A Survey on Data Analytics Framework

M.Sharmila Begum[1] and A.George[2]
[1] Research Scholar, Department of Computer Science and Engineering
Periyar Maniammai University,Vallam
[1]sharmilagaji@gmail.com
[2]Professor, School of Humanities Science and Management
Periyar Maniammai University,Vallam
[2]amalanathangeorge@gmail.com

***Abstract* - In recent years, much interest in developing Data analytics tools has been mainly intense on exploring and understanding massive data formats in social media and growing internet. Such data are a typical form of multimedia data that comprise text, images, videos, and networks. In this survey work, we present a comprehensive overview of the Data analytics tools, their merits and challenges. A taxonomy of prior studies is introduced to classify the state-of-the-art techniques into two categories, namely, Data analytics for gathering information and understanding user behaviors. The proposed taxonomy can provide a coherent vocabulary for researchers to share knowledge and simplify analysis tasks. The Data analytics is rapidly developing with numerous new methods emerging every year. However, the area is still in its infancy with many challenges and open questions. Many of the challenges cannot be addressed using techniques from only one discipline. We believe that multi-disciplinary research that combines Dataization, multimedia, NLP, and human computer interaction will lead to more powerful and enabling approaches and technologies to handle and understand data.**

Keywords – Big data analytic, Tools, Business Intelligence (BI)and Frameworks

I. INTRODUCTION

Managing and analyzing challenges data have for organizations always offered benefits across all industries. The biggest challenges for today's technology lies in the increasing prevalence of data. This is frequently referred to as "big data" [1]. Never before in the history of humankind, have we been able to generate a living history of ourselves. In the process, we are creating new data of immense size and scope. It is indeed a trans formative change to see that within a few decades we have moved from complaining about the lack of data to a data deluge.[2] This makes data analytics even more exciting and indispensable.

Big data analytics is the process of examining large data sets containing a variety of data types – i.e. ,big data to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits. The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI) programs. That could include Web server logs and Internet click stream data, social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things. Some people exclusively associate big data with semi-structured and unstructured data of that sort. The International Data Corporation (IDC) research forecasts that overall details will develop by 50 times by 2020, motivated mainly by more included systems such as receptors in outfits, medical gadgets and components like structures and connects. The research also identified that unstructured details - such as data files, email and video - will account for 90% of all details designed over the next several years. But the number of IT experts available to handle all that details will only develop by 1.5 times the present levels.

II. BIG DATA ANALYSIS PREREQUISITE

Big data frequently involves a development of cultural and technical changes throughout a business or provide new business opportunities to expanding the sphere of inquiry to exploit new acumen's to combine both traditional and big data analytics. It involves making "sense" out of huge volumes of mixed data that in its raw form absences a data model to define what each element means in the context of the others[3]. There are some new issues like discovery, iteration, flexible capacity, mining and predicting or decision management. For ex. many organizations have no idea whether or not data huts light on sales trends or challenge comes with figuring out which data element relate to other data elements, and in what data storage capacity. This process of discovery not only involves sightseeing the data to understand how to use it but also to grasp its relation to traditional database.

Big data analyzing requirements are some of the methods that can be used to find meaning and discover unseen relationships in big data. There are three major significant requirements:

A. *Minimizing data movement* is basically all about conversing computing resources. In traditional analysis scenarios, data is brought into the computer system, processed and then sent to next destination. For Ex. Filpkart sales data might be extracted from e-system, transformed into a relational data type, and loaded into an operational data store structured for generating report. The volume of data increases every time, this type of ETL (Extract, Transform and Load) architecture becomes increasingly less well-organized. It makes more intelligence to store and process the data in the same place.

B. *Using existing skills* means new data or new data sources demands the need to obtain new skills. Most of the time, existing skill sets will determine where analysis can and should be done. Mostly organizations have more and more people who can analyze data using either SQL or MapReduce, it is very important to be able to support both type of processing.

C. *Attending data security* is essential part for many business applications. Basically data warehouse users are habituated not only with judgment defined metrics, dimensions and attributes, but also to a reliable set of management policies and security controls. These rigorous processes are often poor with unstructured data source and open source analysis tools.

