

An outline of Tools and Technologies used for Big Data Analytics

¹Khyati R. Nirmal, ² Dr. M.R.Sanghavi

¹Assistance Professor, Department of Information Technology,

²Professor, Department of Computer Engineering,

^{1,2}SNJB's LS KBJ College of Engineering, Chandwad, MH, India

Abstract

To bring about the recent necessities of Big Data thousands of Big Data Tools and Technologies are out today. As per the skill set and the task one should choose the promising one. The contribution of these tools and technologies is notified at all 5 Vs (Volume, Velocity, Variety, Veracity, and Value) of Big Data. As per their roles, the tools and technologies are categorized in various aspects like Data Collection, Data Extraction, Data Storage, Data cleaning, Data mining, Data visualization, Data analyzing and Data integrating. Ruling out preeminent tools for your application is a crucial job; the target objective of this chapter is to make the entire information available under one roof..

IndexTerms –Big Data, Data Mining, Data Cleaning, Data Visualization, Data Integration.

INTRODUCTION

The cumulative frequency of Data Generation is being increased and gigantic because of the social and network activities of individuals and nonliving objects. The data generated with Big volume defined as Big data which is characterize by Five Vs (Volume, Velocity, Variety, Veracity, and Value) [1]. To handle the different nature of Big Data the traditional technologies are not adequate. The common chalanges identified are Data Collection, Data Extraction, Data Storage, Data Cleaning, Data mining, Data visualization, Data analyzing and Data integrating and many more.

This paper brief about the different tools and techniques available in each category that can assist the Big Data Analyst to choose appropriate tools as per the requirements.

DATA STORAGE AND MANAGEMENT:

The basic requirement of Big Data Storage infrastructure is to manage and retrieve the huge amount of data along with the huge storage of data. The ideal Big Data storage infrastructure is able to handle virtually unlimited amount of data, high rates of random write and read access, support for the various data models, along with structured and unstructured data. To provide the security and privacy the Big Data Infrastructure must work with the encrypted data.

Apache Hadoop:

It is an open source framework for the distributed storage of Big Data. The storage system of Apache Hadoop frame work HDFS (Hadoop Distributed File System) system is very similar to distributed file system. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. Fault tolerance and high throughput access to application data are foremost features of HDFS. While working with the large data it is not efficient to move the data where application is stored, to overcome this HDFS migrates the computation closer to where the data is located. HDFS uses master slave architecture. One node is intended as Name Node and the remaining nodes are intended as Data Nodes. The Name Node stores all metadata about the master slave architecture like information about file locations in HDFS, file ownership and permissions information, Names of the individual blocks, Locations of the blocks, etc. HDFS divides the data into chunks and distributes them to different nodes in a cluster, allowing for parallel processing. HDFS copies each chunk of data multiple times and distributes the copies to different nodes,. As a result, the multiple copies of data is maintained which can allow processing in case of failure of any individual node. The architecture and working of HDFS is described in the figure1.

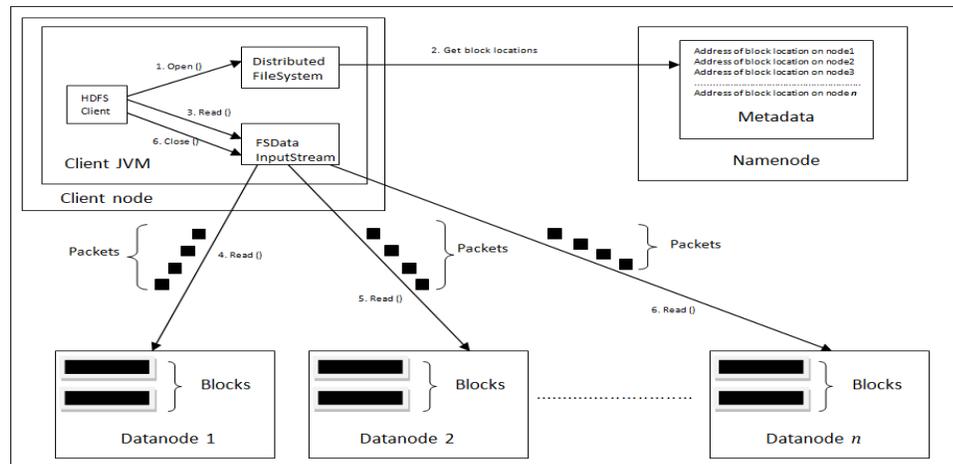


Figure 1 HDFS Architecture

Source: <http://hadooputorials.co.in/tutorials/hadoop/internals-of-hdfs-file-read-operations.html>

Microsoft HDInsight:

Microsoft HDInsight is the services offered by the Microsoft Azure cloud service provider. It is an Apache Hadoop-based distribution running in the cloud on Azure. HDInsight uses the Hortonworks Data Platform (HDP) Hadoop distribution. It is a software framework designed to process, analyze, and report on big data with high reliability and availability. It consists of implementations of Apache Spark, HBase, Storm, Pig, Hive, Sqoop, and so on. HDInsight uses Windows Azure Blob storage as the default file system. The Microsoft Azure HDInsight is able to work with unstructured or semi-structured data and enables users to add capacity on demand. This also provides high availability with low cost. The architecture of HDInsight is described in figure 2.

