Preetishree Patnaik al.International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 2, Issue 2, June 2015, pp. 35-41

An Emerging 3-Tier Architecture model and frameworks for Big Data Analytics

Preetishree Patnaik[1], Pooja Batra Nagpal[2], Ulya Sabeel[3]

[1]Research Scholar, Computer Science Engineering, Amity University, Haryana.
[2]Asst. professors, Computer Science Engineering, Amity University, Haryana.
[3]Asst. professors, Computer Science Engineering, Amity University, Haryana.

Abstract- **Big Data is a term for enormous data sets having larger, more diverse and complex structure that creates difficulties in many data processing activities such as storing, analyze and visualizing processes or results. This process of Research into huge amounts of data to reveal unseen patterns and secret correlations named as Big Data Analytics. In comparison to past decades the primary IT industry has challenged for most organizations was enabling more and faster transactions for business productivity. Today, in the age of the Internet, every user focus on faster delivery of data information (e.g., mails video, audio, medical images, movies, gene sequences, sensor data streams) to systems, PCs, smart phones, tablets and etc.**

Big Data analysis has emphasized the biggest challenges that the IT Industry faced from Big Data and optimizes the best way of finding solution to analyze, monetize, and capitalize the use of Big Data in Business Intelligence. With such revolutionary changes the Data Scientist called it as "the era of Big Data.

In this paper we have proposed a novel 3-tier architecture model for Big Data in Data Mining techniques and rest of paper explains regarding the challenges that Big Data faced today and Big Data security Analytics, including the Big Data Framework models, Components of framework, attributes and Study of application of Big Data mining in the field of Cluster Analysis, and Pattern Evaluation.

Keywords— Big Data, Big Data Analytics, Architecture in Big Data, Big Data Mining

I. INTRODUCTION

The term Big Data encompasses of all forms of data, including Web logs, data from social networking sites, sensor data, tweets, blogs, user reviews, and SMS messages. Big data and big data analytics are in the recent study of information technology and business intelligence. These data are generated from various social networking sites like Facebooks, twitter, etc ,online transactions, emails, videos, audios, images, click streams, logs, posts, search queries, health records, science data, sensors Smart phones and their applications [1]. These data are in different format, hence required for databases to store and analyze the data sets and visualize via typical database software tools.

In comparison to past decades the primary IT Industry has changes a lot, with more fast transaction people are accessing huge amount of data in various pattern e.g. Internet mails, video, images ,audio messages ,sensors data streams and etc with such huge accessibility of data makes a revolutionary change in analysis of data streams patterns . Thus the Data Scientists has announced that we are now in the "Era of Big data" or we are sinking to deep water of big data every day.

Today, we accepted ourselves in the era of digitalization with gigantic progress, development of technologies, web media, social networking sites, online world technologies through internet, Smartphone.etc where every user are Accessing enormous, massive quantities of data from various data sources. Such enormous data sets having massive,

Diverse and complex structure of data is term as "***Big Data***". These massive data creates a lot of difficulties in storing, analyzing, searching and visualization process. But we know that this massive volume of data sets can be useful to user in various aspects and creates lots of confusion in its storing and analyzing. Therefore ,a big massive of data sets(BIG DATA) are need to be store in effective and efficient manner that helps in various type of operations(i.e. analytical operation, process operations, retrieval, reliability of data & etc).

Thus it is most important to execution of these massive data sets into a secrete correlation/pattern/cluster models that makes easy to extract all relevant information from the complex data, such process of extraction of hidden data called as "Big Data Mining" which help in utilization through implementation various types of clustering techniques, Data mining methods. In the below section we have explained the categorization of Big Data, Proposed Framework Model and a 3-tier Architecture Model for Big Data.