III. BIG DATA ANALYZING FRAMEWORK APPROACHES

With the advancement of the frameworks, handling and analyzing big data within a reasonable time has become less difficult. Since the foundation functions of the frameworks to handle and manage the big data were developed gradually, the data scientists nowadays do not have to take care of everything, from the raw data gathering to data analysis by them. They use the existing platforms or technologies to handle and manage the data. The data scientists nowadays can pay more attention to find out the useful information from the data even though this task is typically like looking for a needle in a haystack.

Big data analyzing tools have four key approaches to analyze data and generate analysis data or reports.

A. Discovery tools: This tool is very useful throughout the information lifecycle for intuitive exploration and analysis of information from any combination of structured, unstructured and semi – structured source. These tools permit analysis alongside traditional BI source systems.

b. *BI (Business Intelligence) tools*: This tool is very useful for analyzing & reporting and also for performance management, primarily with transactional data form data warehouse and information systems. With help of BI tools users can draw new reports, come to meaningful or useful conclusion, and make informed decisions quickly.

c. *In-Database Analytics*: These analytical techniques that allow data processing are applied directly within the database. In-Database Analytics include a variety of technique like credit scoring, fraud detection, trends or finding patterns and relationship in data.

d. *Decision Management*: Decision Management consists of predictive modeling, self-learning, and business rules to take informed action based on the existing context. Mostly this type of analysis allows distinct recommendations across multiple channels, maximizing the value of every client to communication.

All of these methodologies have a role to play uncovering hidden relationships. Most of the organization provide analyzing tools and framework which are available to explore big data and are capable to perform analytics. We look at these all tools and frameworks provided by various organization.

IV. BIG DATA ANALYZING TOOLS CHARACTERISTICS

A Big Data platform should give a solution which is designed specifically with the needs of the enterprise in the mind. The following are the basic features of Big Data Platforms- Comprehensive, Enterprise-ready, Incorporated, Open Source based, Low latency flows and updates, Solid and fault-tolerant, Scalability, Extensible, Analysis Web Interface, Hadoop support, Spark support, Allows ADHOC queries, minimal maintenance; etc.

V. ENTERPRISE BIG DATA ANALYZING TOOLS

A. Citus DB

Citus DB scalable and robust analytical tool is built on top of PostgreSQL. Citus DB is designed having parallelism in mind. Citus DB is the first database analytical tool which enables execution of distributed SQL queries on data which is external to the database. Citus DB gives fast and flexible access to massive volumes of data. Event streams, user actions, machine generated data and log files are applicable datasets. Citus DB partitions these massive data and executes queries efficiently on these that involve groupings, look-ups, orderings, and complex selections. Also, Citus DB supports JOIN between multiple small tables and one large table. Citus DB enables real-time responsiveness. For simple queries, run time needed is around 100ms, and increases depending on dataset size and query complexity. Real time insertion and deletion is not available in Citus DB and also it does not support real time analytics.

B. GoogleBigQuery

Google BigQuery uses SQL to analyze big data and gives real time business insights in seconds. It uses a managed data analysis service without the need for server installation or maintenance.

The following are the *features* of Google BigQuer;

Managing data – refers to creation and deletion of tables and is based on a JSON encoded schema and data is imported from Google Storage.

Query – In BigQuery queries are expressed in a Structured Query Language dialect and the results of length around 64MB are returned in JSON. There are some limitations to the usual Structured Query Language queries.

Integration –It is easy to integrate BigQuery with Google Spreadsheets and Google Apps Script.

Access Control – It is done in BigQuery via Google Storage.

C. Greenplum HD

Greenplum HD allows customers to start with big data statistics without the need to develop an entire new venture. It is provided as application or can be used in a preconfigured Data Handling Equipment Component. Greenplum is a 100% open-source qualified and reinforced edition of the Apache Hadoop collection that contains HDFS, Pig, MapReduce, Hbase, Zookeeper and Hive. IT prevails of a finish information research foundation and it brings together Hadoop and Greenplum data resource into only one Data Handling Equipment. Available as application or in a preconfigured Data Handling Equipment Component, Greenplum HD provides a finish foundation, such as set up, training, international support beyond simple appearance of the Apache Hadoop submission. Greenplum HD makes Hadoop quicker, more reliable, and easier to use. We can quickly set up extensive big data statistics remedy using HDFS that brings together Greenplum HD and Isilon scale-out NAS storage systems to provide, extremely effective and versatile information storage and statistics environment.

