Scalable and Flexible Big Data Analytic Framework (SFBAF) For Big Data Processing and Knowledge Extraction

R. Siva Ram Prasad[1], and Chittineni Aruna[2]

***Abstract*----**World Digitalization floods massive amounts of structured and unstructured data through different systems and applications known as "Big Data" today. Social networks, server logs, transactions, search engines, ERP applications, PDAs, sensors and other devices are continuously generating huge amount of structured, semi-structured and unstructured data as a part of their job. This data become more precious than any other and used to assess the future business requirements through analytics. This is a very big opportunity for business entrepreneurs to extract the great knowledge from Big Data to improve their decision making accuracy over emerging business trends. Recent researches were introduced many Big Data analytic mechanisms to analyze the huge unstructured data in terms of searching for valuable business information. Henceforth the data volume and data format inconvenience previous analytics were having scalability and load balancing issues over Big Data processing. In this paper, we introduced Scalable and Flexible Big Data Analytic Framework (SFBAF) architecture for Big Data processing and knowledge extraction. Implementation of this framework addresses the problems of data extraction, data transformation, mapping, analytics performance and result accuracy at design level. Case studies will explain and compare the results of our approach with existing analytic mechanisms in detail.

***Keywords*---** Big Data, data transformation, data analytics, map reduce, data mining

I. Introduction

TODAY Social networks, Search engines, Blogs, Sensors, weblogs, Sensex, E-Commerce and some other applications are generating large volumes of data (zeta bytes) in terms of their routine job. This huge amount of data considered as Big Data[1,2] is different from traditional data in terms of volume, velocity and variety. Volume specifies the amount of data (data size), velocity specifies the data generation rate and variety stands for data structure in general. As per the updated statistics 90 percent of world data generated in last two years [3] and Face book, Twitter, Google, Yahoo, Linked in, Amazon are the companies processing peta bytes of data daily. Face book data growth is increasing 500 TB every day and wall mart has to handle 1 million transactions per hour and the generated log data is equals to 2.5 peta bytes in their databases [4]. Today this digital world is having more than 2.7 zeta bytes of data and this will reach approximately 4.2 zeta bytes by the end of 2016.Instead of batch data generation, today devices like sensors, cameras, weblogs and other forecasting devices are continuously streaming the data as a flood. The current web 2.0 is completely data driven environment [5] composed with the applications like social networks, E-Commerce, search engines, storage clouds and virtual machines as shown in figure.1.Date Importance, increased storage and process capacities and digitalization are also the main causes to form Big Data.

The Big Data is generating from miscellaneous sources is having different formats like structured, semi structured and unstructured data [6].Unstructured data has been occupied the major part of world's complete data approximately 80% and continuing the raise. Mining this Big Data is very important for every organization to extract the valuable information, which is useful to improve the business statistics and to meet the business trends of today. For example Bank authorities has to analyze their transaction logs to implement the timeline graph for transaction load, E-Commerce websites has to analyze their customer forum discussions to mine the customer interests, satisfaction, behavior and recommendations, an advertising and public relations company needs to process the millions of online customer profiles to find target audience, sentiment analysis, customer experience, advertisement priority and mapping etc.

Big Data should be analyzed to extract the valuable information for business analysis and forecasting. Previous researches were introduced various architectures to accomplish the process of Big Data analysis and knowledge extraction by using traditional data analytics [7]. Existing Big Data process architectures and methodologies were not implemented to handle different kinds of data with variety. Process should be different from one data type to other to achieve the accuracy and reliability. In this paper we are introducing an advanced analytic architecture as Scalable and Flexible Big Data Analytic Framework (SFBAF) for Big Data process and pattern discovery. This is an extension to existing traditional analytic frameworks like Apache Hadoop [8], HP

R. SIVA RAM PRASAD[1], Research Director, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.

CHITTINENI ARUNA[2], Associate Professor, Department of CSE, KKR & KSR Institute of Technology and Sciences and Research Scholar, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.

