International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 8 Issue 02 Series. II || Feb 2019 || PP 81-85

A Survey on the usability of available Big Data platforms and its industry trends

Akshaya Kumar Satpathy[1], Sitanath Biswas[2],Mohini Prasad Mishra[3]
[1,2]*Associate Professor, Department of Computer Science Engineering, Gandhi Institute For Technology (GIFT), Bhubaneswar*
[3] *Assistant Professor, Department of Computer Science Engineering, Gandhi Engineering College, Bhubaneswar*

Abstract: *In today's Knowledge economy, Big Data has become the de-facto technology for organization for analysis, prediction, decision making, customer delight and growth. The need for quick understanding and delivering the solution based on the problem dimension is continuously evolving. The selection of the right platform for the right kind of problem is the key to success. Organization is spending time, money and enormous effort in identifying the right kind of platform. This paper provides a comprehensive analysis and assessment of the available platforms to aid the researchers and developers to select the right kind of platform based on the problem domain.*
Keyword: *Big Data, Analytic, Platform, Technology,*

I. Introduction:

"Big Data" is behemoth quantity of data. Data, which is unorganized, discrete, unsorted, but in extremely large volume. Conventional mechanisms are not able to cache them, prepare them and control them.

A. Definition of Big Data

Big Data has evolved and has become the buzz word for all. It refers to the huge collection of data with complexity which the traditional data processing application is inadequate to handle it. The need for collection, collaboration, classifying and calculation in faster and efficient way, has become the need to identify the hidden trends for quick business decision to stay ahead in the competitive market. There are several Big Data platforms available with different characteristics and choosing the right platform requires an in-depth knowledge about the capabilities of all these platforms. [1] Today more than 2.5 Exabytes of data is generated every single day. Organizations are producing data at an astounding rate. Facebook alone collects 250 terabytes a day. According to Thompson Reuters News Analytics, digital data production has more than doubled from almost one zettabyte (a zettabyte is equal to 1 million petabytes) in 2009, and it is expected to reach 7.9 zettabytes in 2015, being estimated to reach 35 zettabytes in 2020. [2]

B. Characteristics of Big Data

i) Volume:-Volume means a huge data generated in any domain.
ii) Velocity:-With the increasing use of technology and upgrading of computer networks (such as 4G speed of internet), data is arriving at high speed.
iii) Variety:-Data coming from heterogeneous resources. With the increase of different types of tech devices data is generating from different sources.(for example cloud computing and different small computing devices)
iv) Veracity:-Big Data is combined of all kind of data such as precise, imprecise, accurate, inaccurate form of data
v) Value:-Value refers to data which needs to be extracted from different resources.
vi) Complexity:-Data management can become a very complex process, especially when volume of data comes from multiple resources.

C. Where Big Data is helping us?

- Based on our business requirement it will help to take right decision.
- Big Data analytics help to come up with correct prediction.
- Big Data helps to improve security measures in networks.
- Recalculate entire risk portfolios in minutes.
- Quickly identify customers who matters the most.
- Use click stream analysis and data mining to detect fraudulent behaviour.

D. Existing Platforms

- JAVA in Apache Hadoop
- R-Language in Apache Hadoop
- Python

- Spark in Apache Mesos, Apache Hadoop
- Scala in Apache Spark
- Hadoop Ecosystem:
- Flume: log files
- Sqoop: Traditiondal database system
- Pig: high level language and execution environment
- Hive: is data warehouse software which provides query language similar to SQL called HIVEQL which manage querying over datasets.

