International Journal of Multidisciplinary Approach and Studies

ISSN NO:: 2348 – 537X

Analyzing Big Data Tools and Deployment Platforms

Pramila Joshi*

***Assistant Professor, Department of Computer Science, Birla Institute of Technology, NOIDA, Uttar Pradesh**

ABSTRACT:

Big data is a latest trend in today's time which is growing exponentially and is very much in demand of smart management. With the advent of the Internet, coupled with complete democratization of content creation and distribution in multiple formats, data is exploding like anything. Not only it is big, both in terms of volume and variety, but it has a velocity component to it as well. It is interesting as well as exciting to be able to extract the nuggets of information embedded in such a huge pool of data, at precisely the time of need. We are migrating to another evolution, popularly called as Big Data. Organizations seeking to find a better way to tap into the wealth of information hidden in this explosion of data around them to improve their competitiveness, efficiency, insight, profitability, and more, to gain them an edge over their competitors. This is the realm of "Big Data." While many companies appreciate that the best Big Data solutions, only a few have figured out how to proceed. In reality, the best Big Data solutions will also help organizations to know their customer better than ever before. It was not easy to select a few out of many Open Source projects. It is a task to choose the ones that fit Big Data's needs most. A new trend in the world of Open Source is that the big players have become stakeholders now for example IBM has done alliance with Cloud Foundry, Microsoft is providing a development platform for Hadoop, Dell is giving Open Stack-Powered Cloud Solution, EMC with VMware are partnering on Cloud, Oracle has released its NoSql database as Open Source. To address these business needs, this survey paper explores various tools to approach this modern problem. The paper diligently describes the challenges of harnessing Big Data and provides examples of Big Data tools and solutions that deliver tangible business benefits.

Key Words : Big Data, Cloud, Hadoop, Business Intelligence, Map Reduce

INTRODUCTION: WHAT IS BIG DATA?

Big Data…it's a new trend today which is varied, growing and moving very fast, and is very much in need of smart management. Today, Data and cloud are energizing organizations across multiple industries and present an enormous opportunity to make organizations more agile, more efficient and more competitive. To capture that opportunity, organizations require a modern Information Management architecture.

Big Data is the latest buzzword which is used to describe a huge volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques. In most organizations the data is too big or it moves too rapidly that it exceeds current processing capacity. Big data can greatly help companies improve operations and make quick, more intelligent decisions. [1]

EXPLAINING BIG DATA:

The world of Big Data is increasingly being defined by the 4 Vs. i.e. these 'Vs' become a reasonable test as to whether a Big Data approach is the right one to adopt for a new area of analysis. These 4 Vs are:

Volume:

The size of the data: This aspect refers to the fact that the amount of generated data has increased tremendously the past years. The quantity of data that is generated is very important in this context. It is the volume of the data which determines the value and potential of the data under consideration and whether it can actually be considered Big Data or not. The name 'Big Data' itself contains a term which is related to size and hence the characteristic. [2] For some companies this might be 10's of terabytes, for others it may be 10's of peta bytes. [3]

Velocity:

The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development. This aspect captures the growing data production rates. There is lots of data being produced and must be collected in shorter time frames. The rate at which data is being received and has to be acted upon is becoming much more real-time. [3]

Variety:

The next aspect of Big Data is its variety. With the multiplication of data sources comes the explosion of data formats, ranging from structured information to free text. Variety of data being processed is becoming increasingly diverse. Gone are the days data sets had to deal with traditional data like Documents, Stock record, personal files, finances etc. Today a variety of data like Photographs, Audio and Video, 3 D models, Simulations, Locations data are being piled. Many such data sources are also unstructured and hence not easy to categorize and process with traditional computing techniques.