II. CATEGORIZATION OF BIG DATA

The huge volume of data (Big Data) is used in various user applications but on other hand it a lot of problematic in storing and analyzing. Therefore, a big volume of data or big data has its own deficiencies. They need big storages and this volume makes operations such as analytical operations, process operations, retrieval operations, very difficult and hugely time consuming. One way to overcome these difficult problems is to have big data clustered in a compact format that is still an informative version of the entire data. Such clustering techniques aim to produce a good quality of clusters/summaries. Therefore, they would hugely benefits everyone from ordinary users to researchers and people in the corporate world, as they could provide an efficient tool that helps with large data such as critical systems to detect cyber-attacks. The following points describe the details of Big Data Categorization:

- **Variety:** Big data come from a great range of sources and a further volume of data source. The data source includes different data format style such as: structured, semi structured and unstructured.

a) Structured Data: - The structures data are organized manner easily sorted to store in database .these variety Data include the abstract data type, web links, pointers etc.

b) Unstructured Data: - The unstructured data are random and difficult to analyze. These are Heterogeneous and raw/incomplete data that are generated from multiple users in different sources. (e.g.: Bitmap images, objects, text, etc).

c) Semi- structured data: - These are the combination of structure and un-structured data and doesn't conforms to a fixed set of tags or others semantics structure of data.

- **Volume:** Volume or the size of data has been larger than terabytes and petabytes. The grand scale and rise of data outstrips fixed store and analysis technique. As the Big data size is massive and huge in nature, so it's a biggest challenge for the data scientist to design the large database for its effective storage and visualization. [1]
- **Velocity:** The range of data used is in max range, Velocity is a necessary parameter not only for big data, but also all processes. For time limited processes to be executed, big data used should be in organization streams to have a maximize value [1]
- **Veracity:** These types of data are generally uncertainty due to inconsistency and ambiguities latency.

The below figure 1: give details the categorization of Big Data:

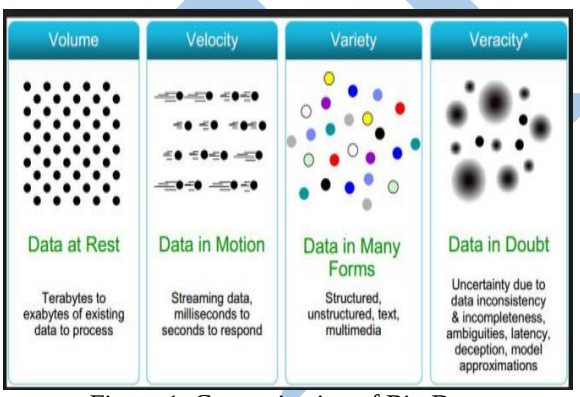

Figure 1: Categorization of Big Data

III. BIG DATA SECURITY ANALYTICS

Big data security analytics is simply a collection of security data sets so large and complex that it becomes difficult (or impossible) to process using on-hand database management tools or traditional security data processing applications. Many enterprise organizations are still struggling to maintain the big data security issue.

Big data security analytics solutions also distinguish themselves based upon three basic characteristics:

- **Scale.** Big data security analytics solutions must have the ability to collect, process, and store terabytes to petabytes of data from an assortment of security analytics activities.
- **Analytical flexibility.** Big data security analytics solutions must provide users with the ability to interact, query, and visualize this volume of data in an assortment of ways.
- **Performance.** Big data security analytics must be built with appropriate computer architecture to process data analytic algorithms and complex Queries and then deliver results in an acceptable timeframe.

Characteristics of Big Data security Analytics:

