Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 3, March 2016, pg.755 – 760

An Exploration on Big Data Architecture, Challenges and Constraints

Dr. Vandana[1], Vaishali Balhara[2]
[1]Asstt. Prof., Comp. Sc. Dept., PT. N.R.S. Govt. College, MDU, Rohtak
[2]Asstt. Prof., PG Dept. of Computer Science, All India Jat Heroes Memorial College**,** MDU, Rohtak
[1] vandanamukeshmalik@gmail.com; [2] vaishali_leo@rediffmail.com

Abstract— The requirement of Bigdata in real and distributed environment is proven from last few years because of its global characterization. Bigdata able to represent large amount of data of different types and format in single or clustered form. It provides the corporated information and access and provide the collective data specific decision making capabilities. In this paper, the capabilities and features of Bigdata are discussed at early phase. Different applications, environment and processes applied on Bigdata are also discussed. As the method requires scientific tools, a methodology driven analysis is also provided in this paper. The paper also provided a framework to define different Bigdata layers extensively. The work also identified the various integrated challenges in Bigdata processing. The scope of this paper is to explore almost each aspect of Bigdata processing and management.

Keywords— Bigdata, KDD, Hadoop, HDFS, Challenges

I. INTRODUCTION

The merger and composition of different organizations, applications, domains and the technologies identified a new era of information representation. This knowledge representation is termed as Big-Data. In the complex hybrid environments, Bigdata[1][2][3][4][5][6] can be considered as well defined data term which can provide all kind of data and its processing. Big data supports the advanced technologies and architecture in single platform. Today each domain is actually reformed in terms of Bigdata. The bigdata is the new reform of traditional databases or warehouses with three integrated features represented by three-V, which is shown in figure 1.

First-V represents here volume which characterizes the size of data. Bigdata deals on bulk of information even in zetabytes, terabytes or gigabytes. The information can be of any organized or raw form can be considered as part of bigdata. The data need not to be defined at one place or in organized structural form. Any kind of related and valuable information can be combined to form the bigdata. The data volume must be able to perform the decision making, knowledge discovery and the optimization concept. The larger information size can be divided in smaller clustered structured form which can be processed individually, collectively or partially. This large scale data can be maintained over web in virtual and distributed environment which can be connected virtually to acquire the information gain.

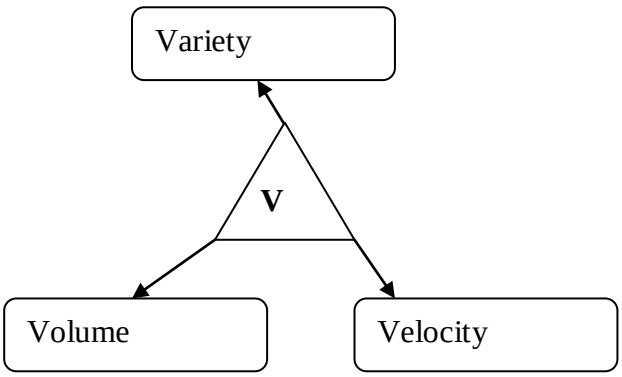

Figure 1 : Bigdata Features

Second-V here represents variety of data. A single object, aspect, operation or behavior can be described by different data forms. These data forms can be numerical, textual, image form, animated or the video form. The data can be in some organized, semi structured or unstructured form. It is not necessary to keep all the related data at one place, instead most of the time different data forms are maintained in different form. But there is some meta-file that can relate all the data forms based on the index specification. The variety of data increases the complexity in the data processing task. It means, there is no one such algorithm or approach that can process all the data forms and provide the associated solution in compact form. Different data forms are required to process under their own limitations and it is also required to combine solution obtained from each data form.

Third-V here represents the velocity or the dynamic nature of the data. The data is captured in real time environment. The timestamp based data is generally considered as part of bigdata. This kind data proves its relevancy but as the data is captured in real time, it is more complex. The complexity is about the organization of data, represented form and the inclusion of various impurities. It is required to process the data in different filtration stages to keep that in an organized form.