Veratica [9] etc. Our architecture implements the data variety respective data processing tools and methodologies to avoid the problems of existing approaches and to improve the scalability and flexibility in architecture design. Case studies will discuss about the advantages of our architecture against existing technologies.

The remaining chapters are organized as follows. Chapter 2 presents related work to summarize past work, including traditional vs Big Data analytics, Big Data analytics usage. Chapter 3 introduces the proposed Scalable and Flexible Big Data Analytic Framework (SFBAF) architecture and implementation. Case studies on SFBAF over other methodologies are presented in Chapter 4. Conclusion and future works are defined in Chapter5.

Fig. 1 Data streaming from various resources as Big Data

II. Related Work

Traditional data analytics versus Big Data analytics: In general, the traditional enterprise structure datais typically stored in a data ware house and processed through structured query language (SQL) [10]. Business Intelligence (BI) tools [11] were introduced later to get the benefits from business data analysis in the form of reports and dashboards. This traditional data is well-formed structured data from different data bases to data ware house through ETL process. Data analytics were designed software tools to extract the business intelligence from raw business data to achieve scalability, consistency, reliability and accessibility. Data analytics can process this structured data by using the interrelationships of data fields and schema information from ware house by using SQL. In this case data analytics were designed for stable centralized ware house environment to mine structured data to extract expected reports and information. If the data analytic servers are at remote location from ware house, than the date need to be transforming from ware house to analytic servers for processing.

Big Data is totally different from traditional data in the way of volume, velocity and variety. Uncontrollable massive amounts (Terabytes to Petabytes and more)of complex unstructured data is generating at each second from various sources like mails, blogs, social networks, news, photos, videos etc. Handling this unstructured data is not possible with traditional data bases and data analytics due to process complexities, hidden correlations, unknown data format and structure. Unstructured data couldn't have any inter relationships explicitly among data values, henceforth designing the data analytics for Big Data is very tedious than relational data analytics. Today making use of these massive amounts of data is very important for business intelligence and become challenging for organizations.

Big Data Analytics usage: The Big Data generating from organizations should be processed to extract the business intelligence and fraud detection, which could have an impact on business performance. In order to process these petabytes of data we need robust analytics with impressive processing structure and algorithms. There are different types of analytics based on target data type like social media data analytics, mail data analytics, video data analytics, structured data analytics, unstructured data analytics and Hybrid data analytics etc. Report generation is an important aspect for business management to identify their performance and growth rates against time intervals. Business analytics [13] for structured data can generate these reports based on existing data from organizations. These reports and indicators mirror the past activities and their results in a streamlined manner. Present business trends are expecting current and future analysis for business improvements by mining the stream of unstructured data using analytics. Henceforth today organizations need the real-time predicative and prescriptive analytics for Big Data mining [12]. These analytics should understand the existing massive resources of data, searching for correlations, navigating and discovering the required patterns to retrieve the intelligence from data sources.

III. Scalable And Flexible Big Data Analytic Framework

Big Data analytics are designed to collect, transform, analyze and extract the business intelligence by discovering useful patterns against Big Data. Recently organizations were started using these analytics to process their Big Data for current business analysis and future forecasting also. This Big Data processing helps the organization to know sentiment analysis, fraud detection, and sales improvement, risk management, customer voice and satisfaction etc. Data variety (extracting from different resources) characteristic of Big Data causes to fail traditional business intelligence tools to process Big Data. Big Data analytics plays a vital role in terms of processing Big Data, henceforth we have to always use only the efficient pattern recovery programs and algorithms along with sophisticated architecture. Enterprises are looking for advanced analytics over traditional to perform "what-if" analysis [14] against multi-dimensional databases for business planning and investment area forecasting.

Data available in the format of Big Data is a collection of different interrelated data types with veracity. The term veracity of Big Data stands for different sources, different formats and different types of data. Here the different sources means OLTP databases, NO SQL data bases, Mail system data bases, social network databases, server logs, sensors, call records, different files etc. Structured data, semi-structured and unstructured data are the basic three different formats of data. Real Big Data types are text, audio, video images etc. Traditional Big Data architecture is using the same mechanism to process the data with veracity. Transformation, analysis, pattern mining, knowledge extraction and result representation is different from one data format to other. This design level drawback at Big Data analytics architecture effects on the scalability, reliability and flexibility.