- The following point depicts the various characteristic features of Big Data Analytics. There are several important factors, but the Single biggest challenge today is finding talent to help program and to manage Big Data are:
- **Sharing and Privacy***:* There are several different integration models. Some user had studied that while providing the security analytics are creating partial copies of data stored in more traditional data mining systems, with the assumption that lower cost commodity storage makes the iterative cost trivial. Whereas some derive data from their existing clusters and import that information into used by Hadoop or their security system. As per studies by SEMI solution in Data privacy and security it is observed that more critical or sensitive data is often made more secure to store in cluster within a cluster using data mining techniques giving the current idea for big data security analytics.
- **Data Encryption***:* This is an important feature to make the big Data more secured to access only with the administrator access rights .it has recommend file/OS level encryption because it scales as you add nodes and is transparent to Nosql operations.
- **Authentication and Authorization***:* To ensure that secure administrative passwords are in place and those application users must authenticate before gaining access to the cluster. Each user has a different type of accessing password (e.g. Developer, user, and administrator roles should all be segregated).
- **Node Authentication:** There is little protection from adding unwanted nodes and applications to a big data cluster, especially in cloud and virtual environments where it is trivial to copy a machine image and start a new instance. Tools like Kerberos help to ensure rogue nodes don't issue queries or receive copies of the data.
- **Key Management:** Data encryption is most important as a key security; so any external key management system is to have secure keys and, if possible, help validate key usage.
- **Logging:** Logging is built into Hadoop and any other clusters. It seems to provide the security to all other network devices and applications and recommend that user built-in logging, or leverage one of the many open-source or commercial logging tools to capture a subset of system events.

- **Network Protocol Security:** SSL or TLS is built-in or available on most Nosql distributions. If privacy is at all important, look to implement protocol security to keep your data private.

IV. COMMON TECHNOLOGIES USED IN BIG DATA

After The unbeaten growth in the data accessibility by users has produced tremendous flow of data streams. This tremendous increase in data accessibility and usage creates problem for processing, analyzing correct information. The Big Data Analytics services not only focus to store and handle large volume of data but it also works for data processing, Analytical study and visualization. There are different types of Big Data technologies used in architecture models which gives an optimize results for many the real time application services.

- **Hadoop**

The Hadoop is a java based framework used in Big Data that helps to run application on system with 1000 of nodes and terabytes of data files i.e. Hadoop can run multiple number of nodes (application) and terabytes of the data can be transferred among the users.
It has own distributed file system that work to allow the system file transfer continuously without any node failure. These types of applications help to reduce the risk of catastrophic system failure in which application is broken down into smaller parts called as fragments or blocks.
The apache Hadoop consists of the following components:

a) Hadoop kernel
b) Hadoop distributed file system (HDFS).
c) MapReduce.
d) HBase.

Further the HDFS associated with three components such as Name node, Secondary Name node and data node.
Hadoop associated with security proposed protocol used as enhanced Linux (SE Linux) for solving multi level secure (MLS) environmental problems in Hadoop. In such environment problems the Hadoop run with multiple sources at different levels and protocols used as extension version of HDFS.
Hadoop is commonly used for distributed batch index building and optimize the index capability in real time applications. Beside this it also supports storage and analysis for large scale processing.
Advantages: Hadoop is distributing storage and computation capabilities, highly scalable, optimized for high throughput, large Block sizes, tolerant to software and hardware failure.
Disadvantages: it is master process are single point failure Hadoop does not offer storage or network level encryption, inefficient for handling secure files.
In the following section give the details of the Hadoop main components in above framework model:

a) **HBase**: it is an open source platform, distributed and non-relational DB system implemented in java. It run above the layer of HDFS which is mostly used for the input and output for MapReduce in will manned structure.
b) **Sqoop**: Sqoop is a command line interface application which provides a platform used for data conversion from RDBMS and Hadoop vice versa.