A) Bigdata Applications

The higher use and significance of big data is in analytical process to generate the hidden patterns and to explore the data relationship. There are number of available tools that provide data representation, visualization, and decision generation and information discovery with support of various phenomenons. The computational observation and its collaboration to real time applications is provided by associated generalized processes. Number of organizations[6][7][8][9] and companies provides the statistical support to big data in different form. Some of the common software systems provided by these companies are listed in table 1.

Table 1: Organizations support to Bigdata

Organizations	**Bigdata supported Software Systems**
Yahoo	Sherpa
Oracle	Oracle Bigdata Appliance
IBM	Hadoop, Infosphere
Cloudera	Cloudera Standard, CDH
Amazon	SimpleDB
Facebook	Cassandra
Microsoft	Dryad
ASF	CloudDB
Google	BigTable

These software worked for different application domains to process big data and to generate analytical results. One such emerging application includes genome data processing. This kind of big data having large data sequences which requires lot of mapping, sub sequence processing, analysis to identify the disease and the symptom based observations. Genomic sequences can be the collection of DNA or RNA based larger sequences which are encoded under some hybrid architecture so that the fast and the accurate decisions will be taken. To secure the information, some encoding measures can be applied at storage or communication time. These sequence processing is also associated to mobile devices also so that the quick and recent information can be processed. The health specification and mining applications uses these genetic information in various associated sub domains. Traffic information management and GPS data collection is also a larger application domain of big data. The automation of traffic lights, vehicle monitoring, automated parking etc are the common big data

applications in such domain. As the traffic communication is heavy with different observation including the vehicle type, frequency, accident case observation etc. These systems also use RFID method for data collection

B) KDD Process on Bigdata

Knowledge discovery[12][13][14][15] on large bigdata pool is defined as a set of structured operational activities to transform the complex data in simpler form and then to generate the required statistical and analytical results. Large volume of hybrid data extracted from various sources is processed under some strategic processes that can explore the challenges within application domain. The knowledge discovery process itself is formed under multiple process stages. At the earlier stage, the data extraction and storage description is defined. This stage includes the parallel data acquisition from various real time sources. Later on, a series of analytical processes with tool specification are applied to generate the meaningful underlying data. The process is defined based on the application and environment so that more significant and realistic decision will be obtained from the work.

II. BIGDATA TECHNIQUES AND TECHNOLOGIES

In this section, some of the major tools used at different stages of big data processing are described. These tools define the associated file system, distributed data management, storage processing, mining processing and the analytical characterization. These approaches and methods includes the intelligent processing in structured form along with technological reformation. The process description combines the statistical, mathematical and economical analysis so that the flexible and robust data processing will be formed. The work is also defined in different dimension with constraint specification so that the organized processing will be formed[1][5][6][7][8]. Some of such tools and technologies are described in this section

A) Hadoop

Hadoop[7][8][9][10] is the open source software project provided by Apache which provides distributed, robust, and reliable data processing. It is a complete web integrated tool that provides the distributed information in clustered formed method. The programming model is also associated to the computation independent modeling so that the large volume of data will be processed. Hadoop based file system is also derived to generate the large information storage and processing. This kind of storage system includes Map Reduce and Google File System (GFS). Hadoop also provide the application development environment with security and the sequencing features so that the distributed integrated communication will be formed. Hadoop not only describe the storage stage but also provides the sharing in public, private and collaborative domain[5][6][7].

B) HDFS

HDFS (Hadoop Distributed File System) is hardware specified distributed file system to provide large volume of data storage. The low level architecture is defined with different node specifications where each node is having specific storage capacity and bandwidth limit. As some data request is performed, the composition of data from these all nodes is done to complete the query. HDFS ensures the fault tolerant and reliable data storage and data processing. HDFS supports master-slave architecture for data storage. To store large information among, the data is divided in smaller segments and stores in separate nodes. The similar node groups in hadoop environment forms a cluster. The controller is defined to hold this data so that the machine will run effectively on available bandwidth and load. HDFS provides scalability for heavy load and heavy data processing[5][6][8].

