires using high-performance computing and large volume storage. Cloud Computing [1, 2] and Big Data technologies [3, 4] provide necessary computing and data processing platform for data intensive and data driven applications in research and industry.

Fusion between Big Data and cloud technologies fuels modern data driven research [5] and provides a basis for modern e-Science that benefits from wide availability of affordable computing and storage resources provisioned on demand. Modern e-Science infrastructures allow targeting new large scale problems which solution was not possible before, *e.g.* research on genome, climate, global warming. Modern e-Science produces a huge amount of data that need to be supported by a new type of e-Infrastructure capable to store, distribute, process, preserve, and curate these data.

This paper presents ongoing research to define the Scientific and Big Data Infrastructure (SDI/BDI) that responds to the demand from research community and industry and incorporates benefits and recent developments of the Cloud Computing and agile applications development technologies. It is based on and extends the main concepts proposed in the earlier authors' papers. The paper explains importance of standardisation in the area of Big Data technologies and briefly refers to the recently published the NIST Big Data Interoperability Framework standards.

The remainder or paper is organized as follows. Section II shortly describes Big Data use in science and industry what provides a basis for discussion on changing Big Data paradigms in Section III. Section IV provides definition of the Big Data Architecture Framework, and Section V provides details and summarises benefits of implementing BDI on clouds. Section VI discusses bioinformatics use cases that illustrate the practical needs and requirements to cloud based BDI. Section VII discusses requirements and existing cloud automation solutions. Section VIII describes functionality of the SlipStream cloud applications automation platform and provides example of bioinformatics use cases implementation.

II. BIG DATA IN SCIENCE AND INDUSTRY

Science has been traditionally dealing with challenges to handle large volumes of data in complex scientific research experiments, involving wide cooperation among distributed groups of individual scientists and research organizations. Scientific research typically includes a 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. New data driven science allows discovery of hidden relations based on processing of large amount of data what was not possible with the past technologies and scientific platforms. 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.

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 technological data, which in general need to be stored or retained to allow effective quality control or diagnostics in case of failure or crash.

With the digital technologies proliferation into all aspects of business activities and emerging Big Data technologies, the industry is entering a new playground. It needs to use scientific methods to benefit from the possibility to collect and mine data for desirable information, such as market prediction, customer behavior predictions, social groups activity predictions, etc.

III. BIG DATA PARADIGM CHANGE AND DATA CENTRIC SERVICES MODEL

A. *New features of the Big Data Infrastructure*

Big Data is not just a large storage, database or Hadoop (as a platform for scalable Big Data processing) problem, although they constitute the core technologies and components for large scale data processing and data analytics. It is the whole complex of components to store, process, visualize and deliver results to target applications. Big data is deemed “the fuel” of all processes, source, target, which together create multiple data value chains.

We will refer to such environment as the Big Data Ecosystem (BDE) that deals with the evolving data, models and requires highly adaptable infrastructure supporting the whole Big Data lifecycle.

Big Data as technology and Data Science as professional area and scientific discipline are becoming a new technology driver and require re-thinking a number of architecture models, infrastructure components, solutions and processes to address exponential growth of data volume produced by different research instruments and/or collected from sensors.

The recent advancements in the general ICT and Cloud Computing technologies facilitate the paradigm change in Big Data Science that is characterized by the following features:

- Automation of all Data Science processes including data collection, storing, classification, indexing and other components of the general data curation and provenance.
- Transformation of all processes, events and products into digital form by means of multi-dimensional multi-faceted measurements, monitoring and control; digitising existing artifacts and other content.
- Possibility to re-use the initial and published research data with possible data re-purposing for secondary research
- Global data availability and access over the network for cooperative group of researchers, including wide public access to scientific data.
- Existence of necessary infrastructure components and management tools that 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 infrastructures and scientific instruments and allow creating trusted secure environment for cooperating groups and individual researchers.

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

Traditional/current IT and communication technologies are client/server based and host/service centric what means that all communication or processing are bound to host/computer that runs application software. This is especially related to security services. The administrative and security domains are the key concepts, around which the services and protocols are built. A domain provides a context for establishing security context and trust relation. This creates a number of problems when data (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 situations 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 is needed to address the following issues:

- Maintaining semantic and referral integrity, i.e. linkage between data at the different stages of their transformation
- Data location, search, access
- Data integrity and identifiability, referral integrity

- Data security and data centric access control
- Data ownership, personally identifiable data, privacy, opacity of data operations
- Trusted virtualisation platform, data centric trust bootstrapping

IV. Big Data Architecture Framework and Components

A. *Defining Big Data Architecture Framework (BDAF)*

In this section, we will discuss the Big Data Architecture Framework (BDAF) that intends to support the extended Big Data definition proposed by authors in their early work [6] and support the main components and processes in the Big Data Ecosystem (BDE). The proposed here BDAF definition follows the ongoing industry standardisation on Big Data at National Institute of Standards and Technology of USA (NIST) (NBD-WG, 2015) [7] and reflects advanced research in scientific community. The presented BDAF is based on and compatible with the standardization and industry best practices in Big Data, Cloud Computing and Information Systems management, in particular, NIST Cloud Computing Reference Architecture [2], NIST Big Data Reference Architecture[3], Intercloud Architecture Framework (ICAF) [8].

