7th National Conference on
"Role of Engineers in Nation Building" organized by VIVA Institute of Technology, Mumbai
(1st and 2nd March 2019)
International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

Big Data and Tableau: The Review

Kalpita Surve[1], Prof. Shreya Bhamare[2]

VIVA School of MCA, University of Mumbai, INDIA

Abstract**—** *Big data is nothing but simple data with huge size collected from various sources. Big Data analytics is about capturing, storing, processing and managing that data continuously. Tableau is current popular Data Visualization Tool used in Business Intelligence. Tableau supports various Hadoop distributions, No SQL Database and Spark and hence it integrates easily with environment, no matter what is source of data or which software used for storing that data. In this paper we will concentrate on working of Tableau. By using Tableau software, we will impose information in order to analyze, extract and manage valuable information.*

Keywords—Big Data, Big Data Analytics, Business Intelligence, Data Visualization tool, Treasure Data

I. INTRODUCTION

Big Data is simple data in a large size ranging from many Terabytes, Petabytes, Exabytes to Zettabytes and Yottabytes. Big data is representation of 3 characteristics as "Volume" which concludes size of data, "Variety" which concludes different sources of data and "Velocity" that depicts amount of data accumulating with time or a simply growth rate & speed of execution. Nowadays Big Data has many use cases in private, science and government sectors. Some facts in Science sector says NASA uses 1.73 Gigabytes every hour, In Government sector NSA Utah Data Center has Yottabyte Capacity and Ebay has 40PB Hadoop Clusters for search, consumer recommendations and merchandising. Even Facebook has 30PB Hadoop Clusters, 50Billion photos and 130 TB of logs every day. So, these use cases are analyzed with some strategies that turn out as Big Data Analytics. It means to capture, store, process, and manage data continuously emerge and evolve. Hence, it's clear that big data analytics strategies need flexibility and agility to meet changing business demands. Tableau is a visualization tool that gives drag & drop functions to analyze data on large scale very easily and quickly. Tableau is a tool to flexibly connect across Big Data platforms, cloud data sources and relational databases to give users the agility they need for analyzing data. Tableau can query relational databases, cubes, cloud databases and spreadsheets to generate dynamic result in terms of graphic charts.

II. BIG DATA OVERVIEW

The Big Data is a combination of datasets from multiple sources stored & transformed together for finest decision making in making of strategies. A big Data platform enables to collect, store and manage more data than ever before. Previously unseen patterns emerge when we combine and cross-examine very large data sets. These patterns contain critical business insights that allow for the optimization of business processes. A Big Data can be SQL or NO SQL Databases considered as a One Big File which is broken into small multiple pieces or file termed as blocks are replicated and distributed over a distributive computing node. Exploring and analyzing big data converts information into insight. However, the massive scale, growth and variety of data are simply too much for traditional databases to manage.

And for Managing Big Data efficiently there are managing frameworks with respective File system. For this reason, businesses are turning towards technologies such as Hadoop, Spark and NoSQL databases to meet their rapidly evolving data needs and Business Intelligence. A Big Data and its problem can be understood by an example of Stocks Dataset where day by day stock information for several symbols for several years having around size of 1 TB as shown in Fig. 1. Now problem is finding out maximum closing price for each symbol. Now for this traditional way will include data access rate, program computation. time and network, bandwidth etc. This could take too much time for execution, for this we can split 1TB File into 100 equal sized blocks and read them in parallel & this would reduce lot of time.