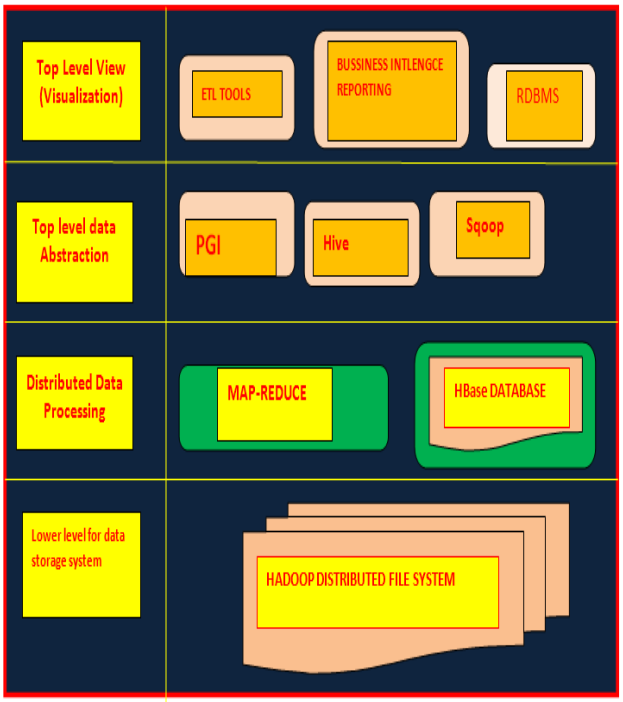

Figure 2: Hadoop Framework model

c) **PIG**: pig is a high-level platform where the MapReduce framework is created which is used with Hadoop platform. Pig is a high level data processing system where the data records are analyzed that occurs in high level language.
d) **Hive**: It is application developed for data warehouse that provides the SQL interface as well relation DB model. Hive infrastructure is built on the top layer of Hadoop so as to help in queries analysis.

- **MapReduce:**

Initially MapReduce was developed by Google with main objective to process and store large datasets on commodity hardware applications. MapReduce model processing is used for the large scale data records into a cluster forms.
The MapReduce is mainly associated with 2 program sets such as: Map( ) function and Reduce( ) function. Users can stimulate their own processing logics having well defined above functions.
Map( ) function : It performs the task as the master node takes the input, divide into smaller sub modules and distributed into salve node. Where a SLAVE node further divides the base problem and passes the result back to master node.
The MapReduce system arrange together all intermediate pairs based on the key values and send a request to Reduce ( ) function for producing the final result.
Reduce( ) Function: The Reduce function act as master node collect all result from sub problems and combine them to give the final result.
MapReduce framework is based on master-salve architecture model is describe below where one master-node handle a number of salve nodes.
Process: first divide the input data into even size equal data blocks for same load distribution. Each data block is then

assigned to one SALVE node and is processed by a map task and result is generated.
The SALVE node is interrupts the master node when it is idle state. The scheduler assign task to the slave node.
In the above architecture model for MapReduce always allows a local data block to a slave node. If the effort failed, the scheduler will assign a rack-local or random data block to salve node instead of local data block.
When the Map ( ) function complete its task, the runtime system gathers all intermediate pairs and launches a set of condense task to produce the final output.
The main drawback in MapReduce large scale data processing is a difficult task, for managing 100-1000 of processors, parallelization and distributed environment. This issue can be solved by the process of Input-output scheduling parallel processing system. It is fault tolerance and supports scalability inbuilt processes for status and monitoring heterogeneous and large datasets.
Using MapReduce framework the efficiency and time to retrieve the data is quite optimize to address the volume aspect new techniques have been proposed to enable parallel processing using above framework.
MapReduce Components:

a) Name Node: it manages HDFS Metadata, doesn't deals with data files directly.
b) Data Node: Stores blocks of HDFS default replication for each block.
c) Job Tracker: Scheduler allocates and monitor; job execution on salves i.e. task tracker.
d) Task Tracker: run MapReduce operations.

- **Hive:**

Hive is a distributed agent platform a decentralized system for building application by networking local system resources. Apache Hive data warehouse component, an element of cloud based Hadoop ecosystem which offers a query ecosystem which offers a query language called HiveQL that translate SQL like queries into MapReduce jobs automatically.
Applications of Apache hive are SQL, Oracle, and IBM DB2. The working procedure in Apache divided into MapReduce oriented execution, Meta data information for data storage and execute on part that received a query from user or application for any query execution.
Advantage: Hive is more secure and implementation to used the data files.
Disadvantages: Hive as it only works for adhoc queries and performance is less compared.

