Big Data Architecture Evolution: 2014 and Beyond

Atif Mohammad
Department of Computer Science, University of Quebec at Chicoutimi
555 Boul De l'Universite, Chicoutimi, Québec, Canada, G7H-2B1
+14185455011
atif.mohammad1@uqac.ca

Hamid Mcheick
Department of Computer Science, University of Quebec at Chicoutimi
555 Boul De l'Universite, Chicoutimi, Québec, Canada, G7H-2B1
+14185455011
hamid_mcheick@uqac.ca

Emanuel Grant
Department of Computer Sciences, University of North Dakota, USA
3950 Campus Road, Stop 9015
Grand Forks, ND 58202-9015
+701.777.4133
grante@cs.und.edu

ABSTRACT
This paper aims at developing the Big Data Architecture, and its relation with Analytics, Cloud Services as well as Business Intelligence. The chief aim from all mentioned is to enable the Enterprise Architecture and the Vision of an Organizational target to utilize all the data they are ingesting and regressing data for their short-term or long-terms analytical needs, while making sure that they are addressing during the design phase of such data architecture for both directly and indirectly related stakeholder. Since all stakeholders have their relative interests to utilize the transformed data-sets. This paper also identifies most of the Big Data Architecture, threat analysis within a Big Data System and Big Data Analytic Roadmaps, in terms of smaller components by conducting a gap-analysis that has significant importance as Baseline Big Data Architecture, targeting the end resultant Architectures, once the distillation process of main Big Data Architecture is completed by the Data Architects.

Categories and Subject Descriptors
D.2 [**Software Engineering**]: Software Architectures – *Data Abstraction, Domain-specific architectures, Information hiding, Languages (e.g., description, interconnection, definition), Patterns (e.g., client/server, pipeline, blackboard).*

General Terms
Design, Reliability, Security, Standardization

Keywords
Architecture, Big Data, Cloud Computing

1. INTRODUCTION
Let us take a Healthcare concern and later we will look at an Automotive Service provider as two examples from our contemporary enterprise structure. In this example our Healthcare concern has selected to undertake 360 degrees change in their architectural transformation to start utilizing all their Big Data utilization and processing needs. It is significant to recognize and address data management issues, since traditional data management is and was done using RDBMS solutions. All the data management, such as data ingestion and egress has been dealt with traditional SQL-oriented applications. A regulated and inclusive methodology as per compliance reasons to healthcare laws applicable to data management is in place for quite some time that has enabled the effective use of data to capitalize for both healthcare and other associated services on its competitive advantages.

The considerations/strategy towards altered path to introduce Big Data Architecture are including the following points.

It is vital to establish a clear characterization of which application components out of the organizational system in the broader landscape, which will be utilized as the system of data storage, where the applications will be referencing from as Organizational Master Data Management.

Establishment of an organization-wide norm that all Big Data System's components (dealing with either data ingestion or egress after transformation), this will include software packages (Hive/Pig/Impala etc.), need to adopt prescriptive data models to have a road map that will be open for any new unforeseen needs of the organizational expansion in the future.

Clearly recognize the diverse need of current organizational components on how data entities are employed by business functions, methods, and services. Such as Clinics Applications in relation to Pharmaceutical and other medicinal concerns of all related stakeholders of our healthcare concern?

Clearly understand the ETL (Extract, Transform & Load) processes, by making sure that CRUD (Create, Read, Update & Load) is taken into consideration in terms of Big Data technologies, such as Hadoop/MapReduce.

The level and complexity is to be determined, sine the Analytical needs, will also be introducing the use of tools, such as Tableau, Jaspersoft or the organization can also opt to use 'R' the Statistical programming language for data transformations, that are required to support the message passing between applications within this new Big Data system?