Applications of Big Data

Big Data is providing an efficient way in predicative models around individual patients. US healthcare system alone already reached 150 exabytes five years ago. Before long, we will be dealing with zettabyte and yottabyte data for countries with large populations such as China and India. This has led to better diagnosis of diseases from DNA, proteins, cells, tissues, organ and treatment. Extensive research in transforming big genome data to diagnostic, therapeutics and imparting anew insights in disease treatment. The specific applications of Big Data in medical application are Genomics Analytics, Flu outbreak Prediction and Control, Clinical Outcome Analytics.[3] A single whole human genome sequencing is 3-200 GB depending on the depth of coverage.Gene sequencing has able to provide personalized health program and many companies has started providing services. Real-life examples such as electronic medical records (EMRs), which are already a successful model in Denmark, prove that the right analysis and application of Big Data in healthcare facilitates improvement in patient care and outcomes. [4] IBM is one of the leading vendors in offering analytics in healthcare solutions. OptumHealth, Oracle, Verisk Analytics are also offering health care solutions. McKinsey projects that the use of Big Data in healthcare can reduce the healthcare data management expenses by $300 billion -$500 billion. David Cameron, then Prime minister of UK in collaboration with the American Biotechnology firm Illumina and Genomics England. has announced a government funding of €300m in August, 2014 for a 4 year project targeting to map 100,000 human genomes by the end of 2017 The main goal of this project is to make use of Big Data in healthcare to develop personalized medication for cancer patients. Big Data is becoming a powerful factor in redefining intelligent, reconnaissance and surveillance in wars and combating terrorist strike. Big Data has applications in Intelligence development, Knowledge management, Decision making, Cyber defence, forensics, warfare methodologies. One of the most important steps of the military decision making process is that accurate and timely analysis of information about enemy. [5]

In manufacturing, data analytics can process the historical process data to identify patterns and relationships to reveal important insights like predict future events, foresee risks, understanding the value chain etc in manufacturing units. Oracle provided the top areas where Big Data can make a difference : Forecasting of products and production, faster service to customers, real time decision making, supplier performance and better interactions. Manufacturers can generate value using Big Data , however in reality yet few of them are close to the Big Data vision.

Exploring the Existing Platforms:

Understanding the right kind of platform for a problem domain is essential. The requirement is profound understanding of the platform.

By December 2015, according to Gartner, there are seven vendors offering Big Data platforms: Hadoop, Amazon, Cloudera, Hortonworks, IBM, MapR, Pivotal and Transwarp. These vendors are adding additional capabilities such as event stream processing, meta data integration, security and governance through partnership and managed development efforts. The open source Apache Hadoop has contributed the dimension of Big data. Hadoop major benefit is the reliability.

Cloudera and Hortonworks based on the Apache Hadoop are the products to start as it is easy to setup and use. Cloudera is today the market leader having a galactic user base and clients. The GUI features of Cloudera helps to diagnosis the problem and manage the clusters. Hortonworks was founded in 2011 and few years it has become one of the leading vendors of Hadoop. It is free of cost and can be downloaded easily. MAP-R is a open source edition of Apache Hadoop that uses its own file system called MapRS. MapRS provides efficient data management, reliability and easy to use. The Hewlett Packard Enterprise's Big data Platform Vertica has advanced SQL database analytics catering to the need of an organization in managing and analyzing data quickly and reliably. Vertica support the standard programming interface ODBC, JDBC, ADO.NET and OLEDB. IBM is relatively new in Big Data. Its Big Data platform has four capabilities - Hadoop based analytics, stream computing, data warehousing and information integration and governance. SAP HANA has been updated with support for streaming data focusing on the real time notification. Microsoft implements Hadoop-based Big Data solutions using the Hortonworks Data Platform (HDP). HDInsight is a cloud based

service available to Azure clusters to run HDP and integrates with Azure storage. Intel has recently turns its attention to Hadoop. It has added the company's Graph Builder and Analytics Toolkit functions to Hadoop.
The ten most common problems in Hadoop are Risk Modeling, Customer Cum Analysis, Recommendation Engine, Ad Targeting, Point of Sale Transaction Analysis, Predict Node failure, Threat analysis, Trade Surveillance, Search Quality, Data Sandbox.
The programming model that is used in Hadoop is MapReduce. One of the major drawback of MapReduce is its inefficiency in running iterative algorithm.[6]

Spark has been growing as one of the leading open source Big Data Framework. Spark in certain circumferences 100 times faster than Hadoop, but it doesn't provide the own distributed storage system. Although there has been not a straight forward pattern to select either spark or Hadoop as there are lots of dependencies like data, working tandem, pattern identification and perception. However there have been many projects where Spark as been installed on top of Hadoop, so that Spark advanced analytics can be made to use using the HDFS. Spark capability of handing real time stream processing like recommendation engines used by retailers, or monitoring the performance of industrial machinery in the manufacturing industry and own machine learning libraries called MLibs, an edge over Hadoop. An application consisting of colossal structured data, don't require machine learning capabilities and advanced streaming analytics, then Hadoop is itself is sufficient.