- **Nosql:**

Nosql database is an approach to data management and main application to handle large distributes data. It is also known as Not-Sql database that provides more significant and growing industry use in Big Data and real time application. It also supports the mechanism to storage and retrieval of data that is modeled to store the data in tabular form like RDBMS.
Nosql includes simple horizontal scaling and fine control availability of huge data. The data structure used informs of document type. The performance of Nosql is faster than RDBMS. The Nosql simple allows executing the sql query statement and provides the consistency (CAP Theorem) availability of Big Data and partition tolerance of Database

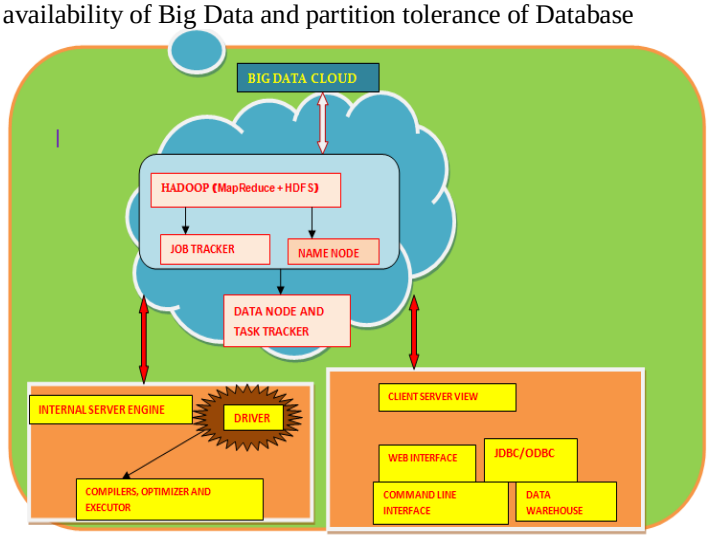

Figure 3: Architecture model for Hive

V. PROPOSED FRAME MODEL FOR BIG DATA IN DATA MINNG

In compression to past decades the IT industry has reckon and the data volume accessibility exceeds the capacity of current online storage and other processing system. The users are now allowed to access massive amount of data and the usage has extended from a range of Exabyte per year to Zettabytes per year. It has been observed that storage and data transport are such technology issue which is need to be focused inorder to optimize the existing problem. Today data scientist has announced that we all are living in the "Era of Digitization" and every fraction of seconds we are sinking to a deep sea of Big Data. Such exponential growth in Big Data cause various problem in storing, sorting, searching, Data Analysis and different hidden correlation data patterns makes the job data analytics as quite difficult.
The persistence growth of computational data analysis has produced a tremendous flow of data streams as compared to past records. This tremendous increased in data accessibility and usage introduce many problem into field of Big Data. Such as the data are now represented in complex, multi-structured, unstructured, heterogeneous format and these data are mostly generated by different data sources. The major issue focuses as data processing, analyzing correct information from massive datasets, handling huge customer data in Big Data warehouse, etc.
Thus inorder to overcome with Big Data issues and challenges that are observing in recent data analysis work can be resolved by the following points:

i. Design an appropriate system to handle the data efficiently.
ii. Investigate and identify the issues that are associated in Big Data storage, management and processing.
iii. Analyze the data using various data mining techniques, KDD that will help to evaluate the various hidden data format and visualization techniques.
iv. Extract the relevant information from the massive datasets using Grid Based clustering techniques in data

mining helps in good pattern evaluation of big data and decision making system.

In the following section we have proposed a novel framework for Big Data Mining using the Grid based clustering techniques. The below figure 4 depicts that the process of execution of Big Data in the field of data mining which helps to overcome with the existing challenges and issues i.e. storing, sorting, searching, analyzing, and visualization into different data patterns .

