"Study and Analysis of Most Sought after Big data Platforms and their Application for Network Analytics"

Thendral N[1],Dr.Jayaramaiah[2],
[12]Dept of Information Science and Engineering, The Oxford College of Engineering, Bengaluru-68
natarajan.thendral@gmail.com[1], djr-hodise@theoxford.edu[2]

Abstract: In Today's world, we are getting technologically "connected", all over with more and more data devices which collect lots information. This has resulted in large quantities of data in the form of images, text, videos and multimedia content, log files, etc. Small and Medium Enterprises (SME) are facing number of problems in collecting, storing, analyzing and exploring these large volumes of data. A number of Big Data Platforms have taken advantage of Hadoop open-source framework and are providing some support to handle the so called Big data of the organizations, Cloudera ,HortonWorks, MapR ,IBM InfoSphere Big Insight ,Pivotal HD are a few Big data platforms currently available in the market. In this paper we carry out a comparative analysis of most sought after Big data platforms based on the operational, functional and performance characteristics of those platforms in general. We suggest that cloudera platform as the one which provides competitive advantage over the other platforms in terms of diagnostics, maintenance and performance analysis to be used as an acceptable tools for network Analytics.

Keywords:* *Big Data, Distribution Hadoop, Diagnostics, Network Analytics.

I. INTRODUCTION

Day after day, new innovations have delivered a lot of information that should be gathered, arranged, classified, moved, investigated, put away, etc. Currently, we are in the Big Data time in which a couple of distributors offer, arranged to-use spreads to manage a Big Data structure, To be particular Cloudera[2], Horton Works[1], MapR[3], IBM Infosphere Big Insights[4], and Pivotal HD[5] are the popular ones. The decision will be made on one or on the other arrangements as indicated by a few necessities. For instance, if the arrangement is open source, Maturity of the arrangement, and so on. A few releases have been supplemented with extra blocks, which make it conceivable to disentangle the task of the stages that retain parts complex due to the quantity of segments required. Accordingly, our work is to make a relative report on the fundamental Hadoop conveyance suppliers to characterize the qualities and shortcomings of every appropriation.

II. BIG DATA ARCHITECTURE

Before beginning with Big Data, one needs to ensure that all the fundamental segments of the design for breaking down all parts of a lot of information are set up. Engineering of a Big Data framework ought to have the capacity to explore the information sources in a quick and economical way. It ought to likewise have the accompanying layers: Data sources, Ingestion Layer, Visualization Layer, Hadoop Platform administration Layer, Hadoop Storage Layer, Hadoop Infrastructure Layer, Security Layer, and Monitoring Layer [11].

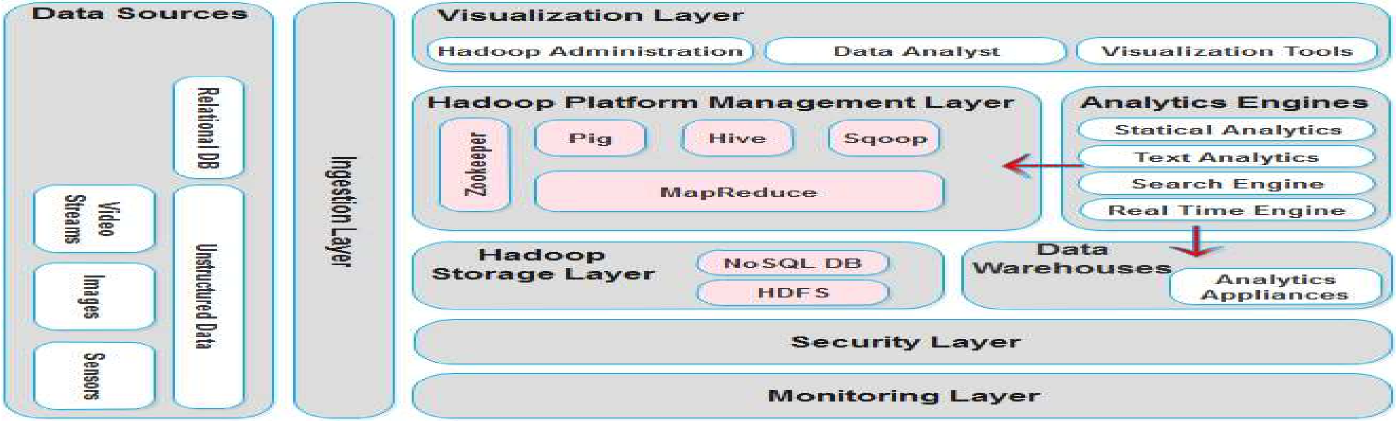


Figure 1: The Big Data architecture

This figure portrays the important segments of the engineering that ought to be a piece of a Big Data framework. It is important to pick open source or authorized structures to take full favorable position of the considerable number of highlights of the diverse segments of a Big Data framework [11].

III. *UNDERSTANDING OF BIG DATA DISTRIBUTION ARCHITECTURES*

In the midst of our investigation, we high light the structures of the particular spread Hadoop. Here there is the case of the five structures: Cloudera appropriation for Hadoop Platform, HortonWorks information platform, MapR Converged Data Platform, IBM Big information Platform, and pivotal HD business. The details are given in the succeeding paragraghs.

