DOI 10.1109/CISIS.2013.87

Data as a Service (DaaS) for Sharing and Processing of Large Data Collections in the Cloud

Olivier Terzo
Istituto Superiore Mario Boella
Turin, Italy
e-mail: terzo@ismb.it

Pietro Ruiu
Istituto Superiore Mario Boella
Turin, Italy
e-mail: ruiu@ismb.it

Enrico Bucci
BioDigitalValley
Milan, Italy
e-mail: enrico.bucci@biodigitalvalley.com

Fatos Xhafa
Universitat Politècnica de Catalunya
Barcelona, Spain
e-mail: fatos@lsi.upc.edu

***Abstract*—Data as a Service (DaaS) is among the latest kind of services being investigated in the Cloud computing community. The main aim of DaaS is to overcome limitations of state-of-the-art approaches in data technologies, according to which data is stored and accessed from repositories whose location is known and is relevant for sharing and processing. Besides limitations for the data sharing, current approaches also do not achieve to fully separate/decouple software services from data and thus impose limitations in inter-operability. In this paper we propose a DaaS approach for intelligent sharing and processing of large data collections with the aim of abstracting the data location (by making it relevant to the needs of sharing and accessing) and to fully decouple the data and its processing. The aim of our approach is to build a Cloud computing platform, offering DaaS to support large communities of users that need to share, access, and process the data for collectively building knowledge from data. We exemplify the approach from large data collections from health and biology domains.**

Keywords-Cloud Computing; Data as a Service; Large Data Collection; Sharing; Health Data Collections, Genomics.

I. INTRODUCTION

Data as a Service (DaaS) is an emerging service on cloud technology. For large users communities in scientific and also in industrial context it becomes a serious difficulty for moving data due to time transfers and network link limitations. Datasets are growing constantly and the process analysis in a large dataset needs high computing capacities. Large dataset dimensions are not suitable for data transfer to the users that will determine in a short-term new approach for new services on data and resources management for data location and for process analysis. This paper intends to propose a different approach on Data as a Services (DaaS) for large communities by introducing new concepts: decoupling data sharing and remote processing by moving applications non data.

The first concept is related to the necessity to create specific services for identifying datasets in share and respective location, in a large dataset context in which moving data is not feasible. The second is related to analysis and data processing, in an ecosystem representing by a large community it can be assumed to not moving data but locating applications and services close to data. The main objective of the paper, therefore, is to delineate a framework for building tools and services for applications migration on cloud. The paper is also aimed to provide users on a large community in Life Science to understanding how it's important to make federations of cloud infrastructure for reducing fragmentation for sharing resources and data. The paper is organized as follows: Section II is on DaaS main requirements and state of the art. Section III is related to large dataset in Life Science field. Section IV explains the necessity to decoupling data sharing and data processing. Section V is on Knowledge as a Service (KaaS), the Section VI makes some considerations to implementation perspective for the new approach. Conclusions are given on the last section.

II. DATA AS A SERVICE (DAAS)

DaaS is an alternative cloud computing service model, different from traditional IaaS, PaaS and SaaS models, where data are made available to users as a service through network. Since data is the value of this model, it is fundamental to be able to manage and process the largest quantity of heterogeneous data in order to enable broad and timely access to knowledge-critical, e.g. business-critical,

information. For this reason DaaS is strictly related to big data and must benefit from its technologies.

Sharing, accessing and processing large data sets is not new. In many scientific communities, for instance, that of High Energy Physics, the Grid technology (in its form of DataGrid) [22]. However, despite big data and cloud computing are two of the fastest-moving technologies [17], current computing technologies are not able to satisfy the needs this explosion in data that is creating challenges and prompting innovation in computer storage and processing, as well as in the design of architectures [12][13].

A. Big data challenges and requirements

Nowadays, in big data terms, researchers define the big data as a dynamic model of Vs: volume, velocity, variety, veracity, and value. Indeed, big data problems raised up when the management and elaboration of data is hampered by specific characteristics like:

- volume, the quantity of data that should be managed;
- velocity, the speed at which data is handled at the production, acquisition and elaboration stage.
- variety, types of data that should be taken in account (e.g. different formats and data structure);
- veracity referes to the trustworthy features of the data
- value refers to the real value behind the data, being it business value, useful knowledge, etc.