The proposed BDAF comprises of the following 5 components that address different Big Data Ecosystem and Big Data definition aspects:

(1) Data Models, Structures, Types
- Data formats, non/relational, file systems, etc.

(2) Big Data Management
- Big Data Lifecycle Management
- Big Data models, semantics and transformation
- Storage, Curation, Provenance, Archiving

(3) Big Data Analytics and Tools
- Big Data Analytics Methods and Applications
- Target use, presentation, visualisation

(4) Big Data Infrastructure (BDI)
- Storage, Compute, High Performance Computing, Network provisioned on demand
- Sensor network, target/actionable devices
- BDI provisioning, operation and automation

(5) Big Data Security and Privacy
- Data security in-rest, in-move, trusted processing environments
- Big Data systems compliance and dependability
- Digital right protections
- Privacy and personal information protection

B. *Data Management and Big Data Lifecycle*

First and the most important task of the data management in the Big Data systems is to ensure data storage and access. This functionality is typically supported by data discovery or search services. We have quite well developed web search of hypertext information. For Big Data search we will need to develop new methods for searching and discovery in large volumes of interlinked data that may be largely distributed and belong to different management domains.

Metadata management is one of the key functions in the data management system and enabler for all other data management functions. Metadata, i.e. data describing data, actually define the semantic meaning of data and relation between data. Metadata reflect data model and data classification. Examples of metadata, related to data are Persistent Data Identifier (PID), filename, URL commonly used for web resources identification and access, record time, author, variable, etc.

The required new approach to data management and processing in Big Data industry is reflected in the Big Data Lifecycle Management (BDLM) model which is shown in Figure 1 that illustrates the main stages of Big Data Lifecycle and their relation to other components of the BDAF [4]. The following data transformation stages are typically present in Big Data applications:

- Data collection and registration
- Data filtering and classification
- Data analysis, modelling, prediction
- Data delivery and visualization.

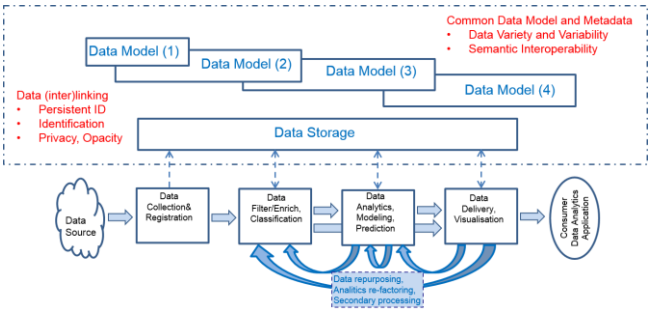

Figure 1. Big Data Lifecycle in Big Data Ecosystem.

The figure also reflect the iterative character of the typical Big Data processing sequence or workflow. The iterative process allows improving data quality and analytics methods by improving data preparation or selection, applying improved or different data analytics methods. Data can be also re-used or used for another data analysis purposes, i.e. re-purposed. It is important to mention that data models may change during data processing and lifecycle, however data management services should provide consistent linking and provenance of all data transformation processes, data sets and models.

The data lifecycle model definition is an important research topic in scientific data management where different scientific domain and data types may bring their own specifics as described in extensive research by CEOS (Committee on Earth Observation Satellites) [9]. The data preservation and curation community defined their own data lifecycle management model that reflects the scientific research cycle that includes research data handling and scientific results publications [10].

Commonly recognized requirement and business rationality to store and re-use data at different stages need to be supported by the consistent metadata, data identification and linkage functionality and infrastructure. When advancing through the data lifecycle stages, data may change their model/structure and consequently metadata; this may be also a consequence of using different systems or platforms at different data processing

stages. Linking data during the whole data lifecycle is one of the important problems in Big Data lifecycle.