The Greenplum HD DCA Component easily combines the Greenplum HD application into a product, offering an enhanced setting designed for performance and stability. The Greenplum Data Handling Equipment has power of Hadoop batch-processing to process unstructured data. This allows businesses to draw out value from both arranged and unstructured data under only one, smooth foundation.

D. Hadapt

Hadapt's flagship product is an Adaptive Analytical tool, which brings a standard implementation of Structured Query Language (SQL) to the open source Apache Hadoop project. Hadapt enables interactive SQL based analytics of large data sets, by combining the scalable and robust architectures of Hadoop into a hybrid storage layer. Hadapt 2.0 delivers Hadapt Interactive Query, Hadapt Development Kit for custom analytics and integration with Tableau Software using Apache Hadoop.

E. IBM InfoSphere BigInsights

IBM InfoSphere BigInsights is a Big Data analytics platform provided by IBM, which support not the same type of analytics under one roof. InfoSphere is built on top of Hadoop to improve its capabilities and provides an interactive UI on it for analyzing the big data. This has built-in analytics capability, with social data analyzer for analyzing social media data, text analytics for in receipt of insights from hug amount textual data, machine data analytics for analyzing machine data like sensors and GPS related data and InfoSphere for the integration with other big data technologies. InfoSphere provides a SQL interface namely, BigSQL, Jaql a Declarative query language and spreadsheet interface called Bigsheets for all tools analyzing and exploring big data easily. BigSQL, Jaql and Bigsheets modules of InfoSphere are some of the core components of the system. BigSQL offer the facility to user to sightsee the database schema and analysis the data using structured query language. Jaql is a Declarative query language to facilitate analyzing of structured, unstructured and semi-structured data. BigSheets a web-based analysis and visualization tool using familiar, excel spreadsheet interface that enable to analyze huge amount of data and long running data collection jobs. This system can upload data

from multiple sources such as web crawlers, database, text files, Json, csv files etc. and can store it in the distributed file system for processing.

F. SAP Big Data Analytics

SAP is one of the most leading provider of big data analytics platforms. SAP is one of the 1st company to introduce in-memory database for analytics. This aims to make Event Stream Processor (EPS) an excellent tool in SAP HANA (High Performance Analytic Appliance). The build in memory database was to provide a single database for both transactional and analytical data processing, mostly refer to as Online Transaction Processing and Online Analytical Processing systems. SAP's ESP is a standalone, overall streaming analytics platform that has a long, rich history as one of the original complex event processing. It has broad base of customer's services in telecommunication, financial, energy, retails, manufacturing, transportation and logistics, and all public or private sectors. SAP provide significant road map item for EPS to integrate it as a service that can run within its SAP HANA in-memory database. This will afford streaming analytics capabilities to SAP HANA's strong analytics capabilities.

G. TERADATA Big Data Analytics

The Aster large data analytics appliance is solution from TERADATA for big data analytics. Aster is essentially a database developed by TERADATA supporting row and column based storage. It's the key element of their big data analytical platform which contains of Aster SQL-MapReduce, Aster Database, which is mainly an interfacing Structure Query Language for Aster Database. It is a combination of hardware and software by connecting several nodes with Infniband in its place of running hadoop on commodity hardware with traditional network connections.

H. Cloudera Big Data Solution

Cloudera was formed in 2008 to help enterprise companies use Apache hadoop to get more valuable output of all of their data. Cloudera's open source platform, is the most popular distributed big data technologies. The big contributions of Cloudera to big data world is the abstraction over different big data technologies such as Hadoop, Hive, HCatalog etc.it provides easy to use technologies without going into technical details and also provides the system management, deployment, configuration, security management, diagnostics and repots generation etc.

I. HortonWorks

This technology funded in 2011 by the ex-Yahoo engineers. Its supports different type of technologies for data integration and data flow control. Thus technologies are Apache Falcon, Sqoop and Flume this are the part of the platform and provides easy and systematic access for handling data in and beyond Hadoop. If we are deploying and configuring HortonWork data platform entails substantial expertise and training to use it along with other technologies experience.