The exponential growth of heterogeneous data due to the diffusion of new technologies (like mobile, cloud computing, Internet of Things, social networks) poses significant problems in the design of infrastructure components, solutions and processes able to store, access and manage big data objects in a feasible way. Moreover big data problems cannot be faced with traditional systems but require new approaches in data analysis algorithms and computing architectures. These issues are driving new technologies as proved by a lot of research communities in scientific and business world that are facing these challenges [11], [12], [13], [14], [15].

For making usable and processing quickly this huge and complex amount of data, will be necessary to meet numerous demanding requirement.

The most important is the data access, should be allowed to single or different groups of users to store, find, access and use data in a trusted way, thanks to reliable fine-grained access controls in collaboration with systems for tracking data use and for keeping data integrity. Moreover, new services should enable and foster global collaboration between users to discover and to share data in order to allow to work on the same data sets, in a collaborative way and build knowledge collectively. Furthermore future technologies should enable users to access data without the need of moving large volumes of data. For addressing all these challenge will be very important to make improvements in network performances and widespread APIs.

Due to the variety and diversity of data formats, types and structure composing big data, interoperability is one of the main challenges in big data management. Standardization should be one of the key to solve the problem, but another aspect to take into consideration is the semantic of the data that becomes an important issue to allow for usability.

In order to manage this huge amount of data, Big Data infrastructure should be also equipped with hardware and systems that provide storage, aggregation and preservation of data for unlimited period of time and at the same time must ensure data security.

From the processing point of view, it should be taken in account that due to heterogeneity of users and data types, the elaboration of data includes a variety of software technologies each with specific computing requirements in term of storage, CPU, memory and networking: this reflects in the requirement of an highly dynamic supporting infrastructure to allow wide data access and distributed processing. Thus the underlying computing infrastructures should be enough flexible and scalable in order to respond to any kind of computing requirements.

B. Architectural aspects

DaaS systems should own extremely dynamic characteristics in order to respond to requirements from different communities of users.

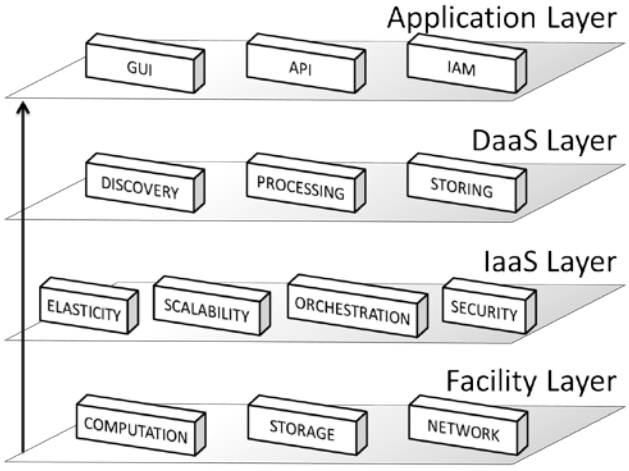

Figure 1 – DaaS generic architectural layers

In Figure 1, it is depicted a generic architecture of a DaaS system, represented through levels of abstraction, from infrastructure to applications:

Infrastructure layer - This is the bottom layer of the architecture and consists of physical systems that are used to process and store data. Hardware components like servers, storage, network and any other physical devices are part of this layer. It may be owned, managed, and operated by the end user, an external provider, or some combination of them, and it may exist on or off premises.

IaaS layer - This layer is in charge to manage computing resources, mainly thanks to virtualization technologies, that allow the abstraction and the pooling of physical resources (CPU, memory, I/O, network) that can be managed and controlled by orchestration tools in a centralized way. Virtualization enables for decoupling physical resources from underlying infrastructure. The pooling and virtualization of physical resources, fundamental aspects in a cloud environment, are essential for achieving the elastic characteristics.

This layer must control resource allocation procedures, balancing between the needs of running applications and the effective capacity of the infrastructure. Thanks to policies and schedulers it can optimize resource utilization improving performances and reducing costs and energy waste. This layer must also provide tools to enhance scalability such as standard interfaces and federation approach in order to increase computational capacity from external resource providers if needed. Another important aspect that should be addressed from this layer is the security. It is needed the integration of different security technologies, permitting to fulfill privacy requirements in a multi-tenancy environment and also to facilitate federation of infrastructures.