Since Big Data system will be the requiring software to support data integration using Cloud to serve organizational customers (patients) and other stakeholders (medical equipment suppliers, pharmaceutical organizations) use of ETL tools, which can be developed in Java, if using open source, in case of Microsoft C# or Scala using a NoSQL, such as HBase, Cassandra or CouchDB for data migration, data profiling tools, such as Hive, Pig or Impala to evaluate data quality.

DIVANet'14, September 21 - 26 2014, Montreal, QC, Canada

http://dx.doi.org/10.1145/2656346.2656358

This paper is organized as follows. Section 2 presents the related works and state of the arts. Section 3 and section 4 describes respectively the big data governance and the big data ingestion strategy. In Section 5, we describe briefly the big data architecture of an application. Section 6 details the challenges of big data. Finally, this paper is concluded and a future remarks are given.

2. STATE OF THE ART WORK

There are inter-organizational and external Architecture inputs as well as non-Architecture inputs, which are needed for a Big Data Architecture. First one of non-Architectural input is a "Communication Plan". It is vital for Big Data Architect to identify the stakeholders both internal and external to an organization for communication needs at the time of architecture design. These communication needs can be classified in terms of as "Critical Success Factors" towards a comprehensive Big Data Architecture containing vision and risks associated to future IT applications/services for the organization.

This communication management also needs to be further distilled in terms of certain mechanisms to share architecture information within internal stakeholders and permit external stakeholders to share their feedback in meetings to have a critique done on the designed architectural components. This plan also needs a detailed timetable, depicting which communiques will follow with which groups of stakeholders with time and physical location constraints.

An architecture describes the basic structure of a system with its elements, the relations between these elements and the relationships of the system to the environment [1]. In addition, principles for the design, development and use of the system should be described [2] are. An architecture of other models (1) differs by a holistic view of a system in terms of the width of the elements under consideration, and (2) by a coarsened consideration.

The data architecture by Zachman [3] represents, a specific part architecture of the information system architecture, and describes the data structure of an information system. In the data architecture such entity types, which we use in our daily data modeling and their relationships to each other are shown. In terms of a coarsened consideration it may be necessary to map only selected entity in the data architecture or summarize the data architecture entity types [4].

For Big Data Architecture, there is no specific way. However, semi-formal notations can be used by practitioners. Big Data Architectures can be referred to as enterprise data architecture to the basic data structures of an entire company or refer as the data architecture of an application system on a section of the company. However Big Data Architectures are to be distinguished from the enterprise data models, which have the aim of documenting the history of a company. This is an approach, which is being utilized from the 90s, however, is uptight with difficulties, since an immense care expense is created by the constant change of a company.

As per the research conducted by the Gartner Research, [5] the global resources of data warehousing currently available at various data centers around globe and several business intelligence solutions provider's managed data, such as Oracle, Microsoft, Yahoo and Google was predicted to attain in 2011 the amount of US$10.8 billion.

A Cloud Data warehouse is created to hold snapshots of production data used for business trending and analysis using SOA, this can be implemented by using SRD described by SOA 3.0 [6] to adopt and work around to establish and create a Cloud Data warehouse. In the contemporary world of business today, it is vital for Cloud Analysts to look deep into a business to formulate the objectives, which is obviously the desired business goal that can be achieved by the Cloud Data warehouse. The first step can be taken as to start a brainstorming session with the related stakeholders of the business needing a Cloud Data warehouse. These stakeholders should be corporate project sponsors, subject matter experts and technology experts associated with the business requiring the Data warehouse.

The storage is the cheapest most utility available and this utility has introduced data related challenge for all related stakeholders. The management of this ever increasing data has scalability beyond imagination due to the electronic use of services, such as B2B, B2C or B2B2C for e-Commerce and such similar challenge is standing for educational institutions of managing large amounts of research data, which we also call as Big Data. The ever increasing data is being accumulated by the use of all sort of electronic gadgets that we use in our daily life and are now pervasive to us, and we all are feeding databases of such organizations, such as Facebook etc., more than several petabytes on regular basis. The Economist reported in 2010 that there were 4.6 billion mobile-phone users using on global basis and over 1 to 2 billion people were subscribers of the internet [7]. This means clearly that in 2012 we are living in data explosion, which is rightly called Big Data.