J. Amazon Big data Analytics Platform

Amazon big data analytics platform provide to analyze very huge data sets requires momentous compute capacity that can differ in size. As per the requirement, it can easily change or resize the environment on AWS to meet the needs without having to delay for further more hardware, or being required to over invest to provision sufficient capacity. Hadoop or Hive both can be accessed via the amazon exposed API's or via launching an Elastic Compute Cloud by Amazon instance.

K. Hewlett Packard Big data Platform

HPE provide Vertica analytics platform for analyzing big data. The Vertica database is a backbone of the HPE product. Instead of fast execution of queries, It execute set type queries for warehouse application due to its CSS (columnar storage structure). This is to provide advance compression techniques to compress our data to store on the disk along with In-database analytics. It supports standards SQL, R and Python based analytics. HP Vertica have wide range of built-in analytical functions, time series, pattern matching, geospatial etc.

TABLE I : Comparison of different Enterprise Big Data Analyzing Tools

Features	Analysis web Interface	Hadoop Support	Spark Support	Analytics	Database Support
HP	No	No	Yes	Yes	Yes
Amazon	Yes	Yes	Yes	No	No
SAP	No	Yes	Yes	Yes	Yes
HW	No	Yes	Yes	No	No
Cloudera	Yes	Yes	Yes	No	No
TD	No	Yes	Yes	Yes	Yes
IS	Yes	Yes	Yes	Yes	Yes
C-DB	Yes	No	No	Yes	Yes
GBQ	Yes	Yes	Yes	Yes	Yes
GHD	Yes	No	No	Yes	Yes
H	No	Yes	Yes	Yes	Yes

Legends **:** HW=Horton works ; TD=TeraData ; IS=InfoSphere ; C-DB=CitusDB ; GBQ=Google Big Query ; GHD= Greenplum HD ; H=Hadapt

VI. OPEN SOURCE BIG DATA ANALYZING TOOLS

A.HIVE Web based interface

This is an alternative option to use the Hive command line interface. The Hive web based interface, abbreviated as HWI or Hive WebUI, is a simple graphical user interface (GUI). This is an interactive web interface which is basically designed for administration purpose of the Hive and as well as for querying the database. In Hive user can create tables and delete tables and also browse the database schema. Mostly more users can execute the queries by supplying it from the webUI even though Hive is not actually an analytical platform easily interacts using the webUI. Hive webUI to work, the user should have Hive configured and deployed on the computer system which requires Hadoop as well for processing.

B. Apache Drill

Apache Drill is a low latency distributed query engine for large-scale datasets, including structured and semi-structured/nested data. Inspired by Google's Dremel, Drill is designed to scale several thousands of nodes and query petabytes of data at interactive speeds that BI/Analytics environments require. Drill includes a distributed execution environment, purpose built for large-scale data processing. At the core of Apache Drill is the 'Drillbit' service, which is responsible for accepting requests from the client, processing the queries, and returning results to the client. A Drillbit service can be installed and run on all of the required nodes. Hadoop cluster to form a distributed cluster environment. When a Drillbit runs on each data node in the cluster, Drill can maximize data locality during query execution without moving data over the network or between nodes. Drill uses ZooKeeper to maintain cluster membership and health-check information.

C. Apache Giraph

Apache Giraph is an iterative graph processing system built for high scalability. For example, it is currently used at Facebook to analyze the social graph formed by users and their connections. Giraph originated as the open-source counterpart to Pregel, the graph processing architecture developed at Google. Giraph adds several features beyond the basic Pregel model, including master computation, sharded aggregators, edge-oriented input, out-of-core computation, and more. With a steady development cycle and a growing community of users worldwide, Giraph is a natural choice for unleashing the potential of structured datasets at a massive scale.

TABLE II : Comparison of different Open Source Big Data Analyzing Tools

Attributes	**Apache Drill**	**Hive**	**Giraph**
Owner	Community	Community	Community
Low Latency	Yes	No	No
Data can be obtained from	HDFS, Hbase, Cassandra, Mongo DB, RDBMS.	HDFS, Hbase	HDFS, Hbase
Hadoop Dependent	No	Yes	Yes
Attributes	**Apache Drill**	**Hive**	**Giraph**
Schema	Optional	Required	Required
Source code	Open	Open	Open
Columnar Storage	Yes	Yes	No
Mode of Operation	On-Premise	On-Premise	On-Premise

VII. BIG DATA ANALYZING TOOL SELECTION

While selecting an analytical tool, the following features must be taken into account.