We introduced Scalable and Flexible Big Data Analytic Framework (SFBAF) architecture for Big Data processing and knowledge extraction as shown in figure 2. Implementation of this framework addresses the problems of data extraction, data transformation, mapping, analytics performance and result accuracy at design level. In English there is an idiom says that "different strokes for different folks", which is best suitable and adoptable for data mining as "different methodologies for different formats of data". The way to handle and process structured data is totally different form unstructured data. In the same manner dealing with mail data is different from dealing with social network data in terms of reliability.

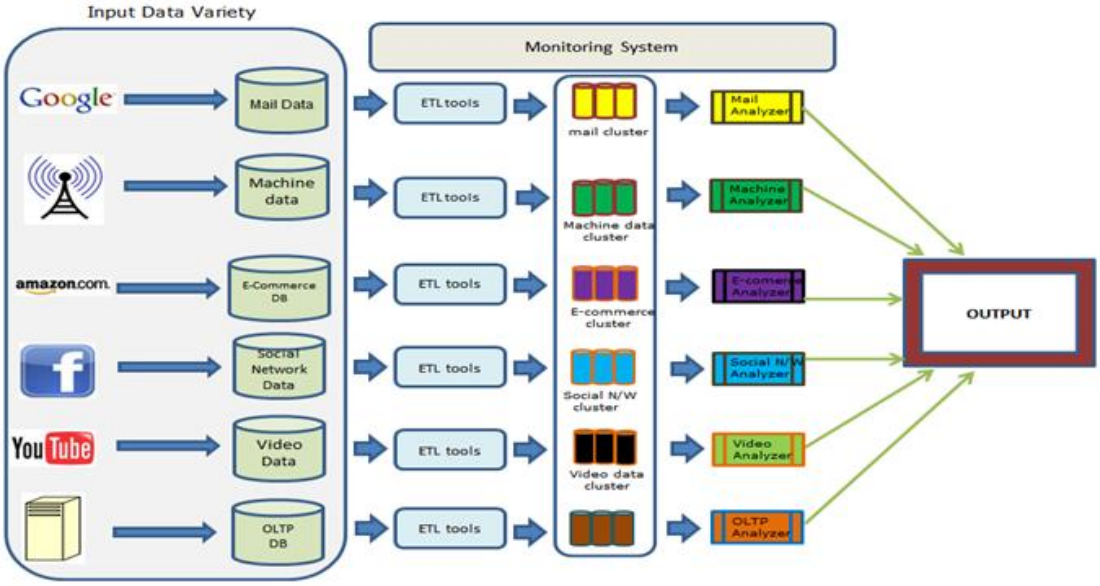

Fig. 2 Proposed SFBAF Framework

Mail data is semi structured, which is having more accuracy and reliability than social network data. Henceforth processing methodology should be different from data type to data type and format also. This mechanism causes little overhead than traditional mechanism but the scalability and flexibility will be higher than traditional one. The above architecture represents our proposed SFBAF for Big Data analytics implementation. This architecture is having Input Data Varity module, ETL tools, Data storage with clusters, Big Data Analyzers, output and monitoring system.

Data Variety Module: this is a collaboration of different data sources and their data storage units. Prominent data sources are search engines, click streams, sensors, call recorders, E-Commerce, social networks, online video warehouses, servers, transaction logs etc. This module data will be uploaded to Big Data storage units through ETL tools for preprocessing and stemming.

ETL tools: In this module the preferred ETL tools are data dependents and process dependents. ETL tools play a vital role in this architecture to achieve the scalability and flexibility. Data preprocessing is the major activity of this module. Our approach is using the respective ETL tools for extraction transform and loading operations. While transforming the data our ETL tolls will convert the data into flexible data structures. This transformation helps to improve the process accuracy and scalability. Once this process completes ETL tools will load the data to Big Data storage area (HBase [15]).