The Big Data represents as an enormous large complex data structure that creates difficulty in data processing and visualization data pattern. This complex structure of massive data has emerged from a Big Data cloud which includes the data in various format and different source for data extraction such as social networking sites, internet application, network data, software devices data, and Smartphone mobile data.

Thus inorder to analyze and store such massive data we need to adopt clustering techniques which give an optimize results in data integration, data analytics and data visualization. Further we need a large big data warehouse to store the optimize data. The warehouses possess the characteristics of subject-oriented, non-volatile, integrate and time variant relationships.

These data supports in various types of decision support system with the sole concept of analyzing data structure format, handling complex data for storage, frequent pattern analysis, evaluating marketing strategy and converting the traditional data warehouse into Big Data Warehouse system including the following features:

- **Preprocessing**:
  This features is executed with a decision making system to produce optimize result with understanding concepts "what data should be stored in data warehouse".
- **Offloading:**

This feature includes the latest trends of the database storage system in Big Data such as Hadoop, MapReduce, Nosql, etc.

- **Exploration:**
  This feature gives the idea for the future analysis of data patterns and discovering various new data mining concepts and techniques.
- **Big Data mining**:
  The term Big Data mining can be defined as a process of extraction of important information from hidden data into a secrete correlation pattern.

The important feature of Big Data mining as:

a) Big Data application in various knowledge development processes.
b) Mining the uncertainty and incomplete data.
c) Mining the complex and dynamic data.
d) Analyzing new hidden data patterns
e) Visualization of data. Further the process of Big Data in data mining helps to gather all important relevant data information into a secrete correlation patterns. These massive data are generates from different data warehouse and data source and need to be analysis.

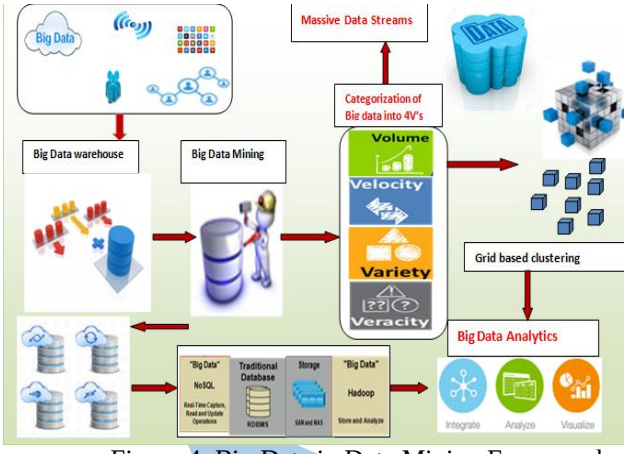

Figure 4: Big Data in Data Mining Framework

In this paper we have proposed the a Nobel framework for Big Data Analytics in Data mining with the sole objective for finding hidden data patterns using different types of cluster algorithms. We have verified that the Grid based clustering Techniques is most well efficient in working on Big Data Analytics. This application of grid Based clustering Algorithm helps in analyzing and visualization of various data pattern with fast processing time.

Thus we studied that the important feature of Big Data mining are as follows:

a) Big Data application in various knowledge development processes.
b) Mining the uncertainty and incomplete data.
c) Mining the complex and dynamic data.
d) Analyzing new hidden data patterns
e) Visualization of data.

VI. A 3-TIER ARCHITECTURE MODEL FOR BIG DATA IN DATA MINING

In general the main issue and challenges exist with Big Data as to handle the massive, unstructured, complex representation of data volume. In proposed framework for Big Data Mining which we have explained the concepts of Big Data Analytical processing for handling the massive data warehouse data, different data accessibility with use in various data dimension techniques and data computation using the grid based clustering algorithm.

In this section we have given a 3-tier Architecture model for Big Data and explained in details under the 3 level of its working.