- Ability to create and publish by the tool users
- Availability of Context-based filters. The Filters will list only choices that have values given the current selection of facts and dimensions.
- Availability of Context-based visualizations. Only visualizations or chart types that are relevant to the data selected will be listed as options.
- Availability of Advanced visualizations. More advanced visualizations include heat maps, scatter plots, bubble charts, histograms, geospatial mapping and combinations of each of these, such as bubbles on a map.
- The tool must have the ability to make annotating analysis to share observations and enabling discussion threads or chats.
- The tool must have the ability to have a dash board view to enable easy data access.
- The tool must allow Offline updates. The analytics tool, when it stores its own copy of the source data in an OLAP cube or in-memory columnar data store, should allow users to schedule automatic data updates.

VIII. CONCLUSION AND FUTURE WORK

To conclude, after the analysis of both enterprise and open source Big Data Tools, it is pretty lucid that it's all about the usage and needs of an individual or the enterprise. It is highly impossible to afford few tools at a personal level because of the prices and complications; while using open source systems might pose problems arising out of out dated features. Security aspects of the tools must also be taken into account. Open source promotes development and innovation and supports developers. We look at all tools and frameworks provide by various organization and comparison table. InfoSphere is best tool so far. For the future work, we plan to develop our own Hadoop system Comparing Apache Spark, storm and Map Reduce with Performance Analysis using K-Means.

REFERENCES

[1] Marcus R. Wigan, Roger Clarke, "Big Data's Big Unintended Consequences" Published by the IEEE Computer Society, pp 46-53,June 2013
[2] K. Balakrishna1,Smt. S. Jessica Saritha2, C. Penchalaiah3 "Extracting Structue Data From UnStructured Data Through HiveQL" Volume 4 Issue 4 April 2015.
[3] David A. Maluf "Managing Unstructured Data With Structured Legacy Systems" Version 2, Updated November 08, 2007.
[4] Alexander Lang, Maria Mera Ortiz, Stefan Abraham "Enhancing Business Intelligence with unstructured data" 2015.
[5] Thu Zar Mon "Design and Implementation of Structured and Unstructured Data Querying System in Heterogeneous Environment" Volume 2, No 5, May 2013
[6] Padmapriya.G, M.Hemalatha "A Recent Survey on Unstructured Data to Structured Data in Distributed Data Mining" March-April 2014
[7] Dr.S.Chitra M.E, Ph.D, Mrs.N.Shunmuga Karpagam M.E, Mr.K.Venkataramanan "Unstructured Data into Intelligent Information Analysis and Evaluation " Vol.2, Special Issue 1, March 2014
[8] Yingcai Wu, Nan Cao, David Gotz, Yap-Peng Tan, and Daniel A. Keim "A Survey on Visual Analytics of Social Media Data" 2016.
[9] A. Venkata Krishna Prasad1 and CH.M.H.Saibaba2 "Mining Big Data: Current Status, and Forecast to the Future for Telecom Data" 2014
[10] Zaharaddeen Karami Lawal, Rufai Yusuf Zakari, Mansur Zakariyya Shuaibu, Alhassan Bala "A review: Issues and Challenges in Big Data from Analytic and Storage perspectives" Volume – 5 Issue -03 March, 2016
[11] C.Yosepu¸ P Srinivasulu¸ Bathala Subbarayudu "A Study on Security and Privacy in Big Data Processing" Vol. 3, Issue 12, December 2015
[12] Abdul Raheem Syed, Kumar Gillela, Dr. C. Venugopal "The Future Revolution on Big Data" Vol. 2, Issue 6, June 2013
[13] Poonam G. Sawant, Dr. B.L.Desai "Big Data Mining: Challenges and Opportunities to Forecast Future Scenario " Vol. 3, Issue 6, June 2015
[14] M.Archana, Ch.Pravallika, J.Ravi Chandra Reddy "Mining Big Data: Current Status, and Forecast to the Future" Volume 3[9], pp: 5004-5008, September 2015