DaaS layer - This layer is composed by all the technologies needed to manage data in order to discover, process and store it from different sources. These tools can be categorized in three groups each with a specific function:

Discovering: are systems suitable to find or collect data from heterogeneous sources, to aggregate data and if needed to feed it into data stores. These tools look for pertinent data in distributed sources (federated datacenters, community clouds, open repositories and other sources) and return location and other information about the most appropriate source.

Processing: are software components for processing huge datasets, like data oriented distributed computing software inspired by MapReduce, or in-memory computing software, that allow to store data in the Random Access Memory (RAM) in spite of slow hard drives, improving performances.

Storing: are scalable technologies designed to face issues posed by big data scale but also by elasticity of cloud computing. The most promising are No-SQL (Not Only SQL) databases. They break from the relational database management system (RDBMS) model, where data is stored in the form of tables, as is the relationship among the data. They can be characterized as non-relational, distributed and horizontally scalable systems.

Application layer - The end users have direct access to this layer through specific interfaces (API) with the aim to exploit the computing infrastructure and the richness of the data provided, allowing performing analysis through provided GUI or building specific applications. Furthermore Identity and Access functionality (IAM) enables to securely control access to the DaaS service and to computing resources of underlying infrastructure. It can be possible to create single users or groups with different utilization permissions.

III. LARGE DATA SETS

A. Genomics

As a first example of the problems associated to dealing with large volumes of data, we will consider the case of genomics, i.e. the data directly generated by or associated to the sequencing of DNA from different biological data sources. While a single genomic sequence do not constitute a large set of data for current standard – a human genome consists in about 100 gigabytes of data – the steady increase in the number of single species sequenced [18], and even more the current effort to sequence thousands of human genomes [19], have already produced petabytes of raw data. In the context of personalized genomics, a scenario consisting of some billion of human genomes is at hand, as more people want their DNA sequenced for diagnostic and prognostic purposes, meaning hundreds of thousands of petabytes. Moreover, raw DNA sequences are annotated, to produce intelligible and rich genomics information, further increasing the overall data volume. Storage of such a large data volume is increasingly becoming a problem per se, leading to the development of efficient compression algorithms related to the structure of sequencing data [20]. However, even in presence of a sufficient storage capacity, the intrinsically delocalized production of high data volumes exceeds our capacity to transfer them at a sufficiently high pace to centralized repositories, so that they often resides in local databases. Thus genomics data are not only large, but

also sparsely distributed over different data repositories, whose synchronization will be increasingly difficult as the volume of new data will increase and the diversity of the data format grows. As a matter of fact, a recent count of only the most important, active and open repositories of biological data, has topped 1500 [21]; many of the datasets in this count are genomics databases, and several others are very useful for a fully inclusive genomics analysis (i.e. they contains data on proteins, instead of DNA sequences, or on diseases related to genetic variants, or on other biological aspects correlated to the genetic sequence of an individual). This leads us to two other challenges related to the deluge of accumulating genomics data, i.e. availability and integration among different formats and different types of data.

B. Patient Data Sensing and Clinical Records

A second realistic and relevant example comes from health domain. Specifically, we refer to data sensing from patients and records. With the fast development in wireless sensor and mobile technologies, it is possible to monitor thousands of patients at hospitals, care-centers and homes. On the one hand, monitoring patients has the advantage of offering more accurate halth service to them. On the other hand, in some relavant cases, remote monitoring is becoming a must in health domain. This is the case of elderly, the population of older people, especially those suffering from mental diseases such as Parkinson, Dementia, etc. As a result of aging societies, the number of potential patients has increased dramatically and the current economic and medical infrastructures are not able to give support to patients via hospitals or care-centers. Remote monitoring of patients has thus become a solution being considered from health domain. The main challenges behind monitoring solutions, are obviously those of big data: volume, velocity, variety, veracity, etc. One can reach easily to gigabytes and terabytes of data if a thousand of patients would be 24x365 monitored even with a few parameters being measured and stored. This amount would explode if full context information would be catered. Further, this amount of data would increase if further layers of data (e.g. complex events built by combination of smaller events). In all, the data sensing from patients monitoring produces big data volumes, which should follow a full cycle of data: capturing, gathering, cleaning, transforming, formatting, storing, analyzing, and visualizing. In some case, and depending on patient's state, all this cycle should be covered in real time. Therefore, technologies other than Hadoop-like batch processing technologies, should be considered. Another data source that caters for data diversity and variety is that of patient clinical records. The data record per patient along many years, increases significantly in elderly population. In all cases, being the data sensored or exising clinical records, the data should be available for access by various teams of doctors, carers, nurses, adminsitrativ and social agents, etc..