Both Hadoop and Spark are scalable. Yahoo has over 42,000 thousand Hadoop Cluster whereas the largest known spark cluster is 8,000. Hadoop provides Service Level Authorization thus ensuring the clients with the right permissions. Spark security is bit thin, however if runs over Hadoop enjoys its security.

In the paper, we are analyzing Big Data Framework providers - Cloudera vs Hortonworks and open source framework Hadoop vs Spark.

Comparison of the Cloudera and Hortonworks

Cloudera and Hortonworks : The Similarities

The core of both Cloudera and Hortonworks is Aapche Hadoop, master-slave architecture and offers enterprise -ready Hadoop distributions. Both provides paid training, services to the newcomers, have establish community to discuss the problem faced.

Cloudera vs. Hortonworks: The Differences

The main difference is on the long term goal. Cloudera aims in becoming an enterprise data hub whereas Hortonworks has partnered with data warehousing company Teradata. Cloudera comes with 60 days trail version while Horton is open source. Cloudera has a larger customer base than Hortonworks. Hortonworks lacks in propriety software as it is completely open source.

Comparison of the Hadoop Map Reduce vs Spark

Features	Hadoop Map Reduce	Spark
Difficulty	Map Reduce is difficult to program and needs the abstractions	Spark is easy to program and doesn't require any abstractions
Interactive mode	There is no inbuilt interactive mode expect Pig and Hive	It has interactive mode
Data Streaming	It is used for generating reports that help in finding the answers to historical queries.	Possible to perform streaming. Batch processing and machine learning all in the same cluster.
Performance	It doesn't leverage the memory of the Hadoop cluster in Maximum	Execute batch processing jobs about 10 to 100 times faster than Hadoop MapReduce.
Latency	Is Disk Oriented completely	Ensures lower latency computations by caching the partial results across its memory of distributed

		workers.
Coding	Writing Hadoop MapReduce pipelines is complex and lengthy process	Writing spark code is always compact than writing Hadoop MapReduce code.

Spark is approximately 2.5x, 5x, and 5x faster than MapReduce, for Word Count, k-means, and PageRank, respectively. [7]

Fig 1 : Hadoop EcoSystem

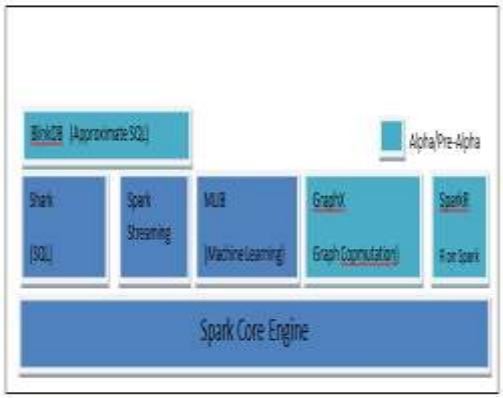

Fig 2 : SparkCore Engine

Trends in Industry :

The choice of Hadoop and Spark depends on the user based case. Spark is a very strong contender and would definitely bring about a change by using in memory processing, Spark will be the de facto framework for a large number of use cases involving Big Data processing. [8] IBM Research Zurich is working on Spark for RDMA; Swisscom has heavily embarked on Spark especially for training. Databricks survey suggest that Spark's growth continues across various industries, building sophisticated data solutions by people in various functional roles. Today, there are over 1000 Spark contributors, compared to 600 in 2015 from 250+ organizations. The key industries using spark are Software, Consulting, Banking, Advertising and Marketing and E-commerce. [9]. According to syncsort, a growing number of companies will look to leverage Hadoop for advanced use cases. It predicts adopting a "Hadoop first" approach to data management – skipping traditional and more expensive platforms.