This becomes a sound reality that we can use service oriented architecture to use by running several services under monitoring of several master services as utilized by map-reduce [8] or Hadoop [9]. The Apache Hadoop software library provides us the framework to process on distributive basis of large data sets (Big Data), which are available or stored on clusters or virtual clusters available in Cloud [9], by some organization. The library can handle failures during this humongous data processing for any analysis needs requested by some stakeholder(s).

We can say that Big Data can be considered in simple words: telecom connection data, statistics of public private intranet and internet sites, RFID enabled logistics data, utility, such as water, energy consumption, financial organizations referral data, such as banks, insurance companies , health and related prescription data, architecture simulation data and atmospheric scientific data, drug discovery for any cure, biological, archaeological, geological exploration data, nuclear physics, molecular biology. All steps associated of the processing – from capture and storage to analysis and visualization are some of the factors, which are posing enormous challenges, but also opportunities to achieve a competitive advantage and for the development of new business models.

3. BIG DATA GOVERNANCE

Data governance considerations is an area in the world of Big Data, which requires significance to ensure that the organization has the essential scopes in sight to sanction the process of transformation. First essential scope is the "Structure", this relates to the organizational structure (departments, such as finance, sales, marketing, human resource etc.) to manage the aspects related to data entity transformation. Second essential scope is the "Information Management System". This system is to deal with standard as well as Big Data for the organization to manage the

governance aspects of data entities throughout its lifecycle within the organizational hierarchy.

Third scope is management of human resource, to align organizational people in relation to their data-related skills and roles, who are to manage and work with their relative data for any transformation processes to utilize or produce every day work. In case the organization lacks such resources and skills, office of CIO/CTO should look into acquiring those critical skills related to Big Data or aligning the training sessions internally to meet the requirements through hiring skilled trainers of Big Data. This entire strategy needs to be documented and stored in an Architectural Repository for any system reengineering needs in future.

4. BIG DATA INGESTION STRATEGY

Since our healthcare concern is already an established organization and using a complete Information System as standard MIS, at the time of the change to Big Data system, there will be a grave requirement to "Extract" current data, which is in the form of Master, Transactional, and Referential entities. The Big Data Architecture must categorize data transfer requirements and also deliver pointers within the Data Ingestion Strategy as to the level of "Transformation" (which should include data cleansing and storing garbage data for any future testing purposes) that will be required to store data in a format (Schemas suited to NoSQL indexing) that meets the requirements (data storage in HDFS – Hadoop Distributed File System and constraints of the target application to extract data using either Mahout or some other machine learning application, designed and developed).

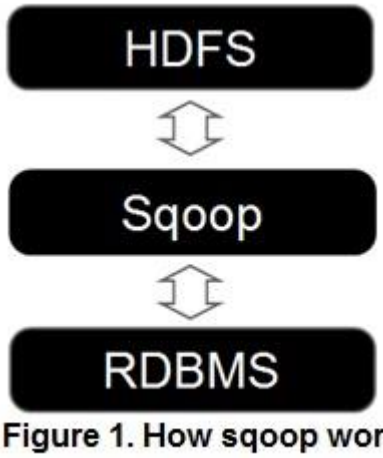

Figure 1. How sqoop works

The main objective of this new Big Data system is being that the data extraction for any every day or analytical needs in the Big Data system's service has quality data loaded at the time of new data store getting cleansed data from RDBMS. While designing data ingestion strategy, it is of quite importance that the architects should establish an organization-wide common data definition to support the transformation processes. The selection of data ingestion tool is extremely important, the tools such as Apache Flume or Sqoop can be utilized.