C. Other Types of Data Resources

There are many other examples of applications and domains that produce big data scenarios. These include data analytics from enterprise domain, learning analytics from Virtual Campuses and Virtual Organizations, Social Networks, World Sensing (GeoSpatial ensing, Smart cities, etc.)

IV. DECOUPLING DATA LOCATION AND DATA PROCESSING

Datasets and applications in Scientifics Disciplines like life sciences are growing constantly and involve large users communities for simulations more often in large-scale experiments [8][9]. Researchers are located over a wide geographic area and needs strong mass of storage systems, high-end computing facilities and networking. This trend require new paradigms for facilitate co-operation, co-ordination between researchers through web technologies.

On the last decades specifics needs were on facilities for transferring large sized of datasets in a short time and facilities for supporting high performance computing analysis, this vision now needs an intensive focuses on investments in IT resources in order to be able to follow requests of analysis from scientific community.

In consideration of this before for accessing on large dataset Data Grid [4] platform were designed for addressing replica of catalog and involved generally terabyte of data clearly more complex for massive dataset characterized by the presence of more repository for transferring, storing outputs and analysis. Data Grid involved a huge size of data in transfer considering the dimensions of dataset and this approach is no longer acceptable today. In fact due to the constant growing of dataset and analysis replica mechanism will face to the problem of latency for data transfers.

The new approach proposes in massive datasets is to change the paradigm by moving progressively from replication services to services for finding, sharing and processing data (see Figure 2). In a scientific community in life science for example specifics needs are on data retrieval, in most case scientist and researchers need to make more tests for the algorithms development with the support of open existing dataset.

Researchers and scientists are facing difficulties for simulations and in particular on:

1. Finding data: this is the first main challenge, in most cases algorithms and applications are developed with a small dataset of information. When the first phase on code consolidation development for an application is completed, the second step is on trial test in a large dataset and often this is a strong constraint, finding dataset and right data for a strong validation.
2. Moving and storing data: the second big constraint is on getting, moving data and the time duration for obtaining data, another strong limitation is on simulations with a huge dataset who need strong capabilities in term of local storage.
3. Processing data: the other aspect is on data processing and this may represent the main limitation, more often code written are not think for an immediate integration on modern computing infrastructure and often the computational available resources are limited.

As explain before three considerations are emerging on e-science. Firstly researchers for finding data need to used services and tools fully integrate in specific DaaS layer. Finding data became a strong problem and represents an additional constraint being given the ever-increasing number of existing database always more numerous, which led researchers to lose a lot of time to isolate the correct database. In first instance to know where data are located each for their specific applications domain, more activities are dedicate for producing datasets and for getting data that will be named data resources but in term of data management for data provenance no more services are developed for dataset discovery. This is a critical issue on the e-science communities for avoiding duplication of datasets for facilitate the searching and location and for knowledge in term of existing dataset in various domains or subdomains on e-science.

On Data as a Services a standard base services for Data discovery will be useful, each dataset will be cataloged to assist the scientific community in dataset search phases and exploitation.

Related to the data discovery considering the high production of tools and applications by the scientist community moving data became a bottleneck due to the dimension of data involved for simulation analysis specifically where more data are involved, moreover on specific fields like genomics in large dataset context [3] moving data will be not acceptable due to limitations on network capabilities and in most cases for local storage limitations.

The sensible evolution is on enabling and facilitates collaboration between researchers by changing scenarios on data sharing and applications processing.

Main challenges and perspectives are on moving applications and algorithms to servers rather than moving data the amount of data and applications developed by scientist. On the few decades datasets and applications grew in an exponential way, this introduces the new approach proposed and the concept on decoupling data discovery and data processing.