1. Cloudera Enterprise

Cloudera Enterprise is a superior minimal effort stage for directing investigation on information [1]. Cloudera Enterprise has the main local Hadoop Search motor and this stage furnishes its clients with dynamic information improvement highlight. Cloudera director incorporates propelled highlights like astute design defaults, modified observing, and powerful investigating which permit simple organization of Hadoop in any condition. Cloudera was right off the bat established by Hadoop specialists from Face book, Google, Oracle and Yahoo. This circulation is to a great extent in view of the segments of Apache Hadoop and it is supplemented by basically house segments for group administration. The point of Cloudera's plan of action isn't just to offer Licenses yet to offer help and preparing also. Cloudera offers a completely open source form of their stage (Apache 2.0 permit) [15].

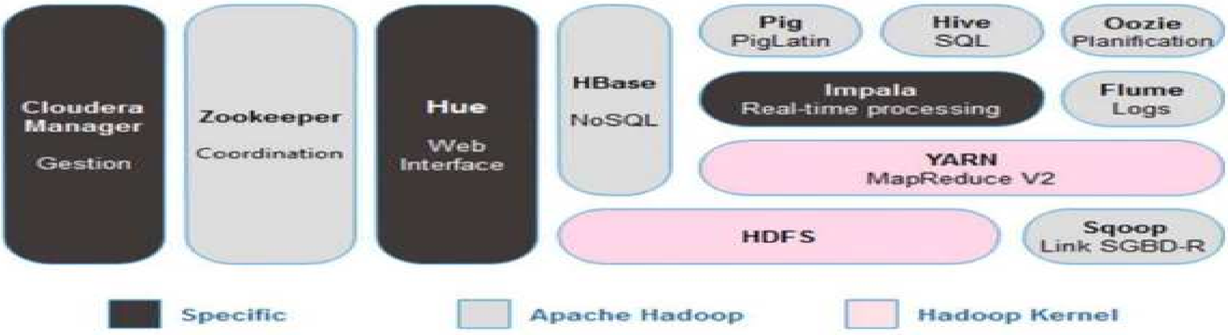


Figure2: Cloudera Distribution for Hadoop Platform (CDH)

2. HortonWorks Distribution

HortonWorks Distribution platforms(HDP) is the business' simply clear secure, undertaking arranged open source Apache™ Hadoop® scattering in light of a united plan (YARN). HDP watches out for the aggregate needs of data still, controls ceaseless customer applications and passes on capable enormous data examination that revive fundamental initiative and improvement [2].

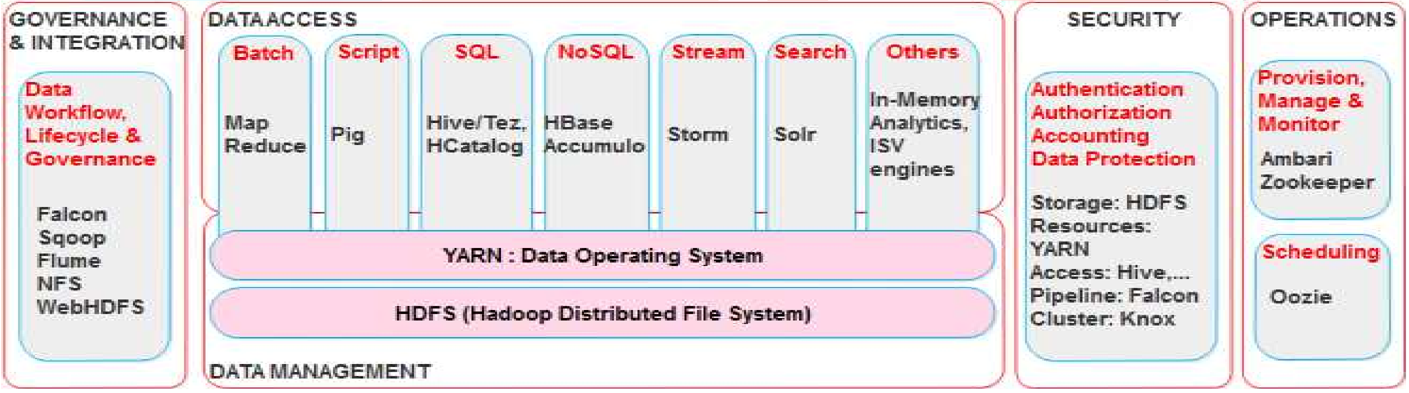


Figure 3: Horton Works Hadoop Platform (HDP)

Hortonworks Data Platform consolidates a versatile extent of taking care of engines that draw in one need to speak with comparable data in various courses, meanwhile. This infers applications for huge data examination can speak with the data in the best way: from gathering to insightful SQL[15] or low dormancy access with NoSQL. Creating use cases for data science, interest and spouting are also supported with Apache Spark, Storm and Kafka.

3. MapR Converged Data Platform

MapR Converged Data Platform is one single stage for enormous information applications. MapR's Platform depends on the idea of□ Polyglot Persistence which permits to use numerous information composes and organizes straightforwardly [2]. MapR, a merged information stage coordinates the energy of Hadoop and Spark with worldwide occasion gushing, continuous database capacities, and endeavor stockpiling, in this way empowering its clients to encounter the colossal energy of information [11].