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

```
ABCSE,B7J,2008-10-28,6.48,6.74,6.22,6.72,44300,5.79
ABCSE,B7J,2008-10-27,6.21,6.78,6.21,6.40,55200,5.51
ABCSE,B7J,2008-10-24,6.39,6.66,6.21,6.40,67400,5.51
ABCSE,B7J,2008-10-23,6.95,6.95,6.50,6.59,59400,5.68
ABCSE,B7J,2008-10-22,6.92,7.17,6.80,6.80,55300,5.86
ABCSE,B7J,2008-10-21,7.20,7.30,7.10,7.10,54400,6.11
ABCSE,B7J,2008-10-20,6.94,7.31,6.94,7.12,45700,6.13
ABCSE,B7J,2008-10-17,6.43,6.93,6.42,6.90,57700,5.94
ABCSE,B7J,2008-10-16,6.61,6.69,6.21,6.53,83200,5.62
ABCSE,B7J,2008-10-15,6.84,6.90,6.36,6.36,78900,5.48
ABCSE,B7J,2008-10-14,7.15,7.32,6.93,6.96,74700,5.99
ABCSE,B7J,2008-10-13,6.00,6.57,6.00,6.57,75700,5.66
ABCSE,B7J,2008-10-10,5.05,5.72,4.79,5.72,158400,4.93
ABCSE,B7J,2008-10-09,6.30,6.41,6.00,6.02,140500,5.18
ABCSE,B7J,2008-10-08,5.60,6.47,5.60,6.28,292000,5.41
ABCSE,B7J,2008-10-07,7.59,7.59,6.66,6.69,89900,5.76
ABCSE,B7J,2008-10-06,7.83,7.90,7.00,7.40,159600,6.37
```

Fig. 1 Data set of Stock Symbols

So, to overcome Big Data challenges like Storage, Computational Efficiency, Data Loss, cost and to process the data a Framework is needed. A framework manages data processing with its ecosystem and respective technologies. There are number of frameworks incorporate in Big Data that provide analytics with connectors for Big Data ecosystem.

2.1 Connectors for Big Data:

- NoSQL: MarkLogic, Datastax
- Cloud: Amazon Redshift, Google BigQuery
- Hadoop: Cloudera Impala & Hive, Hortonworks Hive, MapR Hive, Amazon EMR with Impala & Hive
- Spark: Apache Spark SQL [3]

2.2 Different Datasets:

- Structured Data: This is the most typical easily accessible data with refined form one organization can have. Here aggregates of data are pre-computed and pulled it as extract. That extract becomes base for in memory computing and aggregated for analysis
- Semi Structured Data: It is also known as object storage. In short, a data stored in relational databases, data warehouses and data marts. These are regularly refreshed for entity analysis like transactions, actions taken by individual sales person.
- Raw, unstructured Data: It is a data in cloud storage or data lake. The data created and collected through IOT devices and social network feeds [1].

2.3 Hadoop (High Availability Distributed Object Oriented Platform) :

Hadoop is a most popular framework for distributed processing. It supports huge volume & storage efficiency. It has good data recovery solution and it is cost effective. Hadoop is a like batch processing system that gives quick throughput. Hadoop ecosystem is a power pack of multiple services that stores, analyze and maintains the datasets. It is a platform that is buildup of two main components as Hadoop Distributes File System (HDFS) and Map Reduce. Other than these components Hadoop ecosystem also include Yet Another Resource Negotiator (YARN)and Language platforms like Hive for storing datasets (Fig. 2). HDFS is java based primary file system which provides reliable, fault tolerance, scalable and cost-efficient data storage for Big Data. HDFS consists of Name Node and Data Node which are daemons of Hadoop. It is reliable shared storage. These nodes are java programs that run on specific machines. Data Node has data blocks which are managed by Name Node. A Node is combination of CPU, RAM and Disk and a network to communicate. A Name Node has Metadata of Files & Folders in Disk and contains Block Locations in Memory. A Rack is collection of multiple nodes and a Cluster is Racks interconnected over network [7]. Hadoop is a distributed file system that runs on commodity hardware with help of shell-like commands. Just like ls, mkdir, cp, mv,rm Local File system commands HDFS has Some of basic commands[8]. Map Reduce itself is a framework of two phases as Map phase and Reduce phase. It is core Hadoop ecosystem component which provides data processing with distributed

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

computation. It is distributed programming model for processing large data sets. These are codes implemented by any programming language. Map Reduce is not programming language. Hadoop implements Map Reduce to manage communications, data transfer, and parallel execution across distributed servers. In Map Phase a mapping code is distributed over machines, they work on data presented on