Value:

This highly subjective aspect refers to the fact that until recently, large volumes of data were recorded (often for archiving or regulatory purposes) but not exploited. We need to consider what commercial value any new sources and forms of data can add to the business. [4] The understand ability and management of these sources, the Vs previously described, and then integrate them into the larger Business Intelligence system can provide valuable insights from data and this understanding leads to the "4th V" of Big Data – ***Value***. There is a vast opportunity offered by Big Data technologies to discover new insights that can lead to significant business value. Industries are seeing impact of data in the market and have started reinventing themselves as "data companies", as they feel that information has become their biggest asset. [5]

BIG DATA: SINCE WHEN:

The question arises whether Big Data is a recent trend? Not exactly. Though there is a lot of hype around the topic, big data has been here a long time. Think back of times when you first heard of scientific researchers using supercomputers to analyze large volumes of data. The difference now is that big data is accessible to regular BI users and is applicable to the enterprise. The reason it is drawing attention is because there are more public use cases about companies getting real value from big data (like Wal-Mart analyzing real-time social media data to analyze trends and then using that information to guide online purchases).IDC has determined that the big data technology and services market was worth $3.2B USD in 2010 and is going to skyrocket to $16.9B by 2015.

The big data trend promises that controlling the wealth and volume of information in your enterprise leads to better customer insight, operational efficiency, and better competitive edge. The marketing boost around big data and the pace of research studies, analyst reports, and articles on the subject can be mind startling for companies that want to take advantage of big data analytics but do not know how to separate fact from fiction and determine real use cases for their business. So here's big data elementary information for those just getting in the game.

BIG DATA: FROM WHERE IS IT COMING?

The quantity of computing data generated on planet earth is growing exponentially for many reasons:

- Retailers are building vast databases for recording customer activity.
- Organizations working on logistics, financial sector and health sectors are also capturing more and more data.
- Public Social media like face book, twitter, LinkedIn, YouTube is also creating vast quantities of digital material.
- As vision recognition improves it has become possible for the computers to extract meaningful information from still images and videos.
- Retailers are building vast databases for recording customer activity.
- Organizations working on logistics, financial sector and health sectors are also capturing more and more data.
- Public Social media like face book, twitter, LinkedIn, YouTube is also creating vast quantities of digital material.
- As vision recognition improves it has become possible for the computers to extract meaningful information from still images and videos.
- Finally several areas of scientific computing are also generating huge amounts of data

TOOLS AND TECHNIQUES TO ANALYZE BIG DATA:

Another reason big data is gaining momentum is the fact that the tools to analyze it are becoming more and more accessible. Together, Tera data and IBM have been partnering for

well over a decade to help companies turn data into insights that lead to better and faster decisions. In other words, for decades, Oracle, IBM and Tera data have been providing thousands of companies with terabyte scale large data warehouses, but now there is this recent trend of big data being stored across multiple servers that can handle unstructured data and scale easily. This is due to the increasing use of technologies like Hadoop, which is an open source software project that enables distributed processing of large data sets across clusters of commodity servers. It is designed in a manner to scale up from a single server to thousands of machines and has a very high degree of fault tolerance. These clusters are highly resilient because of their software's ability to detect and handle failures at the application layer rather than relying on high-end hardware. They also allow fast data loading and real-time analytic capabilities. More effectively, Hadoop allows the analysis to occur at a location where the data resides, but it requires specific skills and is not an easy technology to adopt. Arcplan is one such BI software, which connects to Tera data which is a fully scalable relational database system, and SAP HANA, which is a revolutionary platform for real time analytics, allow data analysis and visualization on big data sets. So to be able to make use of big data, companies may need to adopt and implement new technologies, but some traditional Business Intelligence solutions can make the move. Big data is simply a new data challenge that requires leveraging existing systems in a different way.[6] [7]

TOP 50 BIG DATA TOOLS FOR DEVELOPERS

Big Data is everywhere. Even small to medium-sized businesses are seeking ways to gain more insight into processes, adapting additional streams, and derive more actionable visibility from their data. With data traditionally contained in warehouses, information silos within applications or databases, taking useful clues from Big Data was initially a tedious and complex process. But a big thanks to Big Data tools, Big Data management can now be streamlined in a comprehensive Interface.

Sophisticated platforms enable data management and business intelligence, end to end, with solutions for collecting, integrating, analyzing, and even predicting data in ways never before possible. The following Big Data tools , listed in no particular order of importance, for developers offer platforms for quick deployment of apps, the ability to integrated data gathering and analysis from multitudes of sources and applications, and even integrating online and offline data to put actions and events into context. [8]

Data analysis is a do-or-die requirement in today's scenario. We analyze remarkable vendor choices, from a rising Hadoop to conventional database players. Interestingly, many of the best known big data tools available are open source projects. The best known among them is Hadoop, which is proliferating an entire industry of related services and products.