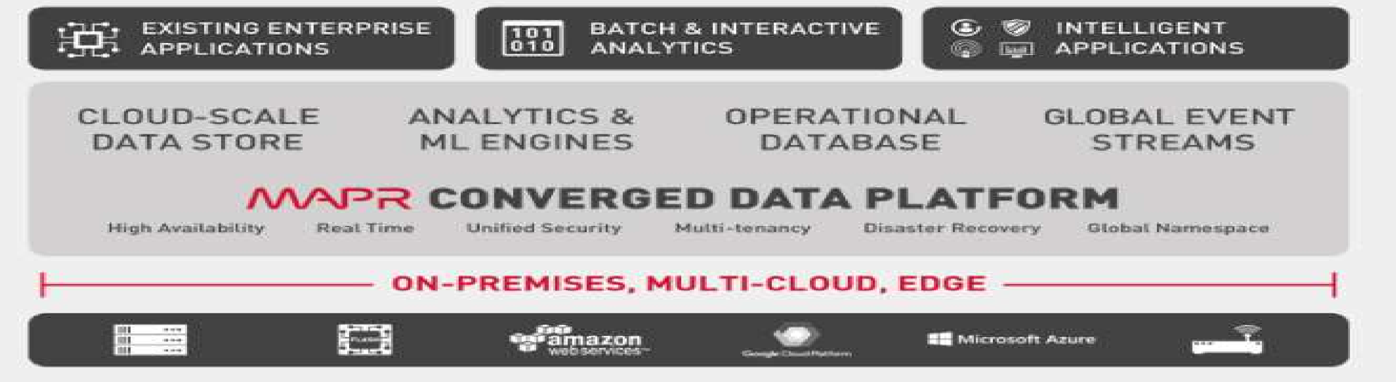


Figure 5:MapR Architecture

The MapR Converged Data Platform tackles the emergency of multifaceted nature that outcomes from persistently sending workload-particular information storehouses. Inside a solitary stage on a solitary codebase, it unites the key advancements that make up a cutting edge information design, including an appropriated record framework, a multi-display NoSQL database, a distribute/buy in occasion spilling motor, ANSI SQL, and an expansive arrangement of open source information administration and examination innovations [16]. The MapR Converged Data Platform conveys speed, scale, and unwavering quality, driving both operational and systematic workloads in a solitary stage.

4. **IBM InfoSphere**

Enormous Insights Distribution Info Sphere Big Insights for Hadoop was right off the bat presented in 2011 of every two forms: the Enterprise Edition and the fundamental adaptation, which was a free download of Apache Hadoop packaged with a web administration support. In June 2013, IBM propelled the Infosphere BigInsights Quick Start Edition [4]. This new version gave enormous information volume investigation abilities on a business-driven stage. It the two joins Apache Hardtop's Open Source arrangement with big business usefulness and henceforth, gives a huge scale investigation, portrayed by its versatility and adaptation to non-critical failure.In short, this distribution supports structured, unstructured and semi-structured data and offers maximum flexibility.

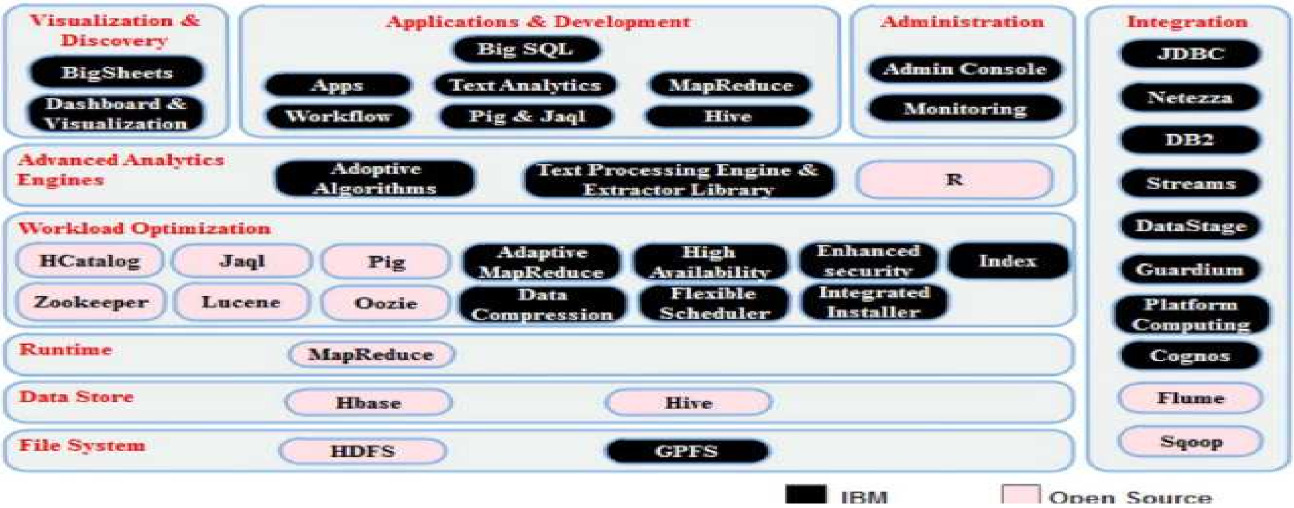


Figure6: IBM InfoSphere BigInsights Enterprise Edition

In spite of the fact that this condition incorporates a full Apache Hadoop stack, it is separated by various IBM segments that address the issues plot above [11]. In Big Insights Version 2.1, which ended up accessible in June 2013, these might be outlined as takes after:

5. Pivotal HD DistributionPivotal Software, Inc.

(Pivotal) is a product and administrations organization situated in San Francisco and Palo Alto, California, with separate all together workplaces. The divisions incorporate Pivotal Labs for counseling administrations, the Pivotal Cloud Foundry improvement gathering, and item advancement assemble for the Big Data advertise. Urgent HD Enterprise is an economically upheld dissemination of Apache Hadoop [5]. The figure underneath indicates how every Apache and Pivotal part incorporates into the general engineering of Pivotal HD Enterprise.