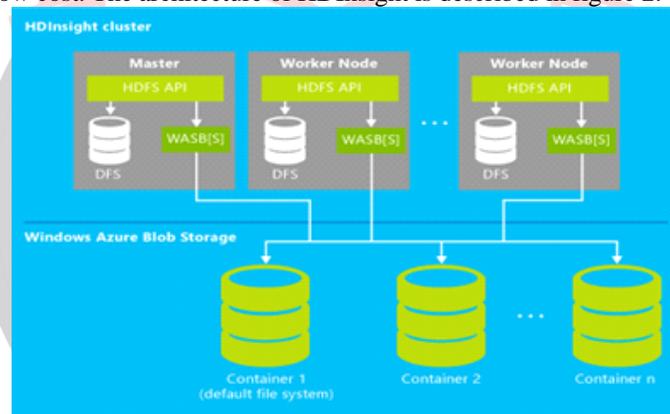


Figure 2 Hadoop clusters in HDInsight

NoSQL:

Big data is unstructured, messy and unpredictable in nature. To manage the Big Data well traditional SQL agree to take challenge like High concurrent of reading and writing with low latency, High scalability and high availability, Lower management and operational costs. NoSQL has come up with solution by proposing support to unstructured data, quick reading and writing data, support mass storage, easy to expand and low cost. NoSQL database pattern are of four types. The NoSQL Database Type along with specification is described in figure 3.

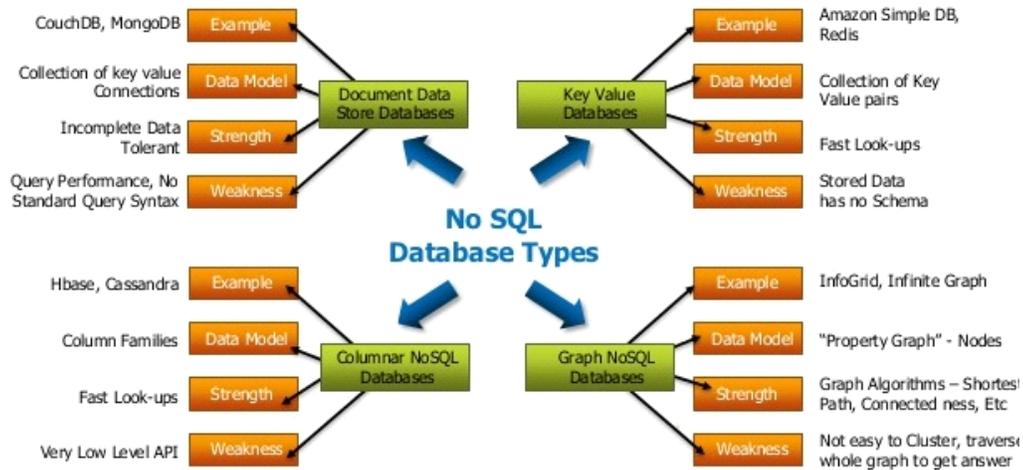


Figure 7.3 No SQL Database Types

Source :<http://www.edureka.co/blog/choosing-the-right-nosql-database/>

Apache Hive:

On the top of the Apache Hadoop the Apache Hive is integrated, which consumes dimensions to read, write and manage large dataset. Apache Hive is having Data ware house like structure and provides SQL like interface to query for Extract, Transform and Load(ETL) large data sets. Hive is not suitable for Online Transaction Processing (OLTP) workloads. Hive is designed to maximize scalability (scale out with more machines added dynamically to the Hadoop cluster), performance, extensibility, fault-tolerance, and loose-coupling with its input formats. Apache hive consist of mainly two Components HCatalog and WebHCat. The table and storage management layer for Hadoop is known as HCatalog. HCatalog consist of data processing tools like Pig and MapReduce .As are result it is more easy to read and write data on the grid. WebHCat provide platforms to run HadoopMapReduce (or YARN), Pig, Hive jobs or perform Hive metadata operations using an HTTP (REST style) interface. The architecture of Apache Hive is illustrated in Figure 4.