C) Map Reduce

MapReduce[12][14][15] describes the network framework to store large volume of data in the form of distributed clusters. It defines a programming model to process larger datasets and to generate the analytical decisions. It is also responsible for generating the key/value pairs for reducing the function and merge them with intermediate key and value specifications. The processing in MapReduce environment is divided in two main stages. In first stage, the storage process is defined by applying the data partition and its distribution. This stage is controlled by master stage which identifies the number of partitions and the location for each partition. Master node also maintains an index table for mapping the data partition to the node. As some data query is performed, the data call is also applied to the complete group. The Map function is applied with query content specification to identify the data partition and relative node[5][6][7]9]. The data architecture of storage process is shown here in figure 2.

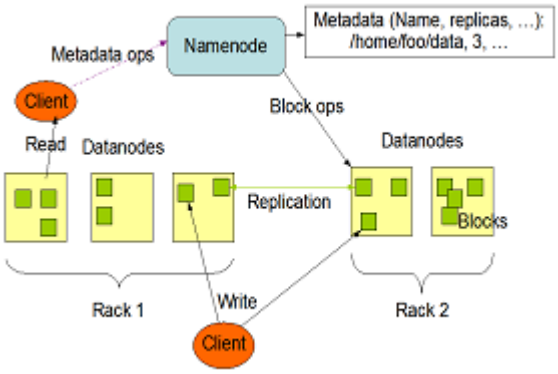

Figure 2 : File System Architecture

The figure signify that the cluster is formed using NameNode, with specification of meta data information. The master node manages the network data and regulates the access of files over the network to both clients and server. The data nodes are defines in each cluster for data storage. The block data information is defined to provide parallel and efficient processing so that large file processing will be done through block sequence processing.

III.BIG DATA PROCESSING FRAMEWORK

In this section, 3-Tier architecture[6][8][9][12][15] is defined to represent the conceptual view of structured process applied by mining processes. All the processes are described separately for each layer or tier. The process driven architecture is described in this section. This architecture covers all the complexities and challenges of large volume and hybrid data processing. The tool independent and technical aspects are defined with this architecture[6][7][9][10][12]. The architecture is based on the rules regulation with process behavior specification. The architecture is here shown in figure 3. Each of the integrated tiers of this architecture is described in this section.

A) Tier 1 : Data Access and Computing

It is the innermost tier which defines the physical characteristics of Big data architecture. The storage aspects and the relative distributed configuration with physical and conceptual constraints are defined in this stage. As described earlier, the larger data can be stored on different nodes of a single cluster. As some data query is performed on big data, it is required to load all data from various locations. This tier is responsible for all kind of data access and computation. The computation is here applied with resource specification and query dependent access. The major access resources are location, data and the processing unit. The scaled computation with memory load method is defined to achieve the parallel computing and programming method. It ensures the coordination between multiple servers, nodes and customers with specification of hidden and organized relationship on diverse data.

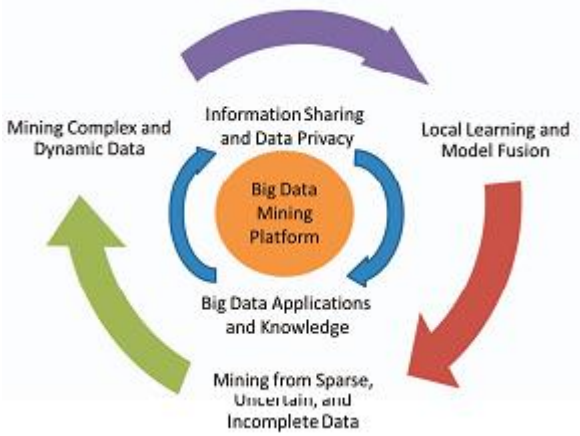

Figure 3 : Bigdata Processing Architecture

B) Tier II : Data Privacy and Domain Knowledge

Bigdata processing completely revolves around the application domain. If the domain information is not known, big data processing cannot be applied effectively. This domain information can be captured through meta-files to generate specific data segments. The domain information also includes the role specification of