C. NIST Big Data Reference Architecture

The NIST Big Data Working Group (NBD-WG) targeted interoperability and scalability of the Big Data infrastructures when developing the Big Data Reference Architecture (BDRA). As summary of this work the NIST Special Publication NIST SP 1500: NIST Big Data Interoperability Framework (NBDIF)[1] [3] has been published in September 2015. The NBDIF publication includes 7 volumes:

Volume 1: NIST Big Data Definitions
Volume 2: NIST Big Data Taxonomies
Volume 3: NIST Big Data Use Case & Requirements
Volume 4: NIST Big Data Security and Privacy Requirements
Volume 5: NIST Big Data Architectures White Paper Survey
Volume 6: NIST Big Data Reference Architecture
Volume 7: NIST Big Data Technology Roadmap

Important to mention that the NBDIF defines 3 main components of the Big Data technology what identifies the main challenges in successful adoption of the new technology for research and industry:

- Big Data Paradigm that includes such concepts as domain related data handling ecosystem, data driven research and production processes or workflows, and required data centric infrastructure and applications design
- Big Data Science and Data Scientist as a new profession
- Big Data Architecture (refer to Volume 6 for details)

It is important to mention that in NIST BDRA the Big Data Application Provider is separated from the Big Data Platform Provider. This reflects new features brought to the modern application infrastructure by cloud computing that, by means of virtualization, allow deploying application on-demand on the general purpose cloud based platforms and underlying infrastructure and ensure data processing scalability and mobility over distributed multi-cloud infrastructure.

V. BIG DATA INFRASTRUCTURE (BDI)

A. Big Data Infrastructure components

Big Data applications are typically distributed and use multiple distributed data sources. Data processing and staging also involve using distributed computing and storage resources. Cloud Computing presents a right choice as a general purpose Big Data platform. Cloud technologies bring the benefit of building scalable infrastructure services that can be provisioned on-demand and dynamically scaled depending on the required workload and volume of data. The major enabling technology for cloud computing is virtualisation of all components of the general computing infrastructure: servers, storage, and network. Virtualisation means physical resources pooling, abstraction, composition, and orchestration under the supervision of a hypervisor (a special software for physical resources virtualisation) and cloud management software which is the main component of any cloud platform.

Figure 2 provides a general view of the major components of the Big Data infrastructure that includes the infrastructure for general data management, typically cloud based, and Big Data Analytics part that support the main stages of the data transformation [4]. The general infrastructure components and services for general data management and user access include:

- General purpose data storage and processing infrastructure that is typically cloud based and provisioned on-demand;
- Big Data Management tools: registries, indexing, data search/discovery, metadata and semantics;
- Security infrastructure (access control, policy enforcement, confidentiality, trust, availability, privacy);
- Federated Access and Delivery Infrastructure (FADI) that allows interconnection and interaction of all BDI components; it is typically provisioned on-demand;
- Collaborative environment to support user groups creation and management.

The Big Data analytics infrastructure components are required to support massive data processing required by Big Data applications and other data centric applications:

- HPC clusters
- Hadoop based applications, streaming data analytics and other tools for large scale data processing
- Specialist data analytics tools (complex events processing, data mining, etc.)
- NoSQL databases for Big Data storage and processing
- Distribute file systems for large scale data storage and processing

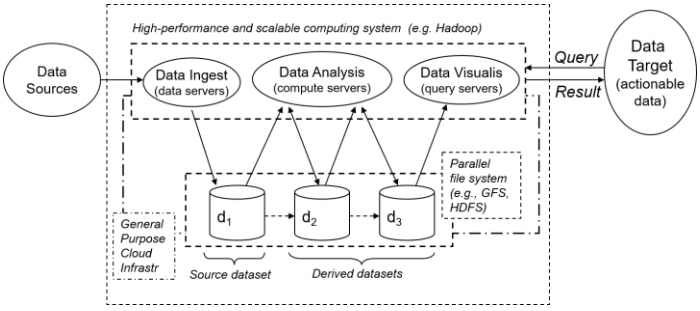

Figure 2. Data processing stages mapped to Big Data infrastructure components: cloud compute and storage resources.

B. Big Data Stack components and technologies

The major structural components of the Big Data stack are grouped around the main stages of data transformation:

(1) Data ingest: Ingestion will transform, normalize, distribute and integrate to one or more of the Analytic or Decision Support engines; ingest can be done via ingest API or connecting existing queues that can be effectively used for

[1] The authors contributed to the NIST SP 1500 development, in particular, to the definition of the Big Data paradigm and Big Data Architecture

handles partitioning, replication, prioritisation and ordering of data

(2) Data processing: Use one or more analytics or decision support engines to accomplish specific task related to data processing workflow; using batch data processing, streaming analytics, or real-time decision support

(3) Data Export: Export will transform, normalize, distribute and integrate output data to one or more Data Warehouse or Storage platforms;

(4) Back-end data management, reporting, visualization: will support data storage and historical analysis; OLAP platforms/engines will support data acquisition and further use for Business Intelligence and historical analysis.

