Reference Architecture for Big Data

Introducing a Model to Predict Current Affairs using Big Data Technology

Megha Bhandari, Smruthi D, Soumya V Bhat
VI sem, B.E. (Information Science)
AMC Engineering College
Bangalore- 560083, India

Bineet Kumar Jha
Asst. Professor, Department of ISE,
AMC Engineering College,
Bangalore-560083, India

***Abstract* – In this paper, processing of huge data with fast access and storage of data is done in such a way that data will be stored with minimal storage space. This paper provides an analytical solution for the organization with the help of pipe and filters and organization of different layers. The reference architecture uses the Hevner's framework for constructing the information system. The solution meets all its quality criteria based on research design in modularity, maintainability, reusability, performance and scalability.**

Keywords: Analytics, Map Reduce, Open Source Framework, Reusability.

I. INTRODUCTION

When dealing with larger datasets, organizations face difficulties in being able to create, manipulate, and manage big data. Huge amount of data produced by the business is very hard to handle as there was no essential tool or procedure to design to search and analyze massive datasets. .Massive amount of data are also produced by data sources such as social media networks, on-line news, open data, sensor data from the "internet of things", log tiles, email, video, sound, images, Which cannot be stored or processed using the traditional BI (business intelligence). Due to the evolution in the technology there are many open source software and frameworks available for the storing and accessing of huge amount of data. The technology provides a good opportunity for the organization that wants to make prediction about the future.

A. Big Data

Day by day the amount of data generated is growing rapidly. For example in 1990's the hard disk capacity was around 14GB-20GB, RAM capacity was around 64MB-128MB and reading capacity was around 10MBps, where as in 2014 the hard disk capacity is increased to 1TB,RAM capacity is increased to 4GB-16GB and the reading capacity is increased to 100MBps.The above example shows the rapid increment of data. The amount of information produced keeps on increasing has roughly doubled every 40 months since the 1980[7] as of 2012, every day 2.5 (2.5×1018) of data are created, which needs huge amount of space or database to be stored. The data beyond the storage capacity and processing power is called big data. These huge amounts of data are produced by various data sources like social media, hospital database, airline database etc. The Challenges included in these type of data are analysis, capture, data curation, search , sharing, storage, transfer, visualization, querying and privacy.

Therefore, the definition of big data according to Gartner is "High Volume, Velocity and/or Variety Information Assets that Demand Cost-Effective, Innovative Forms of Information Processing that Enables Enhanced Insight, Decision- Making, And Process Automation"[5]. Doug Laney introduced this "3V" definition already in 2001 [6]. IBS and IBM provide another definition: "Big data is a term associated with new types of workloads that cannot be easily supported in traditional environments", which indicates the switch from traditional BI to big data and the irrelativeness of the term [7]. Big data technology is to analyze the huge collection of data. A huge change came to the data world when hadoop ecosystem came into existents. Apache hadoop is a framework that allows for the distribution processing of large data sets across collection of computers using a simple programming language. It is an open source data management with huge data storage capacity and data processing. Some important features of hadoop framework are Reliable, Flexible, Economical, and Scalable.

The two main concept used in hadoop are map reduce and HDFS (Hadoop Distribution File System). The concept of map reduce was introduced by Dean and Ghemawat in 2004, map reducing is programming model for processing huge amount of data sets with a parallel, distribution algorithm.

B. Predictive Analysis of Big Data

Predictive analysis is a technique of extracting information from the given data to predict the future outcomes. Predictive analysis will not tell you what will happen in the future, it only tells what might happen in the future. The predictive analysis plays an vital role in big data .For example in an organization their huge amount of customer data are stored which can be used predict the future need of the customer

There are several .methods in which predictive analysis can be done, some of them are association rule learning, affinity analysis, and pattern recognition [13] .some of the organization has provided solution for predictive analysis with respect to big data. Some of the algorithms are K-means, decision trees, deep learning (multi-layered neural networks) and random forests (weighted multiple decision trees based on randomly selected sets of variables.

C. Reference Architectures

Summarizing the concepts used in the actual architecture is called as reference architecture. There are many types of reference architecture like enterprise reference architectures, solution reference architectures, information systems reference architectures, etc .There are many advantages of Reference architecture which can be useful for obtaining solution for the actual reference.

II. RESEARCH METHOD

Herner's Information Systems Research Framework was used for the creation of big data technology which is so popular currently that every life on earth , now either directly or indirectly comes across the name Big Data and runs all applications with the help of Big Data [17] .

Big Data is an information system observed in a scientific investigation that is not naturally present but occurs as a result of investigative procedure based on requirement of business and existing knowledge. This makes it perfectly suitable to design the structure of Big Data Solution Reference Architecture. Research Method involves other models: HP's Big Data Reference Architecture (BDRA) for the deployment of big data solutions, designed to improve the access to big data, deploy solutions and provide flexibility. Similarly, Kazman's Software Architecture Analysis Method (SAAM) and Angelov's framework for designing testimonial architecture. Hevner's model builds and justifies the latter two interpretations.