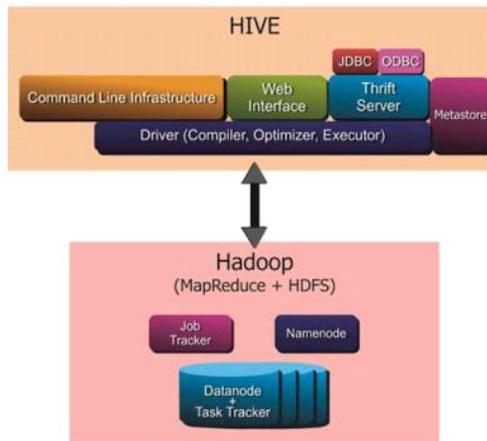


Figure 7.4 Apache Hive architecture

Source:<http://www.hadoopmaterial.com/2013/10/hive-architecture.html>

Apache Sqoop:

Transferring bulk data between Hadoop echo system like HDFS ,Hbase ,Hive and structured data stores such as relational databases ,Oracle etc. is challenging task. Apache Sqoop is working in the same direction. Apache Sqoopis SQL to hadOOP and hadOOP to SQL. Apache sqoop imports and exports on various dimensions like single or multiple tables, complete database, particular columns, filter out selected rows and many more.

Figure 5 Bulk Data Transfer using Sqoop2

Source: Sqoop 2 introduction, by Mengwei Ding at Sqoop User Meetup during Hadoop Summit 2013, San Jose, CA

DATA CLEANING:

Data cleaning is essential preprocessing step of mining large and messy data. Data Cleaning is the method to remove, replace or modify the duplicate, inconsistency, erroneous, noisy, incomplete data from unprocessed data. Incorrect and inconsistency data will lead to inappropriate conclusions. Various tools and techniques are available to refine and reshape the data into a useable data set.

3.1 OpenRefine:

OpenRefine (formerly GoogleRefine) is an open source tool which is designed to clean messy data. Using Open Refine it is easy and quick to explore huge structured or unstructured data sets. OpenRefine is used for the common scenarios like: find out occurrences of particular value in column, distribution of values across data set, manage the different date format and convert it into universal format, Split the one column data sets into multi column data sets, to provide additional data from external data source and many more. Many contributors are continuously providing various techniques to handle various scenarios as a result OpenRefine software is constantly getting better and better.

3.2 DataCleaner:

DataCleaner is equipped with strong data profiling engine that is useful to discover and analyze the quality of the data. Mainly Data Cleaner has the features like searching for particular patterns, dealing with missing values, finding out characteristics of data values. Machine Learning based principles are deployed for Duplicate detection feature that enables to improve inferential matching. DataCleaner handles csv files, Excel Spread sheets, Relational databases and NoSQL databases. It also allows user to define their own cleaning rules whether it may be simple find and replace, regular expressions, and pattern matching or custom transformations. It also supports periodic data quality checks and send notification if quality does not meet to define standards. It is also community driven application and provide plug-in that can easily be integrated.

IV. DATA MINING

Data mining is the method to extract hidden predictive information from data. For Big Data Mining due to Large volume of data, the single personal computer is not adequate to extract hidden information from the data. For this reason it is necessary to deploy Data Mining task on Parallel Processing framework like Map Reduce. Variety of tools and technologies exist that are efficiently retrieving the hidden information from Big Data. Only some tools are listed below with sketchy description

4.1 RapidMiner:

RapidMiner (formerly YALE) is the platform offers software, solution and services for data mining along with data analytics text mining also. RapidMiner analyze the data astutely without human intervention from both structured and unstructured data. It is suitable for research and real world Data Mining Task. Commercially it is available in open source as well as enterprise edition. The intuitive GUI of RapidMiner provides various integrations of elements, without explicitly programming. Big data analytics made easy using RapidMiner Features like:

- Stream Mining: Instead of storing complete data sets in memory only required part is taken in analytical process. Part processes can also be carried out in distributed form, e.g. in RapidAnalytics clusters or even on Hadoop.
- In-database-mining: Instead of relocating data to algorithm it relocate algorithms to the data.
- Radoop: The world's first graphical connection of Hadoop for the management of big data analytics, that unites potency of Hadoop and RapidMiner.
- RapidMiner also provides extensions for R, WEKA and Many more. The architecture of RapidMiner is illustrated in figure 6.