Figure 3 provides overview of the Big Data analytics platforms and tools corresponding to Big Data stack groups. Majority of software tools are Open Source Software, however a number of tools and platforms are proprietary and specific for large public Cloud Computing platforms such as Microsoft Azure, Amazon Web Services (AWS), Google Compute Engine cloud (GCE), HortonWorks, Vertica. More detailed overview of the most popular Big Data cloud platforms is provided below.

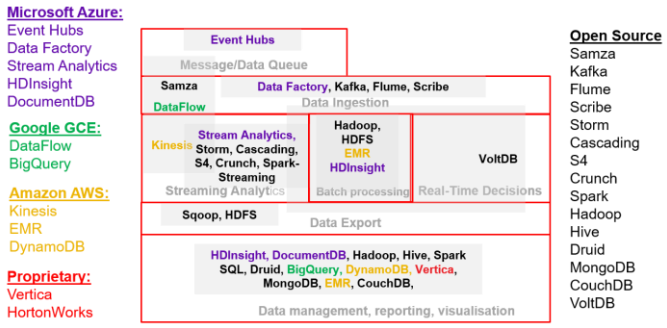

Figure 3. Big Data analytics platforms and tools corresponding to Big Data stack groups.

C. *Benefits of cloud platforms for Big Data applications*

Building Big Data application on cloud brings the following benefits to whole process of applications development, deployment, management and operation:

- Cloud deployment on virtual machines or containers
  - Applications portability and platform independence, on-demand provisioning
  - Dynamic resource allocation, load balancing and elasticity for tasks and processes with variable load
- Availability of rich cloud based monitoring tools for collecting performance information and applications optimisation
- Network traffic segregated and isolation
  - Big Data applications benefit from lowest latencies possible for node to node synchronization, dynamic cluster resizing, load balancing, and other scale-out operations
  - Clouds construction provides separate networks for data traffic and management traffic
  - Traffic segmentation by creating Layer 2 and Layer 3 virtual networks inside user/application assigned Virtual Private Cloud (VPC)
- Cloud tools for large scale applications deployment and automation
  - Provide basis for agile services development and Zero-touch services provisioning
  - Applications deployment in cloud is supported by major Integrated Development Environment (IDE)

VI. Case Study: Bioinformatics Applications Deployment on Cloud

A. *Overall description*

The described here two basic bioinformatics use cases have being implemented in the CYCLONE project [11]. They provide practical example of building cloud based infrastructure for one of the most demanding data intensive scientific domain. The identified requirements include general requirements and specific for bioinformatics applications, in particular strong demand for automated cloud infrastructure provisioning and management that may span multiple institutions and multiple clouds.

Bioinformatics applications process genomic information from DNA sequencers which produce terabytes of information for continuously dropping price (currently less than US$1000 for a human genome) creating a "data deluge" that is being experienced by researchers in this field [12, 13].

Bioinformatics software is characterized by a high degree of fragmentation: literally hundreds of software packages are regularly used for scientific analyses with an incompatible variety of dependencies and a broad range of resource requirements. For this reason, the bioinformatics community has strongly embraced cloud computing with its ability to provide customized execution environments and dynamic resource allocation.

The French Institute of Bioinformatics (IFB) [14] has deployed a cloud infrastructure on its own premises at IFB-core, and aims to deploy a federated cloud infrastructure over the regional PFs devoted to the French life science community, research and industry, with services for the management and analysis of life science data.

B. *Processing human biomedical data (UC1)*

Continuous decrease of the genome sequencing costs (also referred to as NGS – Next Generation DNA Sequencing) allows including genome analysis into day-to-day clinical diagnostic practice. Today, most of genomics analyses are realized on the exome, which is the expressed part (5%) of the genome. However, the full genome sequencing is being envisaged and will be soon included in daily medical practices.

Practical use of genomic applications for processing human biomedical data related to patients will require strict personal data protection and consequently strong and consistent security infrastructure implemented at IFB that should comply with the HIPAA/HITECH regulations [15].

Figure 4 presents the workflow that implemented federated access control: (1) a biomedical user connects to the cloud through the IFB web authenticated dashboard; uses it to (2) run an instance of the appliance containing the relevant pre-configured analysis pipeline. At step (3) the VM containing genome applications is deployed on the cloud (testbed); then (4) the user signs into the web interface of the VM, (5) uploads the patient's biomedical data, and (6) runs the analysis in a secure environment. Finally, (7) the user gets the results.

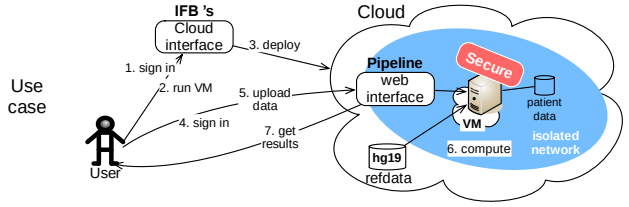

Figure 4: Interaction between functional components in processing human biomedical data.