- **Tier-1: Big Data mining:**

The tier-1 represented the fist layer in the architecture model which is common working principles associated with the accessing of large volume of data and different modes of data computation process. Here the Big data is showed in form of cloud and this Big Data Cloud is generally formed with the collection of different type of complex data such as real time data, networking data and etc. Again the main concern of the first layer is to handle the Big Data storage and management issue exist with the different types of the Data dimension i.e. Data Volume, Variety, Velocity and Veracity. The exponential growth in massive, unstructured heterogeneous complex structure of data representation Big Data creates a lot of

problem in storing, sorting, and visualization pattern of massive data structure. These Big Data required many number of data warehouse to store the large volume of data and helps the user to access data in different format beside this user are also allowed to access and maintain the data semantics , data integrity.

But in some cases the data integrity is difficult to maintain in Big Data warehouses which are located in different location. So inorder to store the massive volume of data here we have used the concept of Big Data warehouse and further to study and analysis the different data patterns we have applied different types of clustering algorithms and proceeding to the next tier.

- **Tier-2: clustering Analysis:**

The tier-2 is associated types of clustering techniques that are applied into Big Data inorder to solve the issue existing with the sorting, analyzing and visualization patterns. We defined the clustering algorithms with the different have emerged as an alternative powerful and meta-learning tool helps to analyze the massive volume of data (Big Data) generated by many application. The user must choose a suitable Clustering Algorithm which can be applied to Big Data volume and different types of distribution pattern in data sets. In general Big Data are combined with different autonomous source, aggregate distributed data source to centralized site system which impacts more on the data structure and semantics.

- **Tier-3: Big Data Analytics and Visualization:**

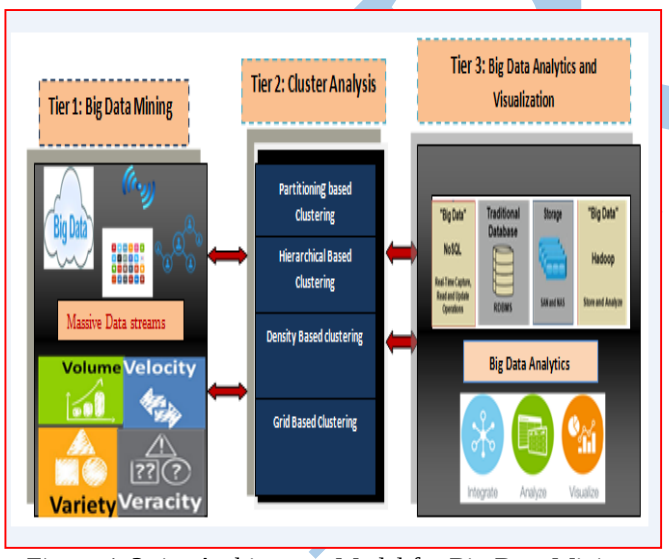

Figure 4: 3-tier Architecture Model for Big Data Mining.

The tier-3 is associated with the output layer for the data processing system. Big Data support various types of database system and storage facilities which is explained in details in section .IV. These layer work for the data visualization of the complex data that has extracted by the Big Data Mining process and data analysis done by using various type of software tools. The figure4: gives the details view of 3-tier architecture model in Big Data Mining

VII. CONCLUSION AND FUTURE WORK

In this paper we have given some important emerging framework model design for Big Data Analytics and a 3-tier architecture model for Big Data in Data Mining. In the proposed 3-tier architecture model is more scalable in working with different environment and also benefits to overcome with the main issue in Big Data Analytics for storing, Analyzing, and visualization. The framework model given for Hadoop HDFS distributed data storage, real-time Nosql databases, and MapReduce distributed data processing over a cluster of commodity servers.

We conclude that the Big Data is intrinsically related to the open source software revolution. Large companies such as Facebook, Yahoo!, Twitter, LinkedIn benefits and contribute to open source projects. In future we will be working for Distributed mining process to work with different distributed versions of some methods, for which a lot of research is needed with practical and theoretical analysis for such Methods.