different users and customers. The data requirements of each customer and query constraints respective to user are also defined. A semantic map on authorization and privacy map is defined in this tier to provide only the authenticated and authorized information. The privacy constraints with access control mechanism and certifications are defined. The inject access with privacy goal specification is defined to provide privileged information. This sharing and authorization is defined under communication control. The privacy can be defined at storage level or at access level or at communication level. The communication constraints such as bandwidth are also defined with this privacy vector.

C) Tier III : Data Mining Algorithms

Once the data is loaded after proving the user authentication and authorization, the next work is to process the user query on loaded data. This layer itself divided number of algorithmic mining processes. These integrated processes mainly divided in three sub processes. As the data is loaded, the pre-processing is applied to transform this raw data to normalized form. Data standardization for heterogeneous, uncertain and incomplete data is done by defining the constraints. These constraints can be domain specification or process specification. Filtration methods are applied in this stage by fusing different algorithms to generate normalized data form. In second stage, the dynamic and complex data is processed to identify the required information data. The segmentation process with rule specification is applied to identify the most required and valuable data. These whole processes adjust the processing model and parameter with process driven specification to extract the knowledge base. In the final stage, the local learning with fusion process is applied to generate the aggregative and analytical results. The results driven from different processes and from different segments can be combined to generate the final result. The featured statistical processing with correlation observations is applied in this stage to determine the relevant data patterns. The process mining is also robust to the environment, completeness and uncertainty and generates the analytical decisions. The stage emerge associated decisions are formed in this stage.

IV. BIGDATA CHALLENGES

As discussed in various section, Big data[6][7][8][9][12][14] works on real time and large volume data. To process and extract information from this large dataset, a complex architecture with multiple layers is defined. But this whole real time processing and analysis acquisition suffers from various associated issues[7][13][14][15][16][17][18]. Some of such critical issues are described in this section.

A) Security

The security is one of critical challenge that affects the information processing at different layers. According to the defined file system, the big data storage is performed on multiple nodes physically or logically. This segmented and distributed storage increases the data criticality and confidentiality. It is required to achieve the security for each node segment at storage level. The storage level security can be achieved by applying the encoding or cryptographic algorithms. But the management of this key sharing processes and index driven data storage need higher key-exchange aspects. The second level, security aspect is defined at the access level. In this stage, the customer and user authentication and authorization is defined. Personal information can be combined with actual data to provide the individual specific data access. Rules can be defined to improve the law enforcement that can increase the chances of information access or knowledge access relative to the adverse consequence. The final level security is defined at the communication level. In this stage, the encoded data communication and local decoding process can be defined. All these security levels provide the secure information storage, processing and delivery.

B) Analytical Issues

As the large and hybrid data is being process, there is the requirement to provide a quality analysis process. Different objective and constraints are setup during this analysis. The requirement is to the algorithm process most adaptable to the application, domain and the process. The selection of number of cluster and node point is also challenging. The bandwidth sharing and the user access restriction are required to set. The composition and localization of different data form must be organized. The data forms can be structured, unstructured and semi-structured. The skills are required to process the data and to provide the application dependent decision making. The integration constraints are also required to setup.

C) Fault Tolerance

The complexity at different Bigdata layers also increases the challenges with technology specifications. Data access is here defined with failure and damage tolerant computation. The failure probability based success information access in cloud environment is defined. The cost observation and checkpoint specification and interval driven access. The recursive method is defined to achieve effective and reliable computation.

D) Heterogeneous Data

Data level criticality in big data system can be identified as availability of real time capturing of data. This data can exist in different formats including structured data form, semi-structured and unstructured data form. The challenge is about to define different algorithmic process for different data form. This processing can be done at storage time at access time. The transformation can also applied to convert each data form to normalized form so that single algorithm can be applied for all data forms.