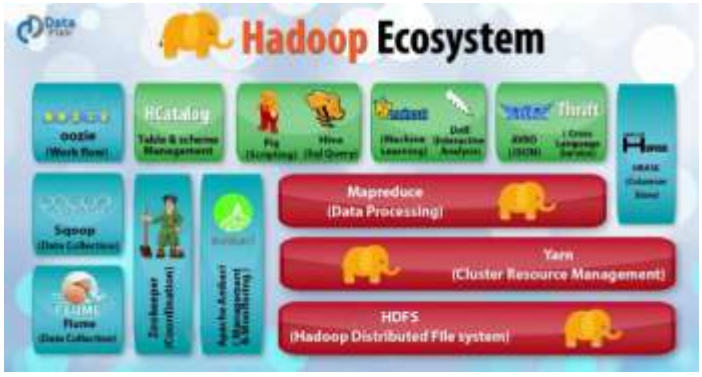

Fig. 2 Hadoop Ecosystem

III. TABLEAU OVERVIEW

Tableau is a powerful data visualization tool that provides drag & drop features to analyze data on large amounts very easily and quickly [9]. The dashboard of Tableau gives dynamic results. It can query relational database, cubes, cloud database and spreadsheets. Tableau generates number of graphs which can be combined into dashboard & shared via network. Tableau reduces the need to pull individual reports from multiple software or databases. See it all in one place and discover new patterns you might never have seen using separate individual reports. Tableau explores this data through the eyes of your entire organization, allowing for more questions to be asked and more discoveries to be made. It deals with democratization of data which is a concept where the people who know the data should be the ones empowered to ask questions of the data. Tableau empowers people throughout the organization to answer questions of their data, large or small, in real-time. The more questions they ask, the more value they extract from the data, leading to smarter business decision every day. Tableau allows you to have one interface for all of your data, regardless of where that data is stored.

Now, businesses are turning towards technologies such as Hadoop, Spark and NoSQL databases to meet their rapidly evolving data needs. From manufacturing to marketing, finance to aviation Tableau helps businesses see and understand Big Data. Tableau supports a specific language which is Visual Query Language (VizQL). It translates drag & drop actions into data queries and then expresses that data in a visual format. The data engine of Tableau is based on advance in-memory technology to speed up ad-hoc analysis of large data in few seconds. The results of query are cached within a memory to provide results quickly on even large number of datasets. Tableau supports only Windows and Mac OS & there is no support for linux. No matter what kind of data consumption or storage software you're using, from various Hadoop distributions to NoSQL databases to Spark, Tableau integrates seamlessly with your infrastructure. Systems like Tableau that support large volumes of both structured and unstructured data will continue to rise. A traditional analysis tool forces to analyze data in rows and columns, choose a subset of your data to present & organize that data into a table, then create a chart from that table [1].VizQL skips all these above things and creates a visual representation of your data right away, giving you visual feedback as you analyze. VizQL allows you limitless exploration your data to find the best representation of it—and with unlimited "undo," there is no wrong path [2].

IV. BIG DATA WITH TABLEAU

Big Data platforms enable you to collect, store and manage more data than ever before. Previously unseen patterns emerge when we combine and cross-examine very large data sets. In Big Data Analytics these patterns contain critical business insights that allow for the optimization of business processes that cross department lines which helps in Business Intelligence. Tableau has focused on providing broad access to Big Data platforms that Enables analysis of Big Data, wherever it lives. Tableau supports

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

over 40 different data sources today as well as countless other sources through our extensibility options. Besides the available measures and dimensions, you can create a calculated expression to develop a new visualization. It may be based on date, mathematical logic, text expressions, input parameters and others. For Data scientists, they can connect tableau with R to enhance the power for analytical inferences. Treasure Data is cloud-based, managed service for data and analytics [5]. Treasure Data empowers data-driven companies to focus on insights, not infrastructure. Treasure Data provides a scalable backend to handle new big data sources (application logs, web logs, mobile data, sensor data, etc), while Tableau provides flexible visual analytics for existing data sources (EDW, CRM, ERP, etc). Users can store trillions of records in the cloud by collecting semi-structured big data in real-time, and aggregate the data by using one of several query engines. Often times, those results will be fed to a data warehouse or reporting server for consumption by additional end-users. By combining Treasure Data and Tableau, you can get the insights from any data sources of any size quickly. Here's a reference architecture diagram. Let's see how it works step by step.