C. *Cloud pipeline for microbial genomes analysis (UC2)*

In the post-NGS research, researchers are will intend to compare large collections of related bacterial genomes (strains). Therefore, this brings increased requirements for automating the annotation of bacterial genomes.

The IFB-MIGALE bioinformatics platform allows annotation of microbial genomes and visualization of the synteny (local conservation of the gene order along the genomes) [16]. The platform automatically launches a set of bioinformatics tools (e.g. BLAST, INTERPROScan) to analyse the data and stores the results in a relational database (e.g. PostgreSQL). These tools use several public reference data collections. A web interface allows the user to consult the results and perform the manual annotation (manual annotation means adding manually metadata and biological knowledge to the genome sequence). Installing the platform requires advanced skills in system administration and application management. Performing the analysis of collections of genomes requires large computing resources that can be distributed over several computing clusters and datacenters.

The cloud infrastructure should allow the life science researchers to deploy their own comprehensive annotation platforms over one or more clouds with the dynamic allocation of infrastructure resources inside a dedicated Virtual Private Cloud (VPC) including distributed user data.

Figure 5 illustrate the use case sequence: a bioinformatician (1) connects to the cloud web dashboard, uses it to (2) run and (3) deploy with one click a genomes annotation platform consisting of many VMs, comprising of a master node of the virtual cluster that provides also the visualization web-interface, associated with several computing nodes. Then the user (4) uses secure communication over SSH to connect to the master and (5) uploads the raw microbial genomic data (MB) to the cloud storage. SCP/SFTP protocols are used from a command line tool or a GUI, to ensure AuthN/Z for the data transfer, and to overcome the performance issues of HTTP for large datasets. Still in command line interface, the user (6) runs the computation to annotate the new microbial genomes. The first step consists of many data-intensive jobs performing the comparisons between the new genome and the reference data.

The results are stored in a relational database (provided by a cloud service or a VM deployed within the platform). Then the scientist (7) signs in the annotated data visualization environment provided by the Insyght web-interface to (8) navigate between the abundant homologues, syntenies and gene functional annotations in bacteria genomes.

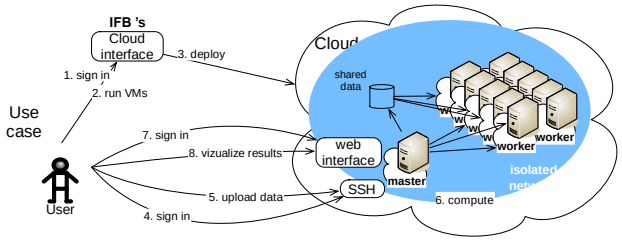

Figure 5: Interaction between functional components in the use case "Cloud virtual pipeline for microbial genomes analysis".

D. *Implementation of use cases and CYCLONE infrastructure components*

1) Deployment UC1 Securing human biomedical data

The first use case "Securing human biomedical data" requires enhanced security features and a deployment done only on certified (by the French Health Ministry) cloud infrastructure. The cloud appliance NGS-Unicancer has been developed as a part of the project NGS-Clinique (INCA - Institut National du Cancer) and is enhanced with the federated access control being developed in the CYCLONE project. The user deploys the NGS-Unicancer appliance via the IFB web interface using the CYCLONE federation provider, gets access to the provisioned VM web interface, uploads their data, run the analysis and obtains the results. In Figure 6, the upper part describes the use case workflow, middle layer represents the workflow steps that are linked to the related CYCLONE software components and services. The bottom part shows the testbed infrastructure components.

2) Deployment UC2: Cloud virtual pipeline for microbial genomes analysis

The second bioinformatics use case requires several components: a user web interface, a relational PostgreSQL database, and a complete computing cluster with a master and several nodes to perform the data-intensive analyses. The application infrastructure requires multiple components provisioning and is implemented using SlipStream cloud automation system, the deployment process is described in a form of SlipStream recipe.