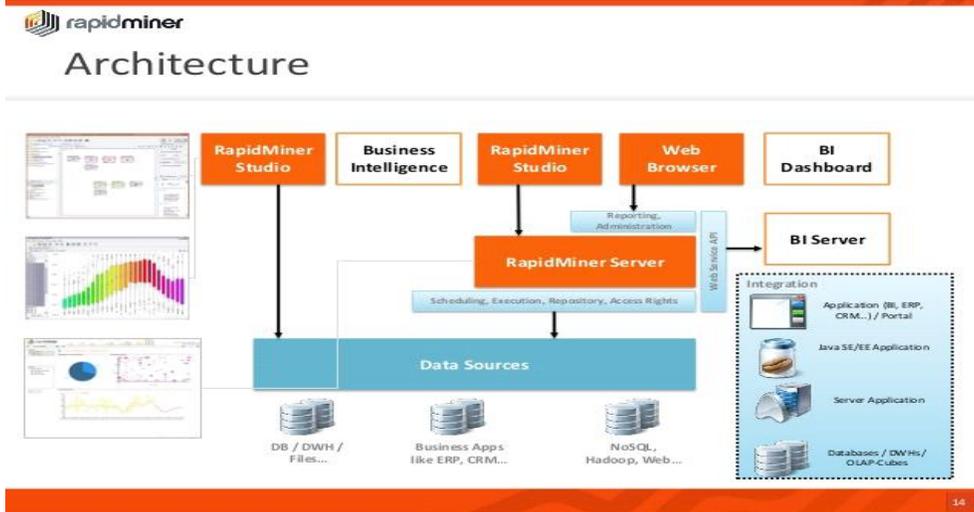


Figure 7.6 Architecture of RapidMiner

4.2 RStudio:

RStudio is the leading integrated development environment for R language. R has vast array of packages that covers the standard algorithm for Data Mining, Machine Learning, Neural Network and exotic specialized algorithm which are based on chaos model. RStudio offers open source and commercial editions on the desktop (Windows, Mac, and Linux) and it also offers a web browser to a Linux server running RStudio Server or RStudio Server Pro. RStudio provides all in one support for work space management, console for direct code execution, syntax- highlighting editor, debugging, plotting and history layout. Shiny and ggvis packages helps for interactive graphics of data visualization. R studio also provides extension for C,C++ and Python languages. R and Hadoop integration is perfect match with the advent of following software solutions.

RHIPE: It is R and Hadoop Integrated Programming Environment. To perform data analytics for Big Data it follows Divide and Recombine technique. In this Strategy data is divided into subsets, Specific R Analytics operation executes computation on these subsets and output is generated. In depth analysis of large data set and to perform the analysis with the help of HDFS are two main objectives of RHIPE. Architecture and working of RHIPE is explained in Figure 7.

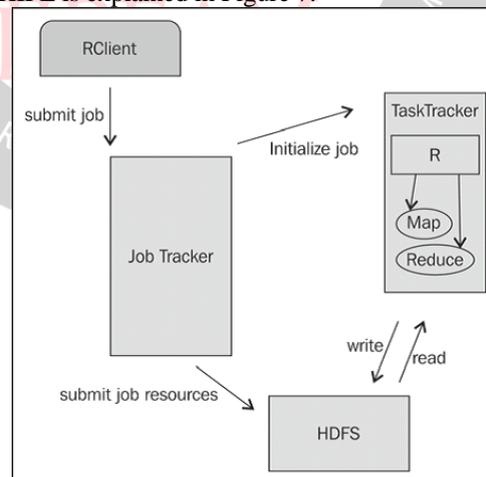


Figure 7 Components of RHIPE

RClient calls the Job Tracker to execute the job. It is the responsibilities parameters like Mapper, Reducer, input format, output format, input file, output file and many more related to MapReduce jobs. JobTracker act as a master node of Hadoop that will initialize and monitor the MAPReduce jobs over the Hadoop Cluster.

RHadoop: It is enriched with mainly 3 packages rhdfs, rmi and rhbase. It was developed by Revolution analytics, which is the leading commercial provider of software based on R.

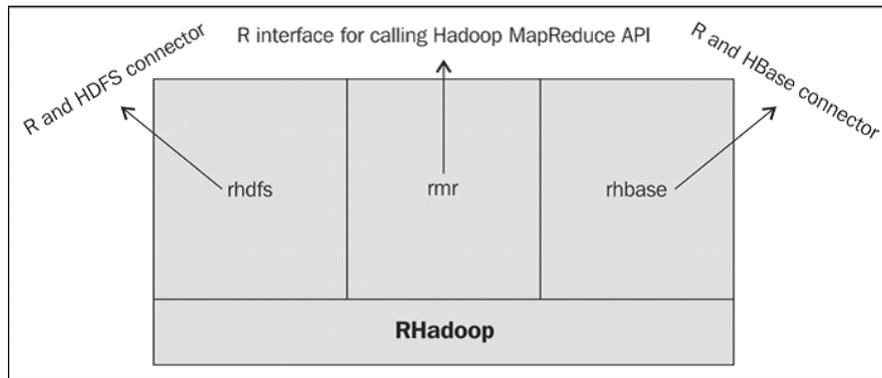


Figure 8 RHadoop Ecosystem

Rhdfs: It is mainly designed to provide HDFS usability from the R console. Along with access to read output from HDFS using Rhdfs It is very easy to perform read write operation on distributed databases. .

Rmr: It is the package designed for of HadoopMapReduce for the R environment. R programmer writes the application logic using map and reduce and submit it using rmi methods. It is the responsibility of Rmr to calls Hadoop streaming MapReduce API with several job parameters as input directory, output directory, mapper, reducer, and so on, to perform the R MapReduce job over Hadoop cluster.