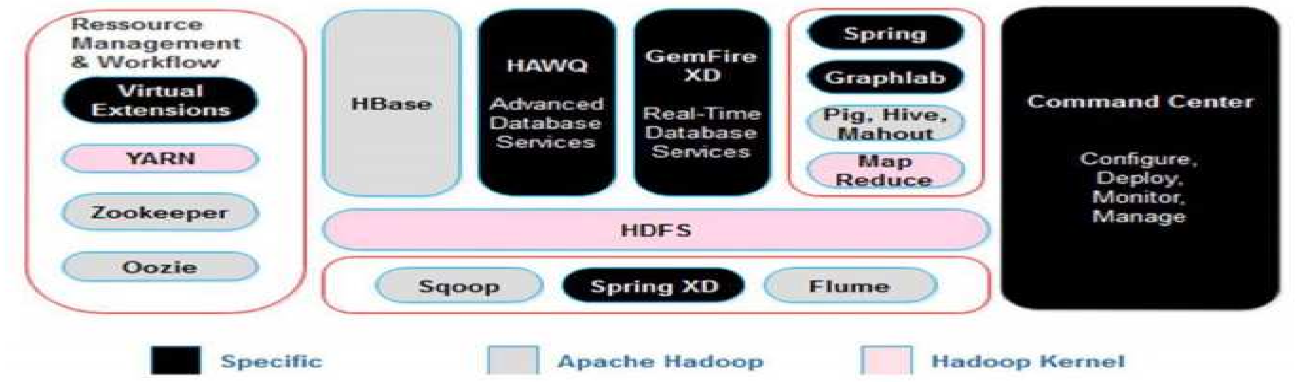


Figure 7: Pivotal HD Enterprise

Cloud Foundry doles out two sorts of VMs: the part VMs that constitute the stage's structure, and the host VMs that host applications for the outside world. Inside CF, the Diego structure passes on the encouraged application stack over the entire host VMs, and keeps it running and balanced through demand surges, power outages, or distinctive changes. To deal with request, various host VMs run duplicate events of a similar application [6]. Cloud Foundry passes on application source code to VMs with everything the VMs need to assemble and run the applications locally.

IV. COMPARATIVE ANALYSIS OF MOST SOUGHT AFTER BIG DATA PLATFORMS BASED ON THE OPERATIONAL, FUNCTIONAL AND PERFORMANCE CHARACTERISTICS

With a specific end goal to assess appropriations, we attempted to recognize the qualities and shortcomings of the five major Hadoop distribution providers: Cloudera, HortonWorks, IBM InfoSphereBigInsights, MapR, and Pivotal.

A. ***comparison based on Functional characteristics:***

Platforms → **↓Operational Characteristics**	**Cloudera**	**Horton Works**	**MapR**	**IBM**	**Pivotal**
Editor and Available Edition	• Express • Enterprise	Hortonworks Data Platform 2.5	• M3(free) • M5 • M7	• Quick Start • Standard • Enterprise	Pivotal Enterprise Edition
Administration Console	Cloudera manger	Ambari	MapR Control Systesm	Web Console	Command center
Software Components	• Cloudera Express • Cloudera Impala • Cloudera Search	• Zeppelin • Ambari User Views • DSX	MapR software.	• Big SQL • Big R • Adaptive MapReduce • IBM GPFS™ FPO	• Command Center, • Virtualization extensions and Isilon support
Ease of use	Powerful deployment, management and monitoring tools which are very much useful.	Very simple and easy-to-use sandbox which helps to getting started rapidly.	The most significant is the support for a native UNIX file system..	Anyone can download the IOP platform for free of charge or select a supported offering and use it on premises	By using Spring Hadoop tool male easy deployment

Product version Evaluated	Cloudera Enterprise: 5.50	Hortonworks Data platform: 2.30	The MapR Distribution including Apache: 4.10	IBM BigInsights for Apache Hadoop: 5.0	HadoopPivotal HD: 3.X

Table 1: Comparison based functional characteristics

The above table explains functional parameters of Available edition, Administration console, software components, ease of use and better manipulating facilities. The Cloudera platform provides better functional characteristics based on Apache Hadoop and projects effective use of open sources associated.

B. Comparison Based On Operational Characteristics:

Platforms → ↓ **Operational Characteristics**	**Cloudera**	**Horton Works**	**MapffR**	**IBM**	**Pivotal**
Open Source	Multiple version : Open source &Licensed	Open source	Licensed	Licensed	Multiple version : Open source &Licensed
Management Tools	Cloudera Manager	Ambari	MapR Control System	IBM Maxico Web console	Cloud Foundry
SQL Support	Impala	Stringer	Drill	IBMBig SQL	SQL
Market Presence	Highest score in market place Based on an evaluation compared to vendors	Next largest competitor with cloudera	Second highest current offering	This is also Strong competitor	Lowest score in market presence
Deployment	Deployement with Whirr toolkit.	Deployement with Ambari. Simple Deployment.	Through AWS Management Console.	IBM PureData System for Analytics.	BOSH and Ops Manager
Integration	Ease of integration using standard APIs and tools.	To ingest new data streams and additional volume as needed	Nagios integration and Ganglia integration.	Transforms data in any style and delivers it to any system.	Some tools available for integration.