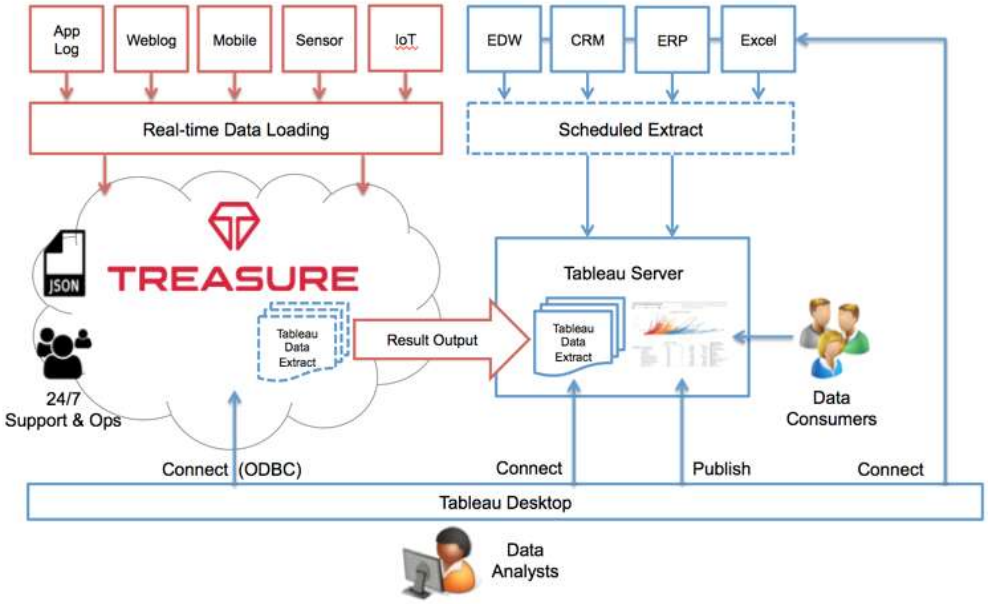

Fig. 3 Reference Architecture Diagram

4.1 Collect Big Data (Treasure Data):

Treasure Data provides four main data collection capabilities as Treasure Agent for streaming data collection, Bulk Loader for parallel bulk loading, JavaScript SDK for website tracking, Mobile SDK (Android, iOS, Unity) for mobile application tracking.

4.2 Aggregate Big Data (Treasure Data):

We can crunch big data into the aggregated format by using one of Treasure Data's embedded query engines. Treasure Data supports 'Tableau Result Output' so you can directly push the aggregated results into Tableau Server.

4.3 Design Workbooks (Tableau Desktop):

To explore raw data and aggregated data too Tableau Desktop is the tool for it.

4.4 Share the Workbooks (Tableau Server):

Analysts can publish workbooks to the server in form of video or document, then the data consumers can see them from their browsers. The beauty is analysts can quickly iterate on the data and reports by having access to all the data sources, so they're now self-reliant.

7th National Conference on
"Role of Engineers in Nation Building" organized by VIVA Institute of Technology, Mumbai
(1st and 2nd March 2019)
International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

You can work directly with your data to create reports and dashboards. Tableau lets you connect live or bring your data into its fast, in-memory analytical engine. An option for fast in-memory data engine or live database connection Tableau has optimized direct connections for many high-performance databases, cubes, Hadoop, and cloud data sources such as Salesforce.com and Google Analytics. With drag & drop, point & click ease to build charts, reports and dashboards, Tableau gets people throughout an organization connected directly to their data. There's no more waiting in an IT queue to answer questions, so now you can begin getting answers from your data. This collaboration does not support functionalities such as embedded BI, metadata management and data preparation as well as predictive analytics functionality to uncover hidden data relationships. It is difficult to interpret complex business rules in Tableau [6]. Tableau enables users to traverse across data sources by blending Big Data with other data sources (e.g. Salesforce, MySQL, Excel files), allowing organizations to keep their data assets where they reside.