III. ANALYSIS AND EVALUATION OF BIG DATA

Initially, to learn the big data architectures, existing literatures were explored.Both scientific and non-scientific sources are used to review the development on big data architectures. Around 30 scientific and non-scientific resources were investigated for this purpose.

Big Data Architecture is an architecture domain that aims to address particular data problems and desires. Big Data solutions architects are trained to describe the organization and performance of the big data solutions. It helps in determining as to how big data solution can be delivered using big data technology such as Hadoop.

One must have proactive experience on Hadoop Applications. For instance, it may be configuring data, administration, debugging, monitoring, performance tuning etc.

A Big Data Solutions Architect generally should have a lot of experience about its architecture before moving on to big data solutions .One must have major ideas and experience with respect to Hadoop Map Reduce, Hive, HBase, MongoDB, Cassandra. Moderately, often they also need to experience in big data solutions like Impala, Oozie, Mahout, Flume, Zookeeper or Sqoop.

The components, architecture ideology, and best efforts found in literature were put ahead in the expert interviews to confirm their positions in the final model. In this way, the literature review in Hevner's framework served to prepare a provisional model of the final Big Data Solution Reference Architecture.

In the end, the big data solutions architect is responsible for the overall design and development of a vision that underlines a projected big data solution.

IV. DEVELOPMENT OF REFERENCE ARCHITECTURE

Angelov's framework focused on reference architecture. He created a model for reference architecture creation and distribution of reference architectures. His model consisted of dimension and each dimension contained a sub- dimension which has a question [18] associated with a sub-dimension. The procedure to obtain the results as follows:

A. Define WHY, WHEN AND WHERE

Angelov's described three W'S and they are as follows:

- Why - This describes the goal of the architecture
- Where - This describes about the operation
- When - This tells about the timing

The aim of the 3 W's is to advice the designer about the solution who can work effectively with big data. Angelov's defined two values for goal sub-dimension and they are standardization and facilitation. The main aim was to provide solutions to the problem and not to work with existing systems and to improve the sub systems and also to provide rules and regulations to solve the problem and also to encourage people who want are capable of solving big data problem. The reference architecture should be updated very often as the programs and appliances used will be outdated.

B. Invite stakeholders

Invite stakeholders who defines people who were involved in the interview about the big data and it also involved group decision regarding the solutions for the reference architecture. The context sub-dimension contains a list of stakeholders who are experts in the field of big data. It involved two groups and they are Requirement designers and

providers. They were piled up with a list of questions before leaving the room.

C. Reference Architecture

1) The Components (Modules) of the Model and Interfaces
2) The Existence of the Modules
3) The Picture of the Modules
4) The Characteristics of the Modules

To develop the reference architecture the developers developed many levels of grounded theory. Grounded theory is defined as the process that was derived from the specifications given and it was organized systematically.

Type 3 Reference architecture defines that it should contain modules, rules and regulations and interfaces .Based on these Angelov's developed a model which contained building design building ideas, building applications, modules .Angelov tells to measure the degree of detail by counting the number of modules. The Reference Architecture defines its modules, rules and regulations in a comprehensive way. The Big data Reference solution use the below given notation

1) Archimedes solution for modules and interfaces
2) TOGAF for building principles

V. BIG DATA SOLUTION REFERENCE ARCHITECTURE

It allows us to understand each and every module of the system in a better way .It was developed after studying the functions of each and every module, interviewing the professionals and by studying the effects of grounded theory. The Reference architecture describes about the rules and regulations. It contains three layers names viz. business layer, application layer and technology layer

1. The Architecture

The design itself was called as Archimate solution to Big Data solution Reference. This is constructed based on the study of the researcher and the effects of grounded theory. Hence the Big Data solution Reference architecture was divided into modules based on their appliances and programs capabilities. The design of the architecture was divided into 3 main layers such as business layer, technology layer and application layer. The layers focuses on what has to be done rather than how it should be done.

2. Architectural Pattern

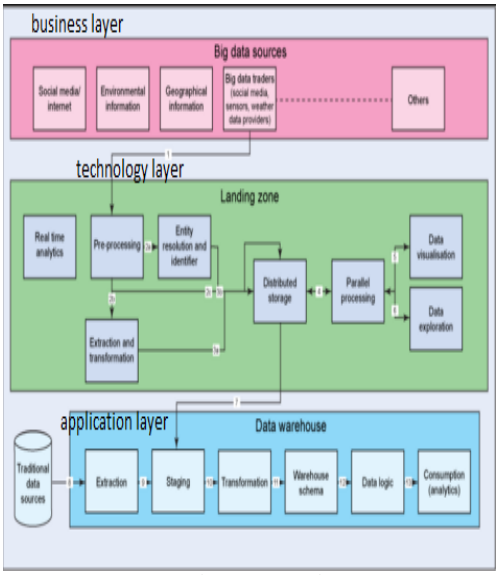

Fig. 1. Components of Big Data Reference Architecture

This brings a relation between the stakeholders, it defines how modules work and it gives solution to architectural problems and also defines how professionals should work with the modules. It involves three important aspects and they are layers, filters and pipes.