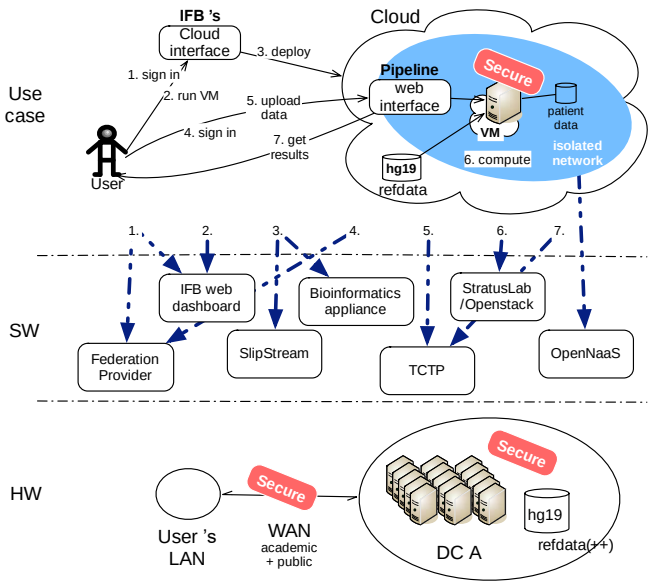

Figure 6: Functional components of the CYCLONE framework involved into UC1 application infrastructure provisioning.

E. Ensuring consistent security of services provisioning on clouds

Due to the generically distributed Bioinformatics resources and applications they may need to be deployed on multiple clouds what brings big challenges in building consistent security infrastructure and services that may span few cloud providers and needs to be integrated with the institution Identity Management and access control system.

Dynamic Access Control Infrastructure (DACI) [17] is provisioned on demand together with the main application and requires security bootstrapping for establishing trust relations between enterprise and cloud security domains [18].

VII. Platforms for Automation of Big Data Applications Provisioning and Management

Modern Big Data applications and infrastructures may include multiple existing cloud platforms and the problem of applications integration and management is becoming critical and requires automation of most development and operation functions, ideally supporting modern development and operation (DevOps) tools [19]. Currently widely used cloud automation tools such as Chef [20], Puppet [21], Ansible [22] allow single cloud provider application deployment but they don't solve problem of multi-cloud services integration and provisioning of inter-cloud network infrastructure.

Intercloud platforms should deliver open integration environment and standardized APIs, protocols, and data formats, allowing for cross-cloud resources interoperability. Practical intercloud platform development should target two major stakeholders and user communities: Application Service Providers (ASPs) as well as their customers to address real life challenges and problems in a consistent and constructive way. The required functionality is provided by the CYCLONE cloud applications deployment platform [11] that leverages the SlipStream cloud automation and management platform [23].

Described above bioinformatics use cases illustrate demand for complex applications deployment and management tools so that they can focus on their main research work/tasks and use complex applications and infrastructure at a "fingertip". This works as a motivator for development of a new concept of applications provisioning, operation and management in clouds defined as Zero Touch Provisioning, Operations and Management (ZTPOM) proposed by the authors in the recent paper [24]. This trend is also recognised by TMForum that have launched the ZOOM (Zero-touch Orchestration, Operations and Management) program to develop best practices and standards for a new generation of service provider support systems that will deliver high business agility and rapid new service development [25].

The proposed ZTPOM framework is being developed in the GEANT project to enable the general (Inter-)Cloud Services Delivery Infrastructure (CSDI) that leverages the advance functionalities of the GEANT network infrastructure [26] and intends to solve the "last mile" problem in delivering cloud services to research campuses and end-users. The ZTPOM enabled CSDI is implemented using architectural and design patterns of the Intercloud Federation Framework (ICFF) [27] and core network infrastructure component the GEANT Open Cloud eXchange (OCX) [28, 29] that serves as customer and cloud provider front-end to access the GEANT network infrastructure. The intended ZTPOM platform development based on the SlipStream will provide smooth integration with other Slipstream applications requiring advanced networking connectivity.

While mentioned above cloud automation tools use recipes or cookbooks that describe the VM and cloud resources configuration in a declarative language, the complete application topology and components interrelationship can be described using a language like OASIS TOSCA (Topology and Orchestration Specification for Cloud Applications) [30], wherein the applications topology and workflow that invokes different cloud based services is described.

VIII. SlipStream: Cloud Application Management Platform

Within CYCLONE, software developers and service operators manage the complete lifecycle of their cloud applications with SlipStream, an open source cloud application management platform. [2] Through its plugin architecture, SlipStream supports most major cloud service providers and the primary open source cloud distributions. By exposing a uniform interface that hides differences between cloud providers, SlipStream facilitates application portability across the supported cloud infrastructures.

[2] Community Edition of SlipStream, is available under the Apache 2.0 license (https://github.com/slipstream)

To take advantage of cloud portability, developers define "recipes" that transform available "base" virtual machines into the components that they need for their application. By reusing these base virtual machines, developers can ensure uniform behaviour of their application components across clouds without having to deal with the time-consuming and error-prone transformation of virtual machine images. Developers bundle the defined components into complete cloud applications using SlipStream facilities for passing information between components and for coordinating the configuration of services.