Clustering Big Data: Transformed data loaded to Big Data storage units by ETL tools after completion of preprocessing. These Big Data storage units are collection of traditional systems or any cloud storage units. In order to group the relevant data here we are using robust KNN clustering algorithm, which finds the correlations among the data values to simplify the process of data clustering. After completion of the correlation finding the relevant data will be formed as clusters. Here we are using a different data storage unit per data format means all social networks data will be stored to a separate unit instead of a single unit for process flexibility.

Big Data Analyzers: After completion of data clustering analyzers will process the data to discover the patterns against

the clustered data. Like hadoop analyzers they also implement the shuffle and sort process for results extraction. Our analyzers are unique at every data type and format to improve the result accuracy by identifying best pattern. Our research strongly believes that single comprehensive analyzer is having the less accuracy than multiple analyzers while finding the match against the data. After completion of the analysis the result will be stored to result data storage region and displayed through retrieval window.

Monitoring System: Monitoring system behaves as a admin module to monitor the data processing mechanism. This system is having some sub modules like resource management system, task manager, connectivity manager and data administration etc. These sub modules helps to monitor the status of executing task and interact with the possible input values.

IV. Case Studies In Detail

This section discuss about our new proposals and the advantages of them over other traditional approaches.

1. Need of data a separation for Big Data mining?

As we discussed in the above section "different strokes for different folks", like that we need different process mechanisms for different types of data. Henceforth data is available in various formats and processing of one type is different from others, we need data separation at initial stage while loading. Data is generating from different sources in various formats is called as data variety. For example if we want find an E-Commerce website related discussion we need to collect the data from same website logs and feedbacks, social networks, Blogs, Mails, Forums etc. if we consider the data format of website feedback, mails is semi-structure and social networks, blogs having unstructured data. In this case the data is relevant in terms of topic but not format. If we mix this structured data and unstructured data that will create process handling problems in future while processing. So our architecture is supporting data separation for Big Data mining.

2. What are the Advantages of Separate ETL tools?

Data extraction, transform and loading are the major concepts of data processing done by ETL tools. Regardless of extraction and loading the other concept transform is little different depends on data type and format. Transforming performs the data cleansing first and later converts the data into the respective data structures. Cleansing is different from structured and unstructured data, however if the data is structured than we need to trim the data spaces but for unstructured data stemming is required. Similarly data storage structures also different among text, video and image data. Different ETL tools will be mapped to respective types at runtime to perform cleansing the data. Traditional Big Data approaches are having the performance problem due to improper data cleansing. Cleansed data makes feasible the burden of pattern discovery and process complexity in Big Data mining.

3. Describe the data dependent clustering and analyzers?

Clustering of our approach is different from other traditional approaches in the way of relevance and variety based clustering. Traditional Big Data mechanism clusters different types of data based on topic relevance and positive ration calculation. Clustering different types leads to ambiguity and error prone of data results. To avoid this problem we are clustering the social network data separate from mail data in a single storage environment. This process helps the result customization and accuracy improvement. Analyzers also different in this approach as we know different data types need the respective analyzers. The training data and process mechanism of unstructured data is different from semi-structured data. Similarly analyzing the text, image and video data is also different in their own way. So we implemented respective data analyzers for the recommended data type and format.

4. How do support this architecture is scalable and flexible?

For example if we want to analyze the next day performance of a stock exchange with Big Data analytics, we need to analyze the data different sources of stock exchange discussion forums. Manually this is very complex but become feasible when approaching Big Data analytics. In general this information is available from social networks, blogs, expert's discussion forums and Sensex logs. Traditional Big Data tools will integrate the data from all resources before processing and extract the knowledge also in same manner. But in above condition the accuracy levels of expert's discussion forum data and Sensex logs will be high than social networks. We can trust the Sensex logs and experts discussion than any other due to their long study and experience. So we have to measure the accuracy not always with help of topic relevance but also with source priority in terms of reliability. This example recommends us to rank the results based on data source priority also. In order to implement the result accuracy based on source priority we need to process each data source results as individual as we implemented. This mechanism supports the high accuracy ratio at data source level improves scalabilityand separate process ETL tools, clustering and analyzers made this approach flexible for data processing in Big Data environment.