V. DATA VISUALIZATION OF OPEN SOURCE DATA SET

In this paper we will impose data set of commodities with their details [4]. A CSV Data set is opened in Tableau for Data Visualization & analysis which can be seen in Fig. 4. Visualization for purchase orders placed for Austin & Baker City in 2012-2018-time spans is shown in Fig. 5. Now visualization for starting 3 months in 2010-2018 spans in Austin & Baker City is shown in Fig 6. Purchase orders placed in last 3 months of 2018 year is visualized in Fig. 7

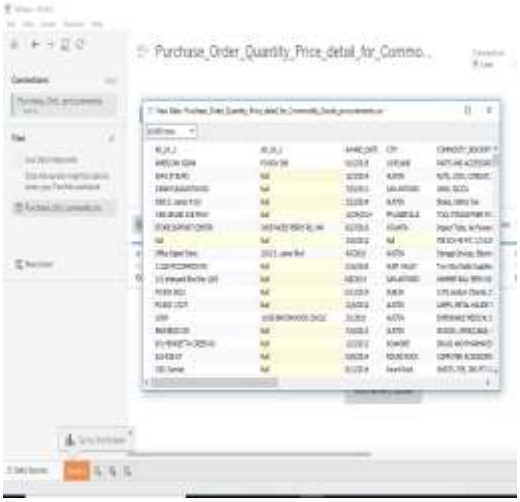

Fig. 4 Data set in Tableau

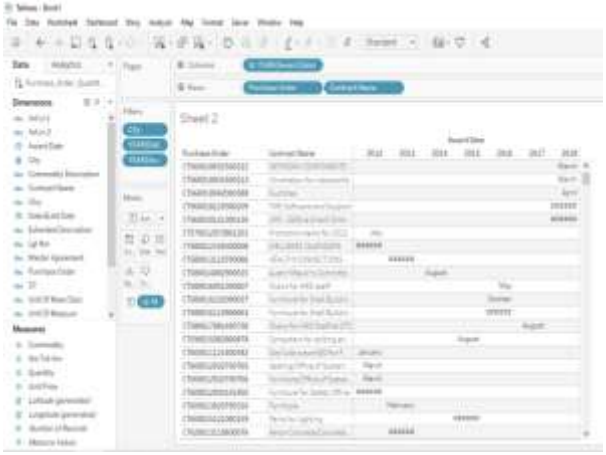

Fig. 5 Purchase Orders for 2 cities in 2012-2018 yrs.

Fig. 6 Commodity for starting 3 months 2010-2018 years

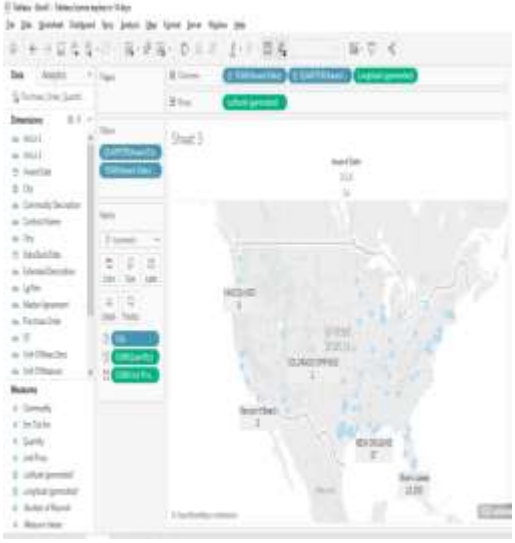

Fig. 7 Commodity in 2018 from Oct-Dec in 2 cities

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

VI. CONCLUSION

In this paper we concluded that Big Data & Tableau collaboration is one of best option for visualization as it supports many data sources, on the fly ETL & custom SQL. No special configuration is required for hadoop & hive in Tableau, which makes it easy & efficient analysis process. In this paper we successfully visualize a big data example of one Open source dataset in Tableau. We have visualized data in three aspects and graphics very easily & quickly with help of Tableau.

REFERENCES

[1] https://www.tableau.com/learn/whitepapers/tableau-big-data-overview
[2] http://hadooptutorial.info/tableau-integration-with-hadoop/
[3] https://www.slideshare.net/mobile/idigdata/big-data-analytics-with-tableau
[4] https://data.austintexas.gov/Budget-and-Finance/Purchase-Order-Quantity-Price-detail-for-Commodity
[5] https://blog.treasuredata.com/blog
[6] https://selecthub.com/big-data-analytics-tools/tableau-big-data-analysis/
[7] www.hadoopinrealworld.com
[8] www.hadoopskills.com
[9] www.hadooptutorial.info