Table 2: Comparison based on operational characteristics

In the above table contains the comparative aspects of the five chosen platforms of Big data based on operational characteristics. The main objective of this comparison is to criticize which is the one for quick and easy deployment and Integrations of various API's.

C. Comparison based on Performance Characteristics:

Platforms Functional Characteristics	Cloudera	Horton Works	MapR	IBM	Pivotal
Flexibility	Offer great flexibility and capability with their services	Apache Tez for interactive access and Apache Slider for long-running applications.	Offer flexibility to Works out of the box with no special configuration required.	flexible data analysis feature s apply to data in a variety of formats	Pivotal Cloud Foundry uses a flexible approach called buildpacks
Security	provide data encryption	provide data encryption	provides encryption of data transmitted to, from and within a cluster	Provides encryption and masking of confidential data.	Secret-key cryptography.
Scalability	They offer great flexibility and capability with their services in such a way that it makes managing our various applications	Needed more support from Hortonworks during implementation and running of platform	Scalable architecture without single points of failure	Highly scalable storage platform to store and distribute very large data sets.	Greenplum running on DCA delivers scalabil ity
High Availability	High Availability With a Load Balancer	Apache Hadoop 0.23.1 and HDFS NameNode high availability	High availability (HA) options for the NameNode and JobTracker.	For using HDFS replicated system based availability only.	Greenplum running on DCA delivers to assure availability and minimize downtime.
Data processing speed	With spark support Data processing, up to 100x in some cases.	Also working on improving computing speed. By using initiated Stinger	Apache Drill, a project backed by MapR to improving data processing speed	IBM InfoSphere Information Analyzer V8.1.1 provides efficient data processing speed.	HAWQ, a proprietary component able to process SQL-like queries 318x faster than Hive.

Table 3: Comparison based on performance characteristics

The above Table describes a few parameters of performance like Flexibility, Data Processing speed, Scalability, High Availability, and Security. After analysing above performance characteristics, we conclude that Cloudera platform will provide reasonably good results for network analytics in terms of availability and processing speed.

V. ANALYZING CLOUDERA DISTRIBUTION FOR NETWORK ANALYTICS:

Cloudera platform provides an investigation stage and the most recent open source innovations to store, process, find, model and serve a lot of information.CDH, the Cloudera Hadoop dissemination, incorporates a few related open source ventures, for example, Hive and Impala. It likewise gives security and coordination a few equipment and programming items [15].The Hive structure in Cloudera platform including Apache Hadoop enables clients to execute intuitive SQL questions straightforwardly against information put away in Hadoop Distributed File System (HDFS), Apache HBase or the Amazon Simple Storage Service.

A. General Architecture for Cloudera for Analytics:

Cloudera is a cutting edge programming arrangement composed particularly for information administration and investigation. The application offers what numerous specialists have marked as the world's speediest, least demanding, and most secure Apache Hadoop stage.

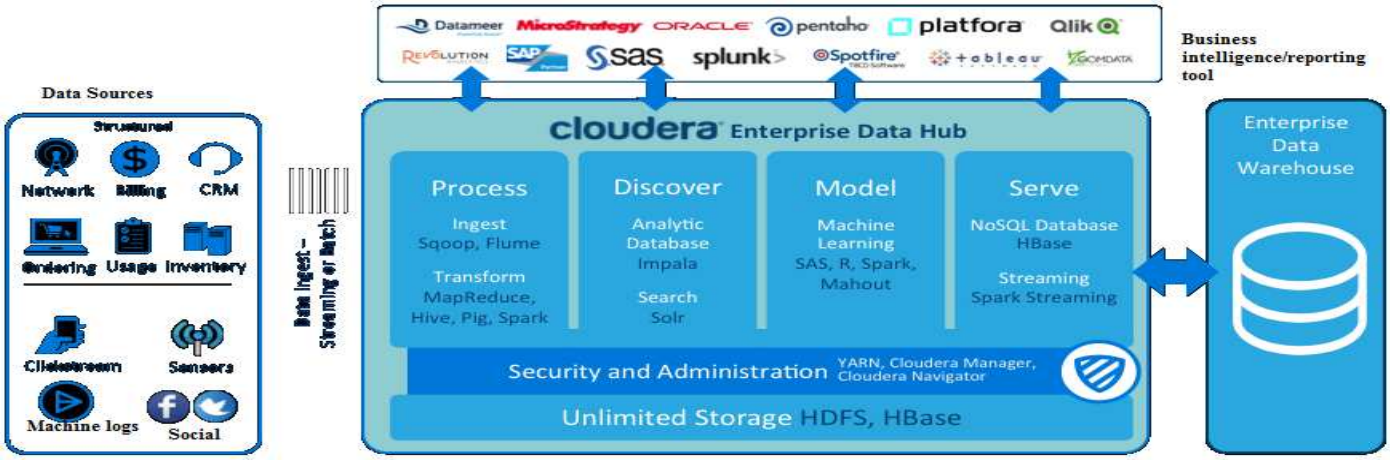


Figure 8: General Architecture for cloudera in analytics

With Cloudera Enterprise Data Hub (EDH), the framework changes the undertaking information administration scene by conveying the primary bound together stage for huge information [1]. The application gives ventures a solitary, bound together place to store, process, and break down every one of their information, engaging them to enhance the estimation of current speculations while empowering principal better approaches to get more an incentive from their information [8].