Rhbase: It is the interface designed for operating the HadoopHBase data source stored at the distributed network via a Thrift server. The rhbase package is designed with several methods for initialization and read/write and table manipulation operations.

HadoopStreaming : It is a Hadoop utility for running the HadoopMapReduce job with executable scripts such as Mapper and Reducer. This is similar to the pipe operation in Linux. With this, the text input file is printed on stream (stdin), which is provided as an input to Mapper and the output (stdout) of Mapper is provided as an input to Reducer; finally, Reducer writes the output to the HDFS directory. The components of Hadoop Streaming is described in Figure 9.

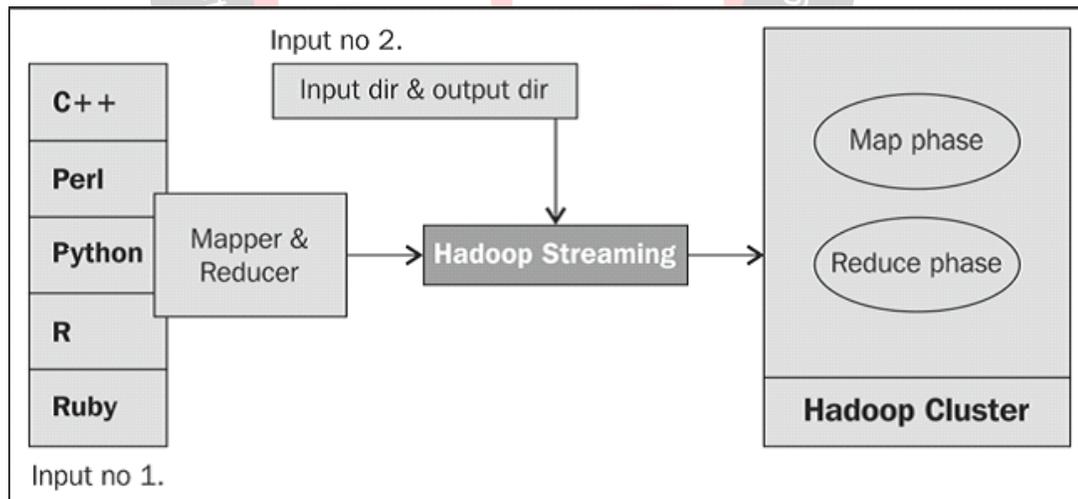


Figure 9 Hadoop Streaming Components

The key feature of theHadoop streaming utility is that it supports Java as well as non-Java programmed MapReduce jobs to be executed over Hadoop clusters. Also, it takes care of the progress of running MapReduce jobs. The Hadoop streaming supports the Perl, Python, PHP, R, and C++ programming languages. To run an application written in other programming languages, the developer just needs to translate the application logic into the Mapper and Reducer sections with the key and value output elements.

4.3 WEKA:

Waikato Environment for Knowledge Analysis (WEKA) consist a set of machine learning algorithms for data mining tasks. Weka is open source software issued under the GNU General Public License Along with the built in algorithm code user can import the java code to the WEKA also. User can also expand the new machine learning algorithms. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. To support the distributed data mining for Big Data a variety of packages are included in the version WEKA 3.8. Few are listed below:

- DistributedWekaBase: It is generalized package to support map and reduce tasks which is not tied to any specific distributed paradigm.
- DistributedWekaHadoop: It is designed particularly to integrate Hadoop Wrapper class and jobs with WEKA.
- DistributedWekaSpark: Same like above package is designed particularly to integrate Spark Wrappers with WEKA.

4.4 Orange:

Orange is an open source tool in which Data mining algorithms are implemented through visual programming or Python Scripting. Orange provides extensions for machine learning algorithm, bio informatics and text mining. Orange has capabilities to remember the choices for frequently used combinations. Orange comes with a variety of machine learning algorithms, user can add new algorithm and wrap it with Orange. The key feature of Orange is to explore the data set in easy way for training purpose.

V. DATA ANALYSIS

Data Mining is to finding out the useful information or unrecognized patterns from data and Data Analysis is to split the data down and evaluate and predict the consequences of those patterns. Basically in terms of database the data analysis is to fire the queries from database and retrieving results that may helpful in future to obtain some significant conclusion. Data Analysis with Big Data is challenging task and for that intention variety of Software services, solutions and platforms are existing to support Big Data Analysis.

5.1 Qubole Data Service

For the Big Data Analytics in Cloud Qubole Data Service (QDS) offers a highly integrated set of tools. QDS takes care of coordinate and managing Cloud resources, and focus on analyzing and using the data. QDS services include:

- Cloud Data Warehouse: QDS provides Apache Hive as a service in the Cloud. This allows you to define schemas over data sets stored in Cloud Storage systems such as Amazon S3.
- Analyze: A user interface that allows to run queries and jobs using data in a Cloud store such as Amazon S3. Resources include MapReduce, Spark, HBase, Presto, and Pig.
- Scheduler: A tool that allows you to run recurring commands in sophisticated workflows.