With Hadoop gaining more traction in the enterprise, there will be a growing demand from end users for the same fast data exploration capabilities they've come to expect from traditional data warehouses.To meet that end-user demand, adoption of technologies such as Cloudera Impala, AtScale, Actian Vector and Jethro Data that enable the business user's old friend, the OLAP cube, for Hadoop will grow – further blurring the lines behind the

"traditional" BI concepts and the world of Big Data.[10]. After years of rapid technology-focused adoption of Hadoop and related alternatives to conventional databases, we will see a shift toward more business-focused data strategies. These carefully crafted strategies will involve chief data officers (CDOs) and other business leaders, and will be guided by innovation opportunities and the creation of business value from data.[11] In a recent survey of 2,200 Hadoop customers, only 3% of respondents anticipate they will be doing less with Hadoop in the next 12 months. 76% of those who already use Hadoop plan on doing more within the next 3 months and finally, almost half of the companies that haven't deployed Hadoop say they will within the next 12 months. [13] The growth in data warehouse has been slowing down.

Gartner has predicted 5 Big Data technologies in 2016 - Smart Machines, Customer Digital Assistants, Internet of Things, Automated Composition Machines, Robo-Boss. This will mark the beginning of new era for Big data analytics.Apache Spark is moving from mere component of Hadoop ecosystem to a big data platform choice for many enterprises.

The Hadoop Big Data analytics market is projected to grow from USD 6.71 Billion in 2016 to USD 40.69 Billion by 2021, at a CAGR of 43.4% between 2016 and 2021 [14]

II. Conclusion:

The paper render a comparison of few of Big Data providers keeping in mind the evolving trends in the industry. Big Data is continuously evolving and penetrating the diverse domains and requirements. Big Data providers are aware of these facts and spending time and effort in giving the best to become slowly the integral part in problem -solving organization.

References

[1]. Agneeswaran Vijay Srinivas, Tonpay Pranay, and Tiwary Jayati. Paradigms for Realizing Machine Learning Algorithms. Big Data. January 2014,1(4):207-214. doi:10.1089/big.2013.0006,

[2]. https://www.infoq.com/news/2013/12/ HadoopUsage

[3]. Lidong Wang and Cheryl Ann Alexander, Department of Engineering Technology, Mississippi Valley State University, USA, Department of Nursing, University of Phoenix, USA "Big Data in Medical Applications and Health Care" - American Medical Journal.

[4]. Dr. Peter Walther is Director of Business Development and Communication for Elsevier Health Analytics. https://www.elsevier.com /connect/ how-big-Data-can-revolutionize -patient-care.

[5]. H.Çintiriz, M.N.Buhur, and E.Şensoy "Military Implications of Big Data" , ICMSS

[6]. Dilpreet Singh and Chandan K. Reddy."A survey on platforms for Big Data analytics"

[7]. Satish Gopalani and Rohan Arora "Comparing Apache Spark and Map Reduce with Performance Analysis using K-Means" International Journal of Computer Applications (0975 – 8887) Volume 113 – No. 1, March 2015

[8]. Juwei Shi, Yunjie Qiu , Umar Farooq Minhas , Limei Jiao, Chen Wang , Berthol, Reinwal, and Fatma Ozcan "Clash of the Titans: MapReduce vs. Spark for Large Scale Data Analytics"

[9]. Apache Spark Survey 2016 Report Highlight , databricks.com/2016-spark-survey

[10]. https://msdn.microsoft.com/en-us/library/ dn749804.aspx

[11]. http://www.information-age.com/top-8 -trends-big-data-2016-123460615/

[12]. https://www.thoughtworks.com/insights/ blog/nine-hottest-data-trends-2016

[13]. http://www.tableau.com/sites/default/files/media/top8bigdatatrends2016_final_2.pdf?ref=lp&signin=a09f463432f61476722fc85 ddd7fa7a6

[14]. http://www.marketsandmarkets.com/ Market-Reports/hadoop-market-766.html