VI. IMPLEMENTING NETWORK ANALYTICS USING CLOUDERA DISTRIBUTION

CDH is the most total, tried, and mainstream dispersion of Apache Hadoop and related activities. CDH conveys the center components of Hadoop – versatile capacity and disseminated registering – alongside a Web-based UI and imperative venture abilities. CDH is Apache-authorized open source and is the main Hadoop answer for offer brought together group handling, intuitive SQL and intelligent inquiry, and part based access controls. Implementing network analytic by using following two tools, which is available in clouderaquickstart virtual machine [9].

- Apache Hive
- Hue

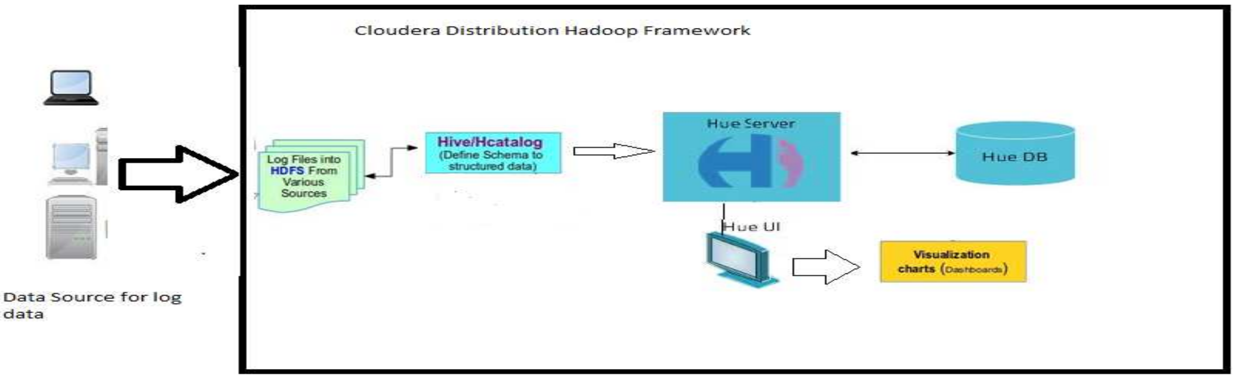
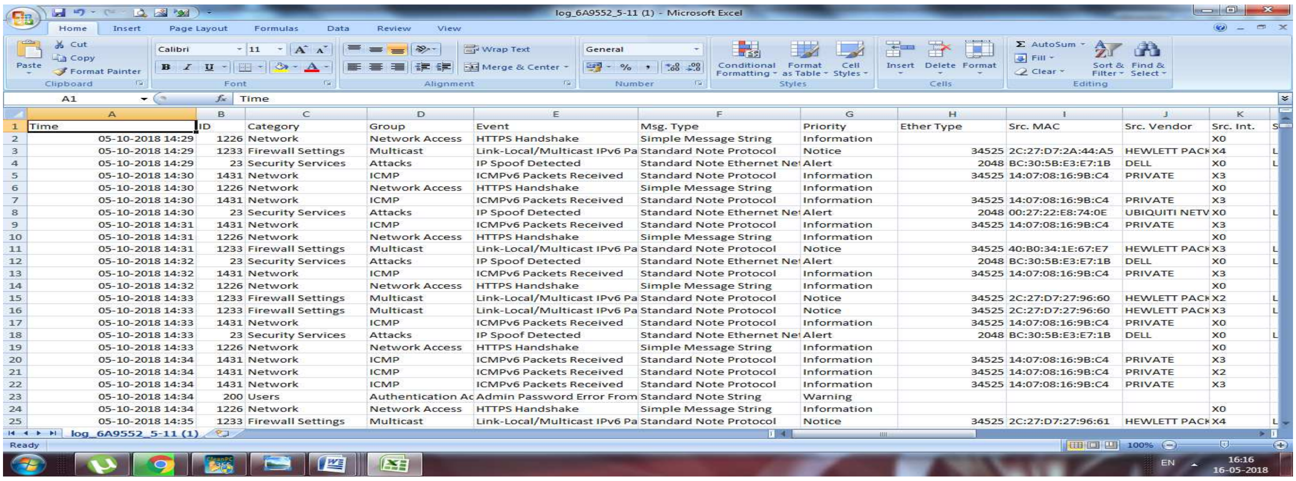

Cloudera Distribution Hadoop Framework
Data Source for log data
Log Files into HDFS From Various Sources
Hive/Hcatalog (Define Schema to structured data)
Hue Server
Hue DB
Hue UI
Visualization charts (Dashboards)

Figure9: Architecture of Network Analytics using cloudera

Logs are computer produced records that catch system and server activities data. They are helpful amid different phases of programming improvement, principally to debug and maintenance purposes and furthermore to manage arrange tasks. Here collecting log files from firewall system in terms of CSV file format. The sample log data for firewall system.