Qubole runs applications written in MapReduce, Cascading, Pig, Hive, and Scalding using Apache Hadoop. Qubole provides Presto as a service for fast, inexpensive, and scalable data processing and Data Analysis. Presto is an open source distributed SQL query engine developed by Facebook. Presto is used for running interactive analytic queries against data sources of all sizes ranging from gigabytes to petabytes.

5.2 BigML

Machine Learning for Big Data (BigML) has made machine learning easy, beautiful and understandable. Using Interactive a GUI for BigML it is simple to import your data and obtain predictions out of it. The entire procedure is broadly accomplished using these features.

- Sources: It is the representation of the raw data along with its features.
- DataSet: It is structured version of raw data. BigML computes a histogram for each numeric or categorical field.
- Models : It automatically finds the correlated patterns from statistical using the variety of task like classification , regression and many more.
- Prediction: Each prediction is associated with confidence value that measures the model's certainty on the prediction.
- Ensembles: It is composed of multiple models over different subsamples of datasets. It reduce error of single machine learning models that might overfit data.
- Evaluations: It provides easy way to estimate the performance of predictive model. Depending on the task solved by ensemble different performance is calculated.

5.3 Statwing

Statwing is primarily designed for data analysts who are not rich with technical knowledge. Statwing has provides everything from attractive visuals to multifaceted analysis. It also offers web based statistical analysis software for business users, data analysts and market researchers. Statwing makes data analysis simple. It provide built in support for commonly used statistical practices , as a result it is convenient for nontechnical to visualize and understand data like experts.

Data Visualization

Data visualization is the pictorial or graphical representation of data. It simplifies the analytics practice and helps decision makers to grab tricky concepts or discover innovative patterns. It is easy to drill down the data and present the data in the form of charts and graphs. For Big Data Scientist it is a big challenge to convey the insights of data are bright and easy way to convey complex data insights. The best part is that coding is not required any more.

VI. Tableau

Tableau is visual analysis tools to explore and analyze data using drag and drop functionalities. It supports data of any volume (peta bytes to tera bytes), Variety of data (Structured and unstructured) , clickstreams, sensor networks web logs and many more. It is to discover big idea from Big Data. Some key features of Tableau are:

- A Simply Delightful Data Visualization Environment
- Finding out Patterns in Big Data Using Tableau's Clustering Feature
- Using Highlighting to Find the Needle in the Big-Data Haystack
- Data Connections to Google Sheets and Shared Filtering

6.2 Silk

Silk is a cloud based application to visualization of data. Silk is a place to publish your data. Each Silk contains data on a specific topic. Anyone can explore a Silk and create interactive visualizations. Some key features of silk are:

- Drag-and-drop tools : Creating pages is fast and fun. One can easily pull in images, videos, tables, maps, charts and more when building out your site.
- Work together: For collaboration invite anyone to author or edit. It's perfect for teams who need to work together.
- Keep info private: User have complete control over who can see your pages. One can decide to Keep Silk private or allow public access.
- Fast data importer : Use it to convert spreadsheets into Silk pages. The information will be tagged and viewable in minutes.
- Auto magical map creation :Silk automatically convert addresses into an online map that your users can navigate with ease. Heat maps, too!

6.3 CartoDB

CartoDB is mainly design to visualize the location based data. It runs on cloud and Amazon Web Service and offeres service on demand. CartoDB is proposed to work with real-time maps that work on World Wide Web and mobile platform. Some features of CartoDB are as follows:

- CartoCSS provides look and feel facilities to easily visualize in data layer.
- A variety of extensions are available for on Google Maps and MapBox mapping layers. These maps include the basic functions –zoom, scroll, etc.
- It also supports Geocoding. Geographical information can be derived from elements different from coordinates.
- By Using CartoDB is easy to import data into tables from the dashboard, querying data via SQL or retrieving information from URLs. Other data collection may be imported from various formats.
- By using PostGIS, CartoDB can query and combine data sets using geospatial data.

The key feature of CartoDB is to visualize location data from spatial database it is not required to code particularly.