Data ingestion is the process geared to obtain data from some resource, import in some data format, such as csv etc., and process this data for later use for any needs related to organizational targets, Big Data is usually stored in NoSQL Databases. An effective data ingestion strategy is to utilize a methodology that validates the data files, initially on individual basis, later by prioritizing the sources for data processing, and the end of this process chain is to validate the end products in terms of trends or some resultant data sets. As we know that there are several data sources available in our industrial sectors in different formats, we must look at the strategy that is to maintain pragmatic promptness and proficiency, which can become a foremost challenge.

5. BIG DATA ARCHITECTURE OF AN APPLICATION

The Big Data Architecture of an application system for an organization can be segregated in the:

- Design,
- Development,
- Maintenance and
- Decommissioning of the application system use.

Depending on the design objective, it may be considered at conceptual level:

- Problem analysis,
- Requirements definition
- Conception of business-relevant data structures

Standard legacy practice for architectural display on multiple conceptual level, are in particular a semi-formal modeling languages, such as the Entity Relationship Model [10]. Whereas at design level, the system technical implementation is conceptually specified. The choice of an appropriate form of description, therefore, depends on the database management system used, as we have witnessed in last three decades. For business application systems relational database management systems has been widely used. To describe the data architecture of a relational database design level is, for example, a relational database schema in tabular form [11].

In addition to the Big Data Architecture can be further sub-architectures specify an application system, which enable a different view of the application, such as:

- A process-oriented perspective or
- Business process modeling or
- A task-oriented view
- Function modeling

Individual views conceptually separated from each other is important to reduce the complexity of the system. Nevertheless, there are dependencies between individual views. Different approaches to application modeling therefore, we should consider not only the data view, but also the interdependencies of the views. Examples of such approaches are the Architecture of Integrated Information Systems or Multi Perspective Enterprise Modeling.

All current standards of our current legacy world of Data Architecture, the data view has been widely regarded as the most important reference point for other views, because it gives a consistent and wide-ranging overview of the components of an application system. In the context of relatively newer paradigms such as object orientation or the service orientation, however, are

encapsulated data to a greater extent in completed items or services. This overview of the nature of the data architecture is limiting in terms of Big Data Architecture, but at the same time supported the demand for more flexibility through the principle of loose coupling calculation.

Data Modeling has been used in the development and design of information systems to be used when it comes to the identification and description of the relevant information objects and their relationships. Data Modeling is the formal mapping of the information objects of the considered universe of discourse by means of their attributes and relationships. The goal is to clarify the definition and specification of the managed objects in an information system, it is required for informational purposes only. The attributes and the relationships between different information objects in order to obtain such an overview of the data view of the information system can distill organizational system needs [12].

6. CHALLENGES OF BIG DATA

Big Data refers to the economically viable extraction and use of decision-relevant knowledge from diverse qualitative and differently structured information that is subject to rapid change and occur in an unprecedented scale, whereas Cloud computing, that provides both storage and compute power is going to be the key to work with.

Big Data presents concepts, methods, technologies, IT architectures and tools available to the exponentially increasing volumes of diverse information in better implementation of sound and timely management decisions and thus improve the inventiveness and competitiveness of enterprises.

The major challenges are given below:

- Security and Privacy
- Market transparency and development of data visualization
- Business models from the perspective of the user
- Consumer oriented application challenges
- Best practices and process models
- Marketing of the Big Data Analytics
- Post Big Data on changes in an inter-organizational change
- Opportunities for the creation of infrastructure with servers, storage, middleware
- Platforms selection, such as Hadoop, Big Cache or In-Memory Technology
- Use of BYOD (bring your own device) such as Mobile Devices
- Usability of such applications
- High Performance Computing
- Big Information Management
- Cloud Services Life Cycle Management Techniques
- Ad Hoc Reporting
- Forecasting methods
- Social Opinion and sentiment analysis
- Consumer Behavior Pattern Recognition
- Traceability of result evaluations

Figure 2 shows a novel model to manage RBAC ID Management Cloud-Services Orchestration.