V. CONCLUSIONS

The paper has provided a clear characterization of various aspects of Bigdata. These aspects include the technology and tool exploration as well as the challenges faces in bigdata processing. The application driven framework is discussed in detail which specification of various associated tiers. Each isolated tier is been explained here in detail.

REFERENCES

[1] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Zomaya and S. Foufou, and A. Bouras, "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis", Trasactions on Emerging Topics in Computing, pp 1-12, 2014

[2] Hua Fang, Zhaoyang Zhang, Chanpaul Jin Wang, Mahmoud Daneshmand, Chonggang Wang, and Honggang Wang, "A Survey of Big Data Research", IEEE Network, pp 6-9, 2015

[3] Matturdi Bardi, Zhou Xianwei, LI Shaui and Lin Fuhong, "Big Data Security and Privacy: A Review", China Commpunication, Vol 2, pp 135-145,2014

[4] Edmon Begoli, James Horey, "Design Principles for Effective Knowledge Discovery from Big Data", Joint Working Conference on Software Architecture & 6th European Conference on Software Architecture, pp 215-218, 2012

[5] Aditya B. Patel, Manashvi Birla and Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce", Nirma University International Conference on Engineering, NUiCONE 2012, pp 1-5, 2012

[6] Dan Garlasu, Virginia Sandulescu, Ionela Halcu, Giorgian Neculoiu, Oana Grigoriu, Mariana Marinescu and Viorel Marinescu, "A Big Data implementation based on Grid Computing", IEEE Conference, pp 1-4, 2012

[7] Marcus R. Wigan and Roger Clarke, "Big Data's Big Unintended Consequences", IEEE Computer Society, pp 46-53, 2013

[8] Xin Luna Dong and Divesh Srivastava, "Big Data Integration", ICDE Conference, pp 1245-1248, 2013

[9] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding, "Data Mining with Big Data", IEEE Transactions on Knowledge and Data Engineering, pp 1-26, 2013

[10] Seref Sagiroglu and Daygu Sinanc, "Big Data: A Review", IEEE Conference, pp 42-47, 2013

[11] Yuri Demchenko, Paola Grosso, Cees de Laat and Peter Membrey, "Addressing Big Data Issues in Scientific Data Infrastructure", IEEE Conference, pp 48-57, 2013

[12] Antonia Azzini and Paolo Ceravolo, "Consistent Process Mining Over Big Data Triple Stores", IEEE International Congress on Big Data, pp 54-61, 2013

[13] Zibin Zheng, Jieming Zhu, and Michael R. Lyu, "Service-generated Big Data and Big Data-as-a-Service: An Overview", IEEE International Congress on Big Data, IEEE Computer Society, pp 403-410, 2013

[14] Weiqiang Sun, Fengqin Li, Wei Guo, Yaohui Jin and Weisheng Hu, "Store, Schedule and Switch – A New Data Delivery Model in the Big Data Era", ICTON, pp 1-4, 2013

[15] Hyoung Woo Park, Yeon Yeo, Jongsuk Ruth Lee and Haengjin Jang, "Study on big data center traffic management based on the seperation of large-scale data stream", Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, CPS Computer Scoeity, pp 591-593, 2013

[16] Avita Katal, Mohammad Wazid and R H Goudar, "Big Data: Issues, Challenges, Tools and Good Practice", IEEE Conference pp 404-409, 2013

[17] Du Zhang, "Inconsistencies in Big Data", Proc. 12th IEEE Int. Conf. on Cognitive Informatics & Cognitive Comp, pp 61-67, 2013

[18] Samet Ayhan, Johnathan Pesce, Paul Comitz and David Sweet, Steve Bliesner, and Gary Gerberick, "Predictive Analytics with Aviation Big Data", IEEE Conference, pp 1-13, 2013. 311-320 vol.1.