6.4 Chartio

Chartio provides simple, secure connections to the world's most popular data sources. By easy dragging and dropping facility of Chartio's Visual Query Engine it is simple to generate the most optimized SQL queries. Chartio collects data from various sources together into a single analysis or chart without any engineering and find hidden insight without a data warehouse. Chartio is a up to date solution for business intelligence, and its user interface is designed in a way that allows any professional to begin using the platform with minimal training. Some key features of Chartio are:

- Visually Analyze Data : To visualize the data two separate interfaces are designed. Data Pipeline is to designed for non technical user which is able to grab information using drag and drop interface. And SQL interface is designed for technical user the in which user can create the generalized statistical functions to create new metrics and calculation.
- Join Data Sources : Data sources can bring to gather within in web browser using Chartio's web-based tools. Users can combine different types of data with a few clicks, and then automatically prepare the chart of the results for the actions performed by user. Chartio allows the user to generate their own columns, and perform all arithmetic operation within the browser.
- Share Dashboards: To deal with real time data and latest information Chariton dashboards is designed. It has access specifies facility so user can restrict the others to control the private data access. Using Scheduler email reports user have option to share data with clients.
- Collaborate with Colleagues: Collaborations feature are enriches in Chartio which provide facility to create different permission based accounts for Admin, Team Member and so on. It is easy to share charts and schedule reports in team without any specific technical details. Any team member can demands for review data without concerning about the actual location of data.

6.5 Plot.ly

Plotly is the modern platform for agile business intelligence and data science. Visualize Data, Together is the main objective of Plt.ly. To work with Plt.ly one should perform 4 simple steps:

- Import data : It provides the facility to connect your dataset. But if user is not having datasets in proper format it also supports the csv files and manual data entries are also allowed.
- Create charts: It Effortlessly generates fabulous and instructive visuals using refined, open source, visualization library and online chart creation tool.
- Share findings: It is simple to share findings by set in them into a web page and exports the carts into eps formats, presentations .
- Track data provenance: Plotly charts are fundamentally tied to their underlying data. As charts are shared and data changes hands through your organization, Plotly tracks how that data transforms. As a cloud platform, Plotly is the only business intelligence solution to provide heavily audited industries with fine-grained information on data provenance.

VII.Data Integration

Data integration platforms are acting as a middle layer for each platform. While extracting the data using import.io, generating visualization using Tableau or Silk and sharing on network, the data integration plays an important role. Few Data Integration services are described below.

7.1 Blockspring

Wide-ranging of tools and data sources are used in commercial applications. To prevent the data from unauthorized identity Data security is a crucial concern. Blockspring Discovery Hub is a simple portal to explore and discover all of organization's data sources - new or old in the Cloud. Blockspring is functioning to cover following aspects.

- Using easy portal of Blockspring user can make a list of data sources.
- It is easy to explore existing data in organization
- To search out the data sources it provides integrated search engine.
- To provide the safety and security the build in Governance policies are maintained.

Some key features of Blockspring are:

- Executive Utilization Dashboard: It traces out the data resources used in organization. It also keeps the tracking of duration and services used by data resources along with the particular application and user. Thus it is convenient to find out idle services and application and as a result organization can optimize the cost with unused data services and applications. Along with optimization of cost Executive dashboard enhances utilization visibility. Blockspring attaches application to data , automate exports, refreshes, and warehousing.
- Google sheet Excel Tableau Slack Bubble: Blockspring is an exclusive application that facilitates all of the high-tech functionalities of automation services such as If This Then That (IFTTT) and Zapier in used to platforms like Excel and Google Sheets. By writing a simple arithmetic formula user can get connected with whole host of 3rd party programs. Eg User can post Tweets from a spreadsheet, look to see who are followers as well as connecting to AWS, Import.io and Tableau and many more.
- Blockspring is open source but to create user defined functions, an enterprise edition is also available in which user can supplement custom tags for easy search and discovery and set API tokens for whole organization at once.

7.2 Pentaho

Pentaho Data Integration combines and generates data to create a comprehensive depiction of business application that derive remarkable conclusions.. The platform delivers precise and visualized data to end users from any source. The graphical tool of Pentaho does not required any complex coding. Pentaho provides a platform to integrate big data and all data sources to the business and IT users using following features.

- Ease of use with the power to integrate all data: In-built drag-and-drop data integration attached with data connectivity ranging from flat files and RDBMS to Hadoop and beyond.
- Graphical extract-transform-load (ETL) designer to create straightforward formation of data pipelines
- Enriched with libraries of pre-built components to retrieve, organize, and composite data from relational sources, big data stores, enterprise applications, and more
Controlling composition abilities to synchronize and syndicate conversions, with announcements and alarms
- Agile views for demonstrating and imagining data on the glide during the data preparation process
- Integrated enterprise scheduler for organizing workflows and debugger for testing and regulation job implementation
- Big data integration with zero coding required: Compare to hand coding techniques using Pentaho's it is more rapidly design and deploy big data analytic reduce manual programming and scripting from the process by introducing Thorough graphic al big data

- Resilient provision for Hadoop distributions, Spark, NoSQL data stores and analytic databases
- Accessible data prep for a wider audience: Provide directed, on-demand data to analysts and end users in an agile fashion.
- Unified automated data integration solutions for integrating massive volumes and variety of data.
- To retrieve faster analytical insights it automatically creates metadata models.
- It also provides supports for progressive analytic models likem R, Python, and Weka.
- Enterprise platform to accelerate the data pipeline: Its functionalities are not limited to standard ETL to It also provides malleable management for end-to-end data flows.
- To Scale up and out data in cluster and cloud environment the architecture is enriched with multi-threaded data integration.
- It provides support for performance monitoring , job roll-back and restart option, auditing for usages.