Big Data Analysis Platforms and Tools	Databases/Data Warehouses	Business Intelligence	Data Mining	File Systems	Programming Languages	Big Data Search	Data Aggregation and Transfer	Miscellaneous Big Data Tools
1. Hadoop 2. MapReduce 3. GridGain 4. HPCC 5. Storm	6. Cassandra 7. HBase 8. MongoDB 9. Neo4j 10. CouchDB 11. OrientDB 12. Terrastore 13. FlockDB 14. Hibari 15. Riak 16. Hypertable 17. BigData 18. Hive 19. InfoBright Community Edition 20. Infinispan 21. Redis	22. Talend 23. Jaspersoft 24. Palo BI Suite/Jedox 25. Pentaho 26. SpagoBI 27. KNIME 28. BIRT/ Actuate	29. RapidMiner /RapidAnalytics 30. Mahout 31. Orange 32. Weka 33. jHepWork 34. KEEL 35. SPMF 36. Rattle	37. Gluster 38. Hadoop Distributed File System	39. Pig/Pig Latin 40. R 41. ECL	42. Lucene 43. Solr	44. Sqoop 45. Flume 46. Chukwa	47. Terracotta 48. Avro 49. Oozie 50. Zookeeper

1. Hadoop
You simply can't talk about big data without mentioning Hadoop. The Apache distributed data processing software is so pervasive that often the terms "Hadoop" and "big data" are used synonymously. The Apache Foundation also sponsors a number of related projects that extend the capabilities of Hadoop, and many of them are mentioned below. In addition, numerous vendors offer supported versions of Hadoop and related technologies. Operating System: Windows, Linux, OS X. [9]

2. MapReduce
Originally developed by Google, the MapReduce website describe it as "a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes." It's used by Hadoop, as well as many other data processing applications. Operating System: OS Independent. [9]

3. GridGain
GridGrain offers an alternative to Hadoop's MapReduce that is compatible with the Hadoop Distributed File System. It offers in-memory processing for fast analysis of real-time data. You can download the open source version from GitHub or purchase a commercially supported version from the link above. Operating System: Windows, Linux, OS X. [9]

4. HPCC
Developed by LexisNexis Risk Solutions, HPCC is short for "high performance computing cluster." It claims to offer superior performance to Hadoop. Both free community versions and paid enterprise versions are available. Operating System: Linux. [9]

5. Storm
Now owned by Twitter, Storm offers distributed real-time computation capabilities and is often described as the "Hadoop of realtime." It's highly scalable, robust, fault-tolerant and works with nearly all programming languages. Operating System: Linux. [9]

Databases/Data Warehouses

6. Cassandra
Originally developed by Facebook, this NoSQL database is now managed by the Apache Foundation. It's used by many organizations with large, active datasets, including Netflix, Twitter, Urban Airship, Constant Contact, Reddit, Cisco and Digg. Commercial support and services are available through third-party vendors.Operating System: OS Independent. [9]

7. HBase
Another Apache project, HBase is the non-relational data store for Hadoop. Features include linear and modular scalability, strictly consistent reads and writes automatic failover support and much more. Operating System: OS Independent. [9]

8. MongoDB
MongoDB was designed to support humongous databases. It's a NoSQL database with document-oriented storage, full index support, replication and high availability, and more. Commercial support is available through 10gen. Operating system: Windows, Linux, OS X, Solaris. [9]

9. Neo4j
The "world's leading graph database," Neo4j boasts performance improvements up to 1000x or more versus relational databases. Interested organizations can purchase advanced or enterprise versions from Neo Technology. Operating System: Windows, Linux. [9]

10. CouchDB
Designed for the Web, CouchDB stores data in JSON documents that you can access via the Web or or query using JavaScript. It offers distributed scaling with fault-tolerant storage. Operating system: Windows, Linux, OS X, Android. [9