Sample log data for System alert event:



1	Time	ID	Category	Group	Event	Msg. Type	Priority	Ether Type	Src. MAC	Src. Vendor	Src. Int.	S
2	05-10-2018 14:29	1226	Network	Network Access	HTTPS Handshake	Simple Message String	Information				X0	
3	05-10-2018 14:29	1233	Firewall Settings	Multicast	Link-Local/Multicast IPv6 Pa	Standard Note Protocol	Notice	34525	2C:27:D7:2A:44:A5	HEWLETT PACK	X4	L
4	05-10-2018 14:29	23	Security Services	Attacks	IP Spoof Detected	Standard Note Ethernet Net	Alert	2048	BC:30:5B:E3:E7:1B	DELL	X0	L
5	05-10-2018 14:30	1431	Network	ICMP	ICMPv6 Packets Received	Standard Note Protocol	Information	34525	14:07:08:16:9B:C4	PRIVATE	X3	
6	05-10-2018 14:30	1226	Network	Network Access	HTTPS Handshake	Simple Message String	Information				X0	
7	05-10-2018 14:30	1431	Network	ICMP	ICMPv6 Packets Received	Standard Note Protocol	Information	34525	14:07:08:16:9B:C4	PRIVATE	X3	
8	05-10-2018 14:30	23	Security Services	Attacks	IP Spoof Detected	Standard Note Ethernet Net	Alert	2048	00:27:22:E8:74:0E	UBIQUITI NETV	X0	L
9	05-10-2018 14:31	1431	Network	ICMP	ICMPv6 Packets Received	Standard Note Protocol	Information	34525	14:07:08:16:9B:C4	PRIVATE	X3	
10	05-10-2018 14:31	1226	Network	Network Access	HTTPS Handshake	Simple Message String	Information				X0	
11	05-10-2018 14:31	1233	Firewall Settings	Multicast	Link-Local/Multicast IPv6 Pa	Standard Note Protocol	Notice	34525	40:B0:34:1E:67:E7	HEWLETT PACK	X3	L
12	05-10-2018 14:32	23	Security Services	Attacks	IP Spoof Detected	Standard Note Ethernet Net	Alert	2048	BC:30:5B:E3:E7:1B	DELL	X0	L
13	05-10-2018 14:32	1431	Network	ICMP	ICMPv6 Packets Received	Standard Note Protocol	Information	34525	14:07:08:16:9B:C4	PRIVATE	X3	
14	05-10-2018 14:32	1226	Network	Network Access	HTTPS Handshake	Simple Message String	Information				X0	
15	05-10-2018 14:33	1233	Firewall Settings	Multicast	Link-Local/Multicast IPv6 Pa	Standard Note Protocol	Notice	34525	2C:27:D7:27:96:60	HEWLETT PACK	X2	L
16	05-10-2018 14:33	1233	Firewall Settings	Multicast	Link-Local/Multicast IPv6 Pa	Standard Note Protocol	Notice	34525	2C:27:D7:27:96:60	HEWLETT PACK	X3	L
17	05-10-2018 14:33	1431	Network	ICMP	ICMPv6 Packets Received	Standard Note Protocol	Information	34525	14:07:08:16:9B:C4	PRIVATE	X0	
18	05-10-2018 14:33	23	Security Services	Attacks	IP Spoof Detected	Standard Note Ethernet Net	Alert	2048	BC:30:5B:E3:E7:1B	DELL	X0	L
19	05-10-2018 14:33	1226	Network	Network Access	HTTPS Handshake	Simple Message String	Information				X0	
20	05-10-2018 14:34	1431	Network	ICMP	ICMPv6 Packets Received	Standard Note Protocol	Information	34525	14:07:08:16:9B:C4	PRIVATE	X3	
21	05-10-2018 14:34	1431	Network	ICMP	ICMPv6 Packets Received	Standard Note Protocol	Information	34525	14:07:08:16:9B:C4	PRIVATE	X2	
22	05-10-2018 14:34	1431	Network	ICMP	ICMPv6 Packets Received	Standard Note Protocol	Information	34525	14:07:08:16:9B:C4	PRIVATE	X3	
23	05-10-2018 14:34	200	Users	Authentication Ac	Admin Password Error From	Standard Note String	Warning					
24	05-10-2018 14:34	1226	Network	Network Access	HTTPS Handshake	Simple Message String	Information				X0	
25	05-10-2018 14:35	1233	Firewall Settings	Multicast	Link-Local/Multicast IPv6 Pa	Standard Note Protocol	Notice	34525	2C:27:D7:27:96:61	HEWLETT PACK	X4	L

Figure10: server status logs for firewall

Above log data are given as the input for our application in cloudera. The log data are having more than 1000 records for analysis. The records are store in the format of CSV file, and then it will be transferred to HDFS file location in /home/cloudera.

Hive Tool:

Hive tool used to create databases for analytical purpose. In analytical application server status logs data's are taking as the dataset to create tables. The following example use to create table for firewall data [15].

Example:

create table eventlog (eventstring,Src_ipstring,IP_PROTOCALstring,Msg string……) row format delimited fields terminated by',';

After creating table need to transfer data into table by using hive query.

load data localinpath '/home/cloudera/evenlog.csv' into table eventlog;

In hive tool we can query the database table for our network analysis basis. It will produce the data according to time taken for analysis the data. By using hive connectivity tools, visualize analytical data in graphs and charts.

Hue tool:

Hue is a web-based interactive query editor in the Hadoop stack that will helpful visualize and share data [15].