References

- Datos.gob.es. (2013). *DATA PROCESSING AND VISUALISATION TOOLS*. Retrieved from <http://datos.gob.es/sites/default/files/files/2013-07-Visualization-Tools.pdf>
- The Apache Software Foundation. (2014, December 10). Welcome to Apache™ Hadoop®! Retrieved October 18, 2016, from <http://hadoop.apache.org/>
- Atlassian Confluence Open Source Project License. (2016, October 16). Home - Apache hive - Apache software foundation. Retrieved October 18, 2016, from <https://cwiki.apache.org/confluence/display/Hive/Home>
- BigML. (2016). BigML is machine learning for everyone. Retrieved October 18, 2016, from <https://bigml.com/features>
- Blockspring. Blockspring for enterprise - access your own business and data services – Blockspring. Retrieved October 18, 2016, from <https://www.blockspring.com/solutions>
- Chartio. Chartio provides simple, secure connections to the world’s most popular data sources. Retrieved October 18, 2016, from <https://chartio.com/product/data-sources/>
- GetApp. (2016, August 1). Chartio Feature Summary. Retrieved October 18, 2016, from <https://www.getapp.com/business-intelligence-analytics-software/a/chartio/features/>
- GitHub Pages. (2015, June). OpenRefine. Retrieved October 18, 2016, from <http://openrefine.org/>
- Herschel, G., Linden, A., & Kart, L. (2014). An Introduction to Advanced Analytic. *ADVANCED ANALYTICS PREDICTIVE ANALYTICS BUSINESS INTELLIGENCE DATA MINING*. Retrieved 18 October 2016, from <https://rapidminer.com/wp-content/uploads/2014/04/advanced-analytics-introduction.pdf>
- Han, J., Haihong, E., Le, G., & Du, J. (2011, October). Survey on NoSQL database. In *Pervasive computing and applications (ICPCA), 2011 6th international conference on* (pp. 363-366). IEEE.
- Hortonworks Inc. APACHE SQOOP. Retrieved October 18, 2016, from <http://hortonworks.com/apache/sqoop/>
- <https://www.tableau.com/stories/topic/big-data#>. Retrieved October 18, 2016, from <https://www.tableau.com/stories/topic/big-data#>
- Import.io. (2016). Import.io Create your own datasets in minutes, no coding required. Retrieved October 18, 2016, from <https://www.import.io/builder/>
- Machine Learning Group at the University of Waikato. Weka 3 - data mining with open source machine learning software in java. Retrieved October 18, 2016, from <http://www.cs.waikato.ac.nz/ml/weka/>
- Machine Learning Group at the University of Waikato. Weka 3 - mining big data with open source machine learning software in java. Retrieved October 18, 2016, from <http://www.cs.waikato.ac.nz/ml/weka/bigdata.html>
- Neopost All rights reserved. (2016). The premier open source data quality solution. Retrieved October 18, 2016, from <https://datacleaner.org/>
- Pentaho Corporation. (2005). Data integration. Retrieved October 18, 2016, from <http://www.pentaho.com/product/data-integration>
- Plotly. (2016). Plotly is the modern platform for agile business intelligence and data science. Retrieved October 18, 2016, from <https://plot.ly/#use-your-favorite-tools>
- Prajapati, V. (2013). *Big data analytics with R and Hadoop*. Packt Publishing Ltd
- Qubole. (2016). Introduction — Qubole data service 1.0 documentation. Retrieved October 18, 2016, from <http://docs.qubole.com/en/latest/user-guide/qds.html>
- Sarkar, D. (2014). *Pro Microsoft HDInsight: Hadoop on Windows*. Apress.
- Silk. What is silk? Retrieved October 18, 2016, from <https://www.silk.co/product>
- Statwing. (2013). Efficient and delightful statistical analysis software for surveys, business intelligence data, and more. Retrieved October 18, 2016, from <https://www.statwing.com>
- Tableau Software. (2016, August 18). With tableau 10, exploring big data just got even easier. Retrieved October 18, 2016, from <http://www.tableau.com/about/blog/2016/8/tableau-10-exploring-big-data-just-got-even-easier-58017>

- Take control of your R code. (2016, April 29). Retrieved October 18, 2016, from RStudio, <https://www.rstudio.com/products/RStudio/>
- The R Foundation. R: What is R? Retrieved October 18, 2016, from <https://www.r-project.org/about.html>
- Strohbach, M., Daubert, J., Ravkin, H., &Lischka, M. (2016). Big Data Storage. In *New Horizons for a Data-Driven Economy* (pp. 119-141). Springer International Publishing.
- Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1), 97-107.