Once a cloud application has been defined, the operator can deploy the application in "one click", providing values for any defined parameters and choosing the cloud infrastructure to use. With SlipStream, operators may choose to deploy the components of an application in multiple clouds, for example, to provide geographic redundancy or to minimize latencies for clients. To respond to changes in load, operators may adjust the resources allocated to a running application by scaling the application horizontally (changing the number of virtual machines) or vertically (changing the resources of a virtual machine).

SlipStream combines its deployment engine with an "App Store" for sharing application definitions with other users and a "Service Catalog" for finding appropriate cloud service offers, providing a complete engineering PaaS supporting DevOps processes. All of the features are available through its web interface or RESTful API.

A. *Functionality used for use cases deployment*

The bioinformatics use cases described above principally used SlipStream's facilities and tools to define applications and its deployment engine through the RESTful API.

The definition of an application component actually consists of a series of recipes that are executed at various stages in the lifecycle of the application. The main recipes, in order, are:

- **Pre-install**: Used principally to configure and initialize the operating system's package management.
- **Install packages**: A list of packages to be installed on the machine. SlipStream supports the package managers for the RedHat and Debian families of OS.
- **Post-install**: Can be used for any software installation that can not be handled through the package manager.
- **Deployment**: Used for service configuration and initialization. This script can take advantage of SlipStream's "parameter database" to pass information between components and to synchronize the configuration of the components.
- **Reporting**: Collects files (typically log files) that should be collected at the end of the deployment and made available through SlipStream.

There are also a number of recipes that can be defined to support horizontal and vertical scaling that are not used in the defined here use cases. The applications are defined using SlipStream's web interface, the bioinformatics portal then triggers the deployment of these applications using the SlipStream RESTful API.

B. *Example recipes*

The application for the bacterial genomics analysis consisted of a compute cluster based on Sun Grid Engine with an NFS file system exported from the master node of the cluster to all of the slave nodes. The master node definition was combined into a single "deployment" script that performed the following actions:

(1) Initialize the yum package manager.
(2) Install bind utilities.
(3) Allow SSH access to the master from the slaves.
(4) Collect IP addresses for batch system.
(5) Configure batch system admin user.
(6) Export NFS file systems to slaves.
(7) Configure batch system.
(8) Indicate that cluster is ready for use.

The deployment script extensively uses the parameter database that SlipStream maintains for each application to correctly the configure the master and slaves within the cluster. A common pattern is the following:

```
ss-display "Exporting SGE_ROOT_DIR..."
echo -ne "$SGE_ROOT_DIR\t" > $EXPORTS_FILE
for ((i=1; i<=`ss-get
         Bacterial_Genomics_Slave:multiplicity`;  i++
));
do
     node_host=`ss-get
        Bacterial_Genomics_Slave.$i:hostname`
     echo -ne $node_host >> $EXPORTS_FILE
     echo    -ne    "(rw,sync,no_root_squash)    ">>
$EXPORTS_FILE
done
echo "\n" >> $EXPORTS_FILE # last for a newline
exportfs -av
```

IX. CONCLUSION

The paper provides information on the ongoing research and development in the EU funded project CYCLONE that is focused on the creation of the multi-cloud applications deployment and management automation platform that is built on top of the SlipStream cloud automation and management platform. The presented and analysed bioinformatics use cases demonstrate the CYCLONE's platform ability to automate cloud deployment of the typical Big Data applications.

Further development of the CYCLONE and SlipStream platforms will target other Big Data and data intensive applications that involve distributed data sources and cloud resources to be dynamically provisioned. It is also intended to contribute to the definition of a set of the data centric API for BDAF and NIST BDIF that would allow flexible data lifecycle and applications workflow management.

ACKNOWLEDGEMENT

The research leading to these results has received funding from the Horizon2020 project CYCLONE (funded by the European Commission under grant number 644925) and the French programs PIA INBS 2012 (CNRS IFB). The authors are thankful to the partners of the CYCLONE project and to the colleagues who collaborated on deploying selected

bioinformatics applications in the CYCLONE testbed: Christian Baudet (Centre Léon Bérard, Lyon, France) and Thomas Lacroix (IFB-MIGALE, Jouy-en-Josas, France).

REFERENCES

[1] NIST SP 800-145, A NIST definition of cloud computing, [online] http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

[2] NIST SP 500-292, Cloud Computing Reference Architecture, v1.0. http://www.nist.gov/customcf/get_pdf.cfm?pub_id=909505

[3] NIST Special Publication NIST SP 1500: NIST Big Data Interoperability Framework (NBDIF) [online] http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf

[4] Demchenko, Yuri, Peter Membrey, Cees de Laat, Defining Architecture Components of the Big Data Ecosystem. Second International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2014). Part of The 2014 Int Conf. on Collaboration Technologies and Systems (CTS 2014), May 19-23, 2014, Minneapolis, USA

[5] Grey, J. 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/

[6] Demchenko, Y., P.Membrey, P.Grosso, C. de Laat, Addressing Big Data Issues in Scientific Data Infrastructure. First Int Symp 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.