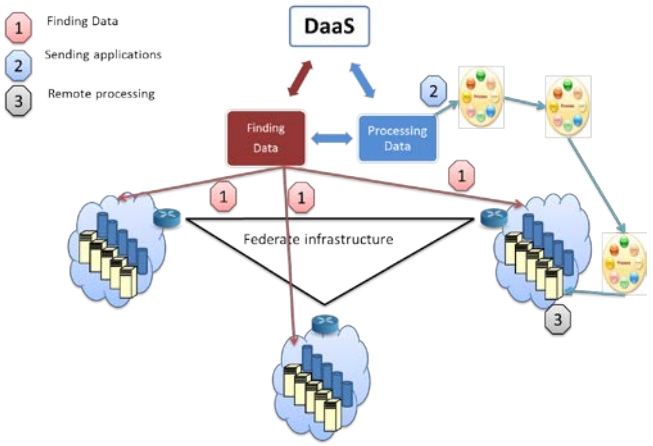

Figure 2 – Managing big data through DaaS

In a community who involve large data collection a first step for improving collaborations and cooperation is on pass through a federation of existing cloud infrastructure for sharing inside the community resources, storage and services [1]. It's mandatory to integrate and federate resources for avoiding fragmentation and duplication of data collection especially in a large dataset context and creating specific services for the community on data discovery, data sharing and data processing.

V. COMMUNITY KNOWLEDGE BUILDING

Here we should indicate how knowledge can be build and shared among members of a community…*knowledge as as a service* (KaaS).

A. Knowledge as a Service (KaaS)

In a community of specialists, scientist, researchers knowledge is the heart of activities and the cost for establish in long term collaborations is high considering the fragmentation of activities, applications and results.

In a federate vision of infrastructures and data shared, it can be easy to delivery knowledge in a short time for increasing cooperation and direct access to information. The approach proposed is an enable framework for improving KaaS by having the capabilities for a quick acquisition of new knowledge inside the community just by considering the capabilities to find existing or new data available inside the community. That can bed names knowledge acquisition. In a traditional service a user must search, browse in a non structured service just considering for example page rank and trend analysis. In a cataloged vision of a shared dataset services users are in condition to be very fast for finding and for knowing existing dataset in share. The knowledge is the added value and it's enable the KaaS. Additionally, knowledge can be built increasingly; the findings of some users can be exposed via services to other users. Additionally, we envision the composition of knowledge services from smaller data sets so that knowledge on larger data sets can be obtained. Therefore, users will not only share the data but also the knowledge they have on data.

VI. IMPLEMENTATION PERSPECTIVE

An accurate catalog of Datasets Identification and resources available will be implemented in addition to the two distinct discovery and processing cloud services on DaaS. The services catalog will be designed for supporting workflow automation from the dataset identification request, applications transfer to run process for data applications analysis.

For enabling a cloud based infrastructure in a DaaS approach where take into account a separation between data sharing and data processing, a redefinition of cloud users services will be analyzed for reducing complexity and improving cloud acceptation specifically on e-science domains who researches need a rapid deployments of data selection and data processing [2].

It's essential to define a standard in creating a Data Discovery Services for capturing data and metadata dataset requirements for a better identification of user needs.

For example in a large community Open standard [6] could give benefit for a flexible and less complex approach in term of cloud computing resources and data declaration and for facilitate sharing of data and resources in an easy way for making collaboration.

On the other hand for data sources catalogation an open standard will help for a better knowledge of datasets available and share inside the community. A common standard for dataset definition and dataset discovery in an implementation perspective will be a useful service for dataset providers and for users. Sharing data its essential especially in a common field of work and it's a necessity for improving collaborations. The methodology for creating a Dataset Discovery Services will be based on several steps (see Figure 3).

First, declaration of each dataset in share for the community for making a global catalog of each dataset available, the declaration consist in describe typology of data, datatype, application domain, dimension and so on.

Second, a user who wants to access on a share dataset must fill a form for a characteristics description of datasets requirements [5].

Third, the Data Discovery Services query the catalog for an identification and dataset discovery. Finally user gets information concerning dataset accessibility and availability.