V. Conclusion And Future Work

In this paper we concentrated on the issues of Big Data analytics architecture especially to improve the performance in terms of scalability and flexibility. Traditional Big Data architecture is using the same mechanism to process the data with veracity. Transformation, analysis, pattern mining, knowledge extraction and result representation is different from one data format to other. This design level drawback at Big Data analytics architecture effects on the scalability, reliability and flexibility. We introduced Scalable and Flexible Big Data Analytic Framework (SFBAF) architecture for Big Data processing and knowledge extraction. Implementation of this framework addresses the problems of

data extraction, data transformation, mapping, analytics performance and result accuracy at design level. Case study confirms the advantage of our approach over other existing Big Data analytics along with vision and pace of our proposed architecture.

In future we want merge this mechanism with cloud to get the cloud benefits like pay-per-use, on-demand, portability, security and extensibility. This approach reduces the infra management burden and storage and process cost also.

REFERENCES

[1] J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and SmartAssets: Ten Tech-Enabled Business Trends to Watch.McKinSeyQuarterly, 2010.
[2] J. Dittrich, J.-A. Quian´e-Ruiz, A. Jindal, Y. Kargin, V. Setty,and J. Schad. Hadoop++: Making a Yellow Elephant RunLike a Cheetah. PVLDB, 3(1), 2010.
[3] http://www.sciencedaily.com/releases/2013/05 /13052 2085217.htm.
[4] http://wikibon.org/blog/big-data-statistics/
[5] J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and SmartAssets: Ten Tech-Enabled Business Trends to Watch.McKinSeyQuarterly, 2010.
[6] S. Das, Y. Sismanis, K.S. Beyer, R. Gemulla, P.J. Haas, and J.McPherson, "Ricardo: Integrating R and Hadoop,"Proc. ACMSIGMOD Int'l Conf. Management Data (SIGMOD '10),pp. 987-998.2010. http://dx.doi.org/10.1145/1807167.1807275
[7] A. Ghoting and E. Pednault, "Hadoop-ML: An Infrastructure forthe Rapid Implementation of Parallel Reusable Analytics," Proc.Large-Scale Machine Learning: Parallelism and Massive Data SetsWorkshop (NIPS '09),2009.
[8] http://hadoop.apache.org/docs/r1.2.1/mapred_ tutorial.html
[9] http://www.vertica.com/documentation/hp-vertica-analytics-platform-7-0-x-productdocumentation/
[10]R. Cattell. Scalable SQL and NoSQL data stores. SIGMODRec., 39:12–27, May 2011. http://dx.doi.org/10.1145/1978915.1978919
[11]Kiron, David, Rebecca Shockley, Nina Kruschwitz, Glenn Finch and Dr. Michael Haydock, "Analytics: The widening divide: How companies are achieving competitive advantage through analytics"IBM&MIT Management Review. October 2010.
[12]J. Zhao, J. Wu, X. Feng, H. Xiong, and K. Xu, "InformationPropagation in Online Social Networks: A Tie-Strength Perspective,"Knowledge and Information Systems,vol. 32, no. 3, pp. 589-608,Sept. 2012. http://dx.doi.org/10.1007/s10115-011-0445-x
[13] A. Ghoting and E. Pednault, "Hadoop-ML: An Infrastructure forthe Rapid Implementation of Parallel Reusable Analytics," Proc.Large-Scale Machine Learning: Parallelism and Massive Data SetsWorkshop (NIPS '09),2009.
[14] "Big Data Analytics: Advanced Analytics in Oracle Database" An Oracle White Paper, March 2013.
[15] http://hbase.apache.org/book/quickstart.html