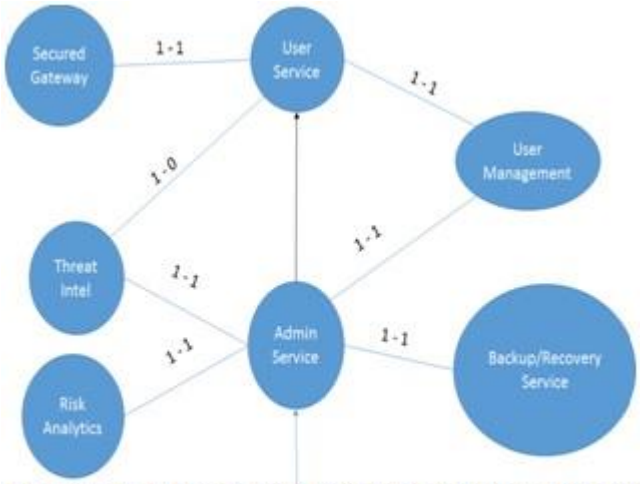

Figure 2. RBAC-Big Data Analytics Threat Manag. Deterministic Finite Machine

- Security and Privacy {Admin Service}
- Business models from the perspective of the user {Secured Gateway + User Management}
- Consumer oriented application challenges {Backup/Recovery Service}
- Best practices and process models {Threat Intel}
- Platforms selection, Hadoop, Big Cache or In-Memory Technology {Secured Gateway}
- Use of BYOD such as Mobile Devices {Secured Gateway + User Management}
- Usability of such applications {Secured Gateway + User Management}
- Big Information Management {Risk Analytics}
- Cloud Services Life Cycle Management {Risk Analytics + Backup/Recovery Service}
- Ad Hoc Reporting {Risk Analytics}
- Forecasting methods {Threat Intel}
- Social Opinion and sentiment analysis {Risk Analytics}
- Consumer Behavior Pattern Recognition {Secured Gateway + User Management}
- Traceability evaluations {Threat Intel + Risk Analytics + Backup/Recovery Service}

The combination of algorithms [13][14] that deal with BDBIM are the secret sauce to deal with Big Data Threat, RBAC, and Risk Analytics are major players in this solution. The architecture that we are looking into is associated with Big Data and is based on

generic hardware, that contains a motherboard with a processor also known as a core these days, a power supply, ram also known a random access memory, a hard drive a network card and a casing to make a complete server to serve as a node in a cluster to be used for either storage or compute purpose for traditional and/or Cloud service provider. It is less costly, and is replaceable, somewhat scalable for a certain time frame. (Obviously within a decade, there might be changes, which will make these current settings either legacy, still working or totally obsolete).

Figure 3. Big Data Business Intelligence Management {BDBIM}

These nodes can be added as an when needed to make users to take advantage of scalability for their data storage needs connected to a switch to transport this data to nodes and processed results back to user, which is also known as analytics or data mining. The throughput of data also can be increased, if data resides in same node, as being processed under a name node, or master node of the cluster managing any sorting process desired by the user or users of this Big Data storage and processing facility.

The combination of algorithms that deal with BDBIM (Big Data Business Intelligence Management) are the secret sauce to deal with Big Data Threat, RBAC, and Risk Analytics are major players in this solution. Due to the nature of data that we are working and dealing within the contemporary world of Big Data and Cloud, BDBIM will revolutionize the Analytics by providing eventual machine learning for web/cloud services to create decisions and in case of doubt consult the decisions for approval with Risk Analyst and Threat Analyst before taking any actions. These decisions will involve CIA (Confidentiality, Integrity & Authentication) in general but are not limited to:

- Fraud
  - Detection
  - Prevention
- Attack
  - Predictions
  - Detection
  - Prevention
- Known Threat Monitoring
- Unknown Threat
  - Detection
  - Prevention
- Identity Management
  - Role Based Access Control
    - Levels of Access
  - User Log Management
    - Login
    - Time Management
    - Logout