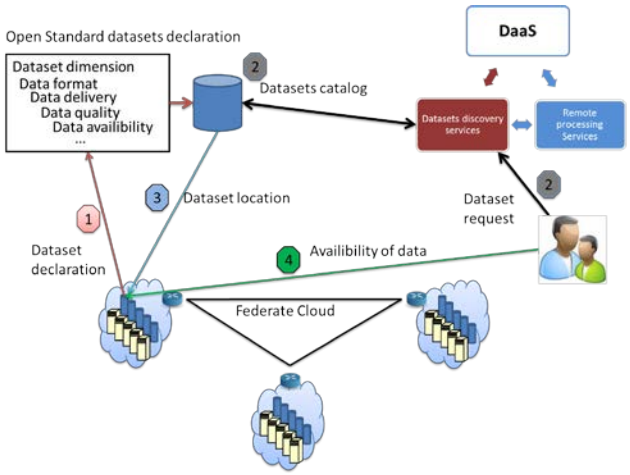

Figure 3 - Dataset Discovery Services

On the same spirit, in parallel to the Datasets Discovery Service it's necessary to make a strong focus on interoperability in computational resources for data analysis and processing. The new approach consists in moving applications and algorithms to datasets and requires a strong formalization in term of computing capabilities offers by community.

On the same way a standard for resources capabilities will be define for an implementation perspective on this sense the catalog of services for processing data could help users for getting the right resources computing capabilities available. After Dataset Discovery Services the data processing services will in charge to select the right resources available where the dataset is located corresponding to the resources requirements declared by users (see Figure 4). The approach will be resume in:

First, the community declare resources capabilities in term of hardware, instances available like Amazon [8][9] philosophy for making a global catalog of computing resources put at the users disposal by the community.

In a typical scenario after the declaration of dataset identification users declare what kind of resources is needed for the processing phases. They must fill a description or a selection of template and instances number needed, duration of uses, capabilities in term CPUs and RAM of each instances. In this way the Remote Processing Services query the catalog of resources available and provide to give the maximum availability of resources for the user at the moment. Finally the user is in condition to use the dataset and to provide to transfers the application environment for starting the processing phase.

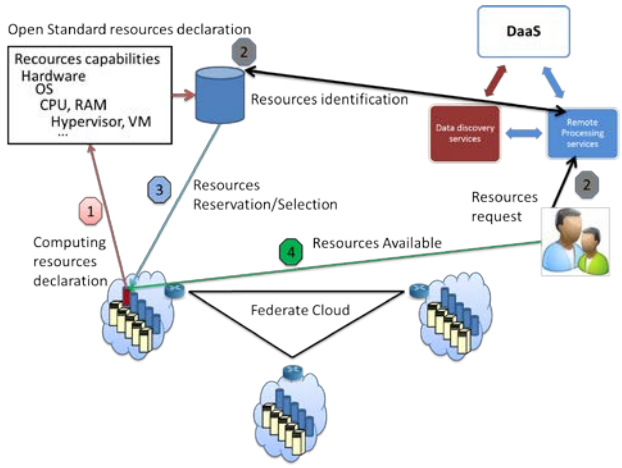

Figure 4 - Resources Location Services

The framework implementation proposed aim to give a new perspective for performing a quick processing of data in a large dataset context. This main challenge need to define new methods and technics for increasing sharing, collaborations on large scale dataset processing in a community [7] of users.

VII. CONCLUSIONS AND FURTHER WORK

This paper propose a new approach in managing large amount of data through a novel DaaS architectural model. It has been delineate a possible framework for building tools and services for applications migration on cloud. The main innovation consists in decoupling data from processing by moving applications instead of data.

We give some examples of large datasets in Life Science such as Genomics and Clinical Records and of challenging arising from Big Data management as data access, interoperability, infrastructure elasticity.

The core of the new approach proposed is composed by tree main steps: finding , moving and storing, processing data. The proposed framework consists of two distinct Services: the first is a Dataset Discovery Service, for finding data in distributed location through a catalog with pen and well-defined metadata formats and protocols; the second is the Resource Location Service, for selecting suitable computing resources in order to elaborate datasets.

A consideration arise from this architecture: in order to elaborate data without moving it, it is needed that the site (eventually a cloud node) that hold selected dataset possess also enough computing resources in order to satisfy the computing requirements of delivered applications.

REFERENCES

[1] SNIA: Advancing storage & information technology, Implementing, serving and Using Cloud Storage. Cloud storage Initiative, http://www.snia.org/sites/default/files/2010-10-WP-ImplementingServingandUsingTheCloud.pdf

[2] Dana Pectcu, José Luis Vàsquez-Poletti, “European research Activities in Cloud Compting”, 2012

[3] Frédéric Magoulès. Jie Pan, and Fei teng “Cloud Computing, Data-Intensive Computing and scheduling”, Multidimensional data analyis in a cloud data center, 2012,

[4] Srikumar Venugopal, Rajkumar Buyy, Ramamohanarao Kotagiri, “A taxonomy of Data Grids for Distributed Data Sharing, Management and processing”, ACM Computing Surveys Volume 38 Issue 1 article N° 3, 2006.