- Pipes and pattern filters.

It shows how sequence of data is transferred via modules. It shows the Big Data solution Reference architecture module should be built in a way that each module should carry out a sequence of task. Each module in the Big Data Reference Architecture carries task independently. Each module in a layer is considered as a filter .Each filter carries out a task not depending on another filter. The flow of the data is represented by a Flow Archimedes relation

- Layers

The architecture design shows that application, business and technology as layer. The layers works on the principle of loose coupling. Loose coupling is defined as the process in which two modules of the same layers are held together but they are independent of one another .For example: The layers pattern is implemented by showing each module in business application and technology layer.

3. Architecture Principles

The architecture principle is based on the two fundamentals and they are interoperability and loose coupling. Interoperability is defined as the process in which in two or more modules work together as a result of which there is no pressure on humans. Loose coupling-It defines the solution in such a way that 2 modules should be loosely coupled. Big

Data is a field where the modules need to be changed very frequently. The frequent change of modules is required and the modules need to adjust to such environment for this purpose we need to have a loosely coupled property which makes the Big Data field to be used in a very effective way. The modules should be selected in such a way that it should function properly and should have the capacity to overcome the situation.

- Interoperability - The programs and the appliances should work on specific standards. The standards should be specified in a way that it should satisfy all the user needs and also should satisfy the changing big data. The process which establishes standard should minimize the cost and it should help the different merchants to purchase the products.

The two best practices for Referential architecture are as follows:

1. Open source architecture: some of the advantages are given below:
   a. Security-the open source software provides the best security and correction and reduction of error can be done easily.
   b. Quality-The open source software provides software according to user's requirement which can also be manipulated by the user according to their needs.
   c. Customizability- As mentioned above the organization can copy the software and modify according to their needs. As it is an open source modifying code is very simple.
   d. Flexibility-The open software can run on any version of the system. There is no need to keep updating the software, updating can be done whenever user wishes to update.
   e. Audibility-The closed software cannot provide any guaranty of secure software as the source code is not available for the user. In open source this issue does not exist as the source code is visible, user can be sure of the software which he is using.
   f. Support options-open source software's have various support options like newsgroups, online chat, documentation, and mailing.
   g. Cost: For building new software the expenses will be more intend the organization can use inbuilt software which is less in cost.
2. *Agile development*

The solution for big data can be easily obtained by agile method. Hence the project manager should use both hardware and software interactively and make some changes to the obtained solutions. Some of the best examples for agile development are Scrum [50], Kanban [51], Lean [52], and XP [53].All the above software have one thing in common that is they are short iterative focusing on delivery and high quality solutions. Hence all organization should use agile method for the best solutions. Some of the disadvantages are testing is done throughout the life cycle, requirements are not sufficient, requirement keeps on increasing and frequent delivery is required.

VI. EVALUATION

The Big Data Solution Reference Architecture has been investigated for its quality by every 10 data experts out of number of 50. This investigation on questionnaire indicates that the solution meets all its quality criteria based on research design in modularity, maintainability, reusability, performance and scalability. The table given below is the average scores based on the quality criteria

Modularity	3.13
Reusability	3.00
Performance	2.75
Scalability	2.75

Overall, it signifies that the model can be qualified as'reasonably good' reference architecture. It will have its use in the architecture community, and perhaps be adopted once published.

VII.CONCLUSION

This paper performs research regarding solutions based on big data reference architecture. The qualitative data analysis and gounded theory are the two methods used for designing the reference architecture based on questionnaire with respect to quality criteria.

A. Observation and Calculation

With a group of big data experts the reference architecture result was justified and calculated .The Quality of the model was described using 5 criteria: Modularity, Maintainability, Scalability, Reusability and Performance.

REFERENCES

[1]. CSC,Big Data beginning to explode.
[2]. K.Ashton,That" IOT thing,"22 June 2009.
[3]. Lohr,The Origins of 'Big Data': An Etymological Detective Story," February 2013.
[4]. The Economist, Data,data everywhere," 25th February 2010.
[5]. Gartner, Big Data, Bigger Opportunities: Investing in Information and Analytics," 2013.
[6]. M.Ferguson, Architecting A Big Data Platform for Analytics Intelligent Business Strategies, Wilmslow, UK ,2012.