7. CONCLUSION AND FUTURE WORKS

We are witnessing a paradigm shift in Data Environment. In recent years, Big Data has risen on the technology horizons and is under the aspect of efficient and cost effective management and analysis of vast amounts of data for both public and private organizations. There are several organizations, which are trying to harness this continuing data stream, and in 2014, several of these organizations will go about making this data available in real time. Any organization, that want to take advantage, no matter they are automotive, financial or healthcare concerns, the latest developments in the field of in-memory data management are blurring between, computing algorithms over operational data and analytical data between datacenters and data warehouses, there will be any of the medium term otherwise will more likely be eliminated entirely .

The central concept to Big Data Architecture context is that data is either streaming in or some ETL processes are in progress with an organizational environment with which they have some sort of relationship in terms of trend production or some other analytical or other organizational needs, such as transaction processing on several diverse data types. The environment that an organization (directly or indirectly related stakeholders, who require or produce these data streams [both input and output, using Hadoop/MapReduce, Hive, Pig or some other such tool]) has an effect on how the end results will be produced.

8. ACKNOWLEDGMENTS

This work is supported by the Department of Computer Science and Mathematics at the University of Quebec at Chicoutimi, Quebec Canada.

9. REFERENCES

[1] ISO: ISO 15704:2000 Industrial automation systems - Requirements for enterprise-reference architectures and methodologies. Geneva, 2000.

[2] IEEE: IEEE Recommended Practice for Architectural Description of Software Intensive Systems (IEEE Std 1471-2000), 2000.

[3] Zachman, John A.: A Framework for Information Systems Architecture. *In IBM Systems Journal* 26 (3), pp. 276-292, 1987.

[4] Bosshammer, Manfred; Winter, Robert: Formal validation of compaction operations in conceptual data models - For a consistent compression of graphics and text documentation of the database business application systems using the example of SAP R / 3 PP and SD scheme. In: *King, Wolfgang (ed.): Economy computer science '95. Physica, Heidelberg*, pp. 223-241, 1995.

[5] S. Moore, Gartner forecasts global business intelligence market to grow 9.7 percent in 2011, *Gartner Research*, Sydney, Australia, Feb. 18 2011.

[6] Atif Farid Mohammad and Emanuel S. Grant. Cloud Computing, SaaS, and SOA 3.0: A New Frontier, *International Conference on Cloud Computing and Visualization 2010 (CCV2010)*, Prince George's Park, Singapore, May 2010.

[7] The Economist, Data, data everywhere, a *special report on managing information*, Feb. 27 2010.

[8] Palden Lama, Xiaobo Zhou. AROMA: automated resource allocation and configuration of mapreduce environment in the cloud. *ICAC '12: Proceedings of the 9th international conference on Autonomic computing*. September 2012.

[9] Di Xie, Ning Ding, Y. Charlie Hu, Ramana Kompella. The only constant is change: incorporating time-varying network reservations in data centers. *SIGCOMM Computer Communication Review*, Volume 42 Issue 4. September 2012.

[10] Chen, Peter Pin-Shan: The Entity-Relationship Model - Toward a Unified View of Data. In: *ACM Transactions on Database Systems 1* (1), pp. 9-36, 1976.

[11] Kemper, Alfons; Eickler, André: database systems an introduction. *6th edition*, Oldenbourg, Munich, 2006.

[12] Ferstl, O.K.; Sinz, E. J.: Fundamentals of computer science economy. *5th edition Munich et al.* , 2006.

[13] Atif Farid Mohammad, Emanuel Grant, Ronald Marsh, Scott Kerlin. Cloud Computing Monitoring Gateway for Secured Session Management of Big Data Analytic Sessions. *CGAT 201*3, Singapore April 2013

[14] Atif Farid Mohammad, Emanuel Grant, Scott Kerlin. Cloud Computing Monitoring Gateway for Big Data Secured Analysis using Live Signature – *TPALM. CGAT 2013*, Singapore April 2013