- Editor

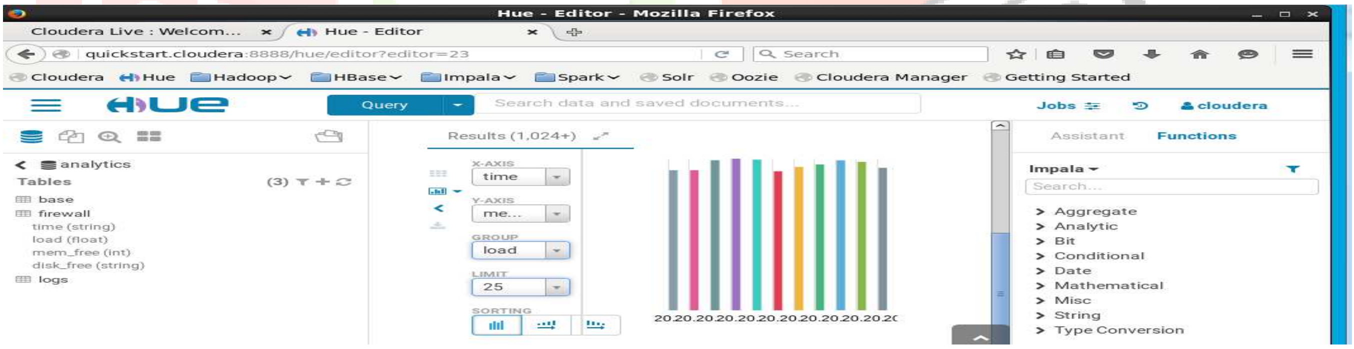


Figure 12: Graphs for System error status

Therefore overview of cloudera Distribution for Network Analytics efficiently analyzed huge record with graphical manner.

The objective of Hue's Editor is to make information questioning simple and gainful. It centers around SQL yet additionally bolsters work entries. It accompanies a shrewd auto finish, seek and labeling of information and question help.

VII. CONCLUSION AND FUTURE WORK

Many of the Big data platforms, and architecture frameworks differ in terms of their approach and level of details. Some are just proposed guidelines, whereas others have specific methodologies and critical aspects to follow. The majority of the platforms are abstract and generic in nature and hence the ability to work accurately is questionable. In this paper we analyzed a few open source Big data platforms like Cloudera, Horton Works, MapR, IBM and Pivotal. Our evaluation is based on both subjective measures like the ease of use and objective measures like the performance of each distribution, enabling users to make more informed decisions. According to our evaluation Cloudera offers additional management software as part of the commercial distribution so that Hadoop Administrators can configure, monitor and tune their hadoop clusters. Integrating the tools with Cloudera platform, will give best form of diagnostics and performance analysis. This is important to identify network failure and maintenance issues on prediction basis. Our future work is to expand this research to include more complex network analysis as well as multidimensional data to assist faster, diagnostics and improved performance.

References

[1] HortonWorks Data Platform HortonWorks Data Platform: New Book. (2015).
[2] Menon, R. (2014). Cloudera Administration Handbook
[3] Dunning, T., & Friedman, E. (2015). Real-World Hadoop
[4] Quintero, D. (n.d.). Front cover implementing an IBM InfoSphereBigInsights Cluster using Linux on Power.
[5] Pivotal Software, I. (2014). Pivotal HD Enterprise Installation and Administrator Guide.
[6] Sarkar, D. (2014). Pro Microsoft HDInsight. Berkeley, CA: Apress.
[7] ThibaudChardonnens, “Big Data analytics on high velocity streams: specific use cases with Storm”, Software Engineering Group, Department of Informatics, University of Fribourg, Switzerland, 2013.
[8] McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity. Paper, June 2011. 7, 9, 10, 11
[9] Nauman Sheikh, “Big Data, Hadoop, and Cloud Computing, Implementing Analytics”, Morgan Kaufmann, 2013.
[10] C. Dobrea, and F. Xhafa b, “Intelligent services for Big Data science”, Future Generation Computer Systems, Volume 37, 2014, pp. 267-281.
[11] Sawant, N., & Shah, H. (Software engineer). (2013). Big data application architecture &amp; Aa problem-solution approach. Apress.
[12] Lenovo, I. (2015). Lenovo Big Data Reference Architecture for Cloudera Distribution for Hadoop, (August).
[13] Read, W., Report, T., & Takeaways, K. (2016). The Forrester WaveTM: Big Data Hadoop Distributions, Q1 2016.
[14] Gates, Alan, and Daniel Dai. Programming Pig: Dataflow Scripting with Hadoop. 2 edition. O’Reilly Media, 2016.
[15] Capriolo, Edward, Dean Wampler, and Jason Rutherglen. Programming Hive: Data Warehouse and Query Language for Hadoop.1 edition. Sebastopol, CA: O’Reilly Media, 2012.
[16] Ting, Kathleen, and JarekJarcecCecho. Apache Sqoop Cookbook: Unlocking Hadoop for Your Relational Database. 1 edition. Sebastopol, CA: O’Reilly Media, 2013.
[17] Murthy, Arun, Vinod Vavilapalli, Douglas Eadline, Joseph Niemiec, and Jeff Markham. Apache Hadoop YARN: Moving beyond MapReduce and Batch Processing with Apache Hadoop 2. 1 edition. Upper Saddle River, NJ: Addison-Wesley Professional, 2014.