[7] NIST Big Data Working Group [online] http://bigdatawg.nist.gov/

[8] 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

[9] Data Lifecycle Models and Concepts. [online] http://wgiss.ceos.org/dsig/whitepapers/Data%20Lifecycle%20Models%20and%20Concepts%20v12.docx

[10] DCC Curation Lifecycle Model [online] http://www.dcc.ac.uk/resources/curation-lifecycle-model

[11] Demchenko Y., et al, CYCLONE: A Platform for Data Intensive Scientific Applications in Heterogeneous Multi-cloud/Multi-provider Environment. In Proc IEEE IC2E Conf. 4-8 April 2016 Berlin

[12] Marx, V., Biology: The big challenges of big data. Nature, 2013 vol. 498 (7453) pp. 255-260

[13] Stephens Z.D., Lee S.Y., Faghri F., Campbell R.H., Zhai C., Efron M.J., *et al.*, Big Data: Astronomical or Genomical? PLoS Biol, 2015 vol. 13 (7): e1002195.

[14] French Institute of Bioinformatics – CNRS IFB UMS3601, http://www.france-bioinformatique.fr/

[15] The U.S. Health Insurance Portability and Accountability Act (HIPAA) and Health Information Technology for Economic and Clinical Health (HITECH) [online] http://www.hhs.gov/ocr/privacy/hipaa/administrative/statute/hipaastatutepdf.pdf

[16] Lacroix, T., Loux, V., Gendrault, A., Hoebeke, M. and Gibrat, J.F. Insyght: navigating amongst abundant homologues, syntenies and gene functional annotations in bacteria, it's that symbol! Nucl. Acids Res., 2014 vol. 42 (21): e162.

[17] C.Ngo, P. Membrey, Y.Demchenko, C. de Laat, Policy and Context Management in Dynamically Provisioned Access Control Service for Virtualised Cloud Infrastructures. The 7th Int Conf on Availability, Reliability and Security (AReS 2012), 20-24 August 2012, Prague, Czech Republic. ISBN 978-0-7695-4775-6

[18] 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.

[19] Davis, Jennifer; Daniels, Katherine (2015). Effective DevOps. O'Reilly. ISBN 978-1-4919-2630-7.

[20] Chef: Cloud Automation deployment and DevOps platform [online] https://www.chef.io/chef/

[21] Puppet: Cloud Automated Provisioning and Management https://puppetlabs.com/

[22] Ansible IT automation tool [online] http://docs.ansible.com/ansible/

[23] SlipStream Cloud Automation [online] http://sixsq.com/products/slipstream/

[24] Demchenko, Y., et al, ZeroTouch Provisioning (ZTP) Model and Infrastructure Components for Multi-provider Cloud Services Provisioning. In Proc IEEE IC2E Conf. 4-8 April 2016 Berlin

[25] Zero Touch Network-as-a-Service: Agile, Assured and Orchestrated with NFV, TMForum, July 2015 https://www.tmforum.org/events/zero-touch-network-as-a-service-naas-agile-assured-and-orchestrated-with-nfv/

[26] GÉANT pan-European network [online] http://www.geant.org/Networks/Pan-European_network/Pages/Home.aspx

[27] Demchenko, Y., C. Lee, C.Ngo, C. de Laat, Federated Access Control in Heterogeneous Intercloud Environment: Basic Models and Architecture Patterns. In Proc IEEE International Conference on Cloud Engineering (IC2E), March 11, 2014, Boston, USA

[28] Yuri Demchenko, Cosmin Dumitru, Sonja Filiposka, Taras Matselyukh, Damir Regvart, Migiel de Vos, Tasos Karaliotas, Kurt Baumann, Daniel Arbel, Cees de Laat, Open Cloud eXchange (OCX): A Pivot for Intercloud Services Federation in Multi-provider Cloud Market Environment, IEEE 4th International Workshop on Cloud Computing Interclouds, Multiclouds, Federations, and Interoperability (Intercloud 2015), In Proc IEEE Int Conf on Cloud Engineering (IC2E), March 12, 2015, Tempe, USA

[29] Sonja Filiposka, et al, Distributed cloud services based on programmable agile networks. Proc TERENA Networking Conference (TNC16), 13-16 June 2016, Prague, Czech Republic.

[30] Topology and Orchestration Specification for Cloud Applications, Version 1.0. Candidate OASIS Standard. 11 June 2013. [online] http://docs.oasis-open.org/tosca/TOSCA/v1.0/TOSCA-v1.0.html