REFERENCES

[1]. Big Data: A Review by Seref SAGIROGLU and Duygu SINANC Gazi University,IEEE 2014.

[2]. The Critical Role of the Network in Big Data Applications Sponsored by: Cisco Systems Lucinda Borovick Richard L. Villars April 2012

[3]. Oracle NoSQL Database and Fusion's ioDrive2 Offer Cost-effective, Extreme Performance for Big Data Environments.WWW.FUSIONIO.COM.

[4]. C. Eaton, D. Deroos, T. Deutsch, G. Lapis and P.C. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, Mc Graw-Hil.

[5]. Yuri Demchenko "The Big Data Architecture Framework (BDAF)" Outcome of the Brainstorming Session at the University of Amsterdam 17 July 2013. [2] Tekiner F. and Keane J.A., Systems, Man and Cybernetics (SMC), "Big Data Framework" 2013 IEEE International Conference on 13–16 Oct. 2013, 1494–1499.

[6]. Dong, X.L.; Srivastava, D. Data Engineering (ICDE)," Big data integration" IEEE International Conference on , 29(2013) 1245–1248.

[7]. Jian Tan; Shicong Meng; Xiaoqiao Meng; Li ZhangINFOCOM, "Improving ReduceTask data locality for sequential MapReduce" 2013 Proceedings IEEE ,1627 - 1635 [7] Yaxiong Zhao; Jie Wu INFOCOM, "Dache: A Data Aware Caching for Big-Data Applications Using the MapReduce Framework" 2013 Proceedings IEEE 2013, 35 - 39 (Volume 19)

[8]. Sagiroglu, S.; Sinanc, D.,"Big Data: A Review",2013,20-24

[9]. Minar, N.; Gray, M.; Roup, O.; Krikorian, R.; Maes, "Hive: distributed agents for networking things" IEEE CONFERENCE PUBLICATIONS 1999 (118-129) [10] Garlasu, D.; Sandulescu, V.; Halcu, I.; Neculoiu, G,"A Big Data implementation based on Grid Computing", Grid Computing, 2013, 17-19

[10]. Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce", 2012, 6-8

[11]. Jefry Dean and Sanjay Ghemwat, MapReduce:A Flexible Data Processing Tool, Communications of the ACM, Volume 53, Issuse.1,2010, 72-77.

[12]. Defintinig Big Data Architetcure Framework: Outcome of the Brainstorming Session at the University of Amsterdam, 17 July 2013. Presented at NBD-WG, 24 July 2013 [Online]. Available: http://bigdatawg.nist.gov/_uploadfiles/M0055_v1_7606723276.pdf .

[13]. Haixun Wang , Wei Wang, Jiong Yang ,Philip S. Yu, "Clustering by Pattern Similarity in Large Data Sets".

[14]. R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Authomatic subspace clustering of high dimensional data for data mining applications. In SIGMOD, 1998.

[15]. Seref SAGIROGLU and Duygu SINANC Gazi University Department of Computer Engineering, Faculty of Engineering Ankara, Turkey ss@gazi.edu.tr, duygusinanc@gazi.edu.tr . Big Data , A survey

[16]. JERZY STEFANOWSKI Institute of Computing Sciences Poznan University of Technology Poznan, Poland Lecture 7 SE Master Course 2008/2009. Data Mining clustering techniques lectures .

[17]. XiaoCai,FeipingNie,HengHuang∗ University of Texas at Arlington Arlington, Texas, 76092 xiao.cai@mavs.uta.edu, feipingnie@gmail.com, heng@uta.edu- Multi-View K-means clusteringonBigData

[18]. Future Wei Fan Huawei Noah's Ark Lab Hong Kong Science Park Shatin, Hong Kong david.fanwei@huawei.com Albert Bifet Yahoo! Research Barcelona Av. Diagonal 177 Barcelona, Catalonia, Spain abifet@yahoo-inc.com . Mining Big Data: Current Status, and Forecast to the Future.