[5] Greg Shulz, “Cloud and Virtual Data Storage Networking”, CRC press, Taylor & Francis Group, 2012

[6] Andy Edmonds, Thijs Metsch, Alexander Papaspyrou, Alexis Richardson, “Toward an Open Cloud Standard”, Internet Computing, IEEE, Volume 16, p15-25, Issue 4, 2012

[7] Samira Daneshyar, Majid Razmjoo, “Large Scale data Processing Using Mapreduce in Cloud Computing Environment”, International Journal on Web Service Computing (IJWSC), Vol.3, No.4, December 2012.

[8] Strategic Plan for a Scientific Cloud Computing Infrastructure for Europe, http://cds.cern.ch/record/1374172/files/CERN-OPEN-2011-036.pdf

Riding the wave “How Europe can gain from rising tide of scientific data”, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[9] Rudolph Strijkers, Willem Toorop, Alain van Hoof, “AMOS: Using the Cloud for On-demand execution of e-Science Applications”,

[10] Amazon Inc., "Amazon Elastic Compute Cloud (Amazon EC2)". [Online]. Available: http://aws.amazon.com/ec2/ ,2013

[11] Cisco Global Cloud Index: Forecast and Methodology, 2011–2016 http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns1175/Cloud_Index_White_Paper.pdf

[12] PARADE - Partnership for Accessing Data in Europe

Strategy for a European Data Infrastructure (White Paper) published on Fri 2nd October 2009

http://www.csc.fi/english/pages/parade/

[13] GLOBAL RESEARCH INFRASTRUCTURE SUB GROUP ON DATA,

e-Infrastructure activity, Draft Report, 28 October 2011

http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/g8.pdf

[14] IBM Research - Understanding System and Architecture for Big Data

Anne E. Gattiker, Fadi H. Gebara, Ahmed Gheith, H. Peter Hofstee, Damir A. Jamsek, Jian Li, Evan Speight, Ju Wei Shi, Guan Cheng Chen, Peter W. Wong

http://domino.watson.ibm.com/library/CyberDig.nsf/1e4115aea78b6e7c85256b360066f0d4/f085753cf57c8c35852579e90050598f!OpenDocument%26Highlight=0,rc25281

[15] Audrey Watters The Age of Exabytes, Tools And Approaches For Managing Big Data, ReadWriteWeb and HP, 2010, http://readwrite.com/2012/03/05/big-data

[16] Yuri Demchenko, Zhiming Zhao, Paola Grosso, Adianto Wibisono, Cees de Laat,

"Addressing Big Data Challenges for Scientific Data Infrastructure

System and Network Engineering", Group University of Amsterdam

Amsterdam, The Netherlands

2012 IEEE 4th International Conference on Cloud Computing Technology and Science

[17] Gartner's 2012 Hype Cycle for Emerging Technologies Identifies "Tipping Point" Technologies That Will Unlock Long-Awaited Technology Scenarios

STAMFORD, Conn., August 16, 2012

http://www.gartner.com/newsroom/id/2124315

[18] "How Many Species Had Their Genomes Sequenced? | UA Magazine." [Online]. Available: http://www.united-academics.org/magazine/earth-environment/good-to-know-how-many-species-had-their-genomes-sequenced/. [Accessed: 10-Feb-2013].

[19] G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean, "An integrated map of genetic variation from 1,092 human genomes.," *Nature*, vol. 491, no. 7422, pp. 56–65, Nov. 2012.

[20] M. Hsi-Yang Fritz, R. Leinonen, G. Cochrane, and E. Birney, "Efficient storage of high throughput DNA sequencing data using reference-based compression.," *Genome research*, vol. 21, no. 5, pp. 734–40, May 2011.

[21] X. M. Fernández-Suárez and M. Y. Galperin, "The 2013 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection.," *Nucleic acids research*, vol. 41, no. D1, pp. D1–7, Jan. 2013.

[22] BUNN, J. AND NEWMAN, H. 2003. Grid Computing: Making the Global Infrastructure a Reality. Wiley Press, London, UK, Chapter Data Intensive Grids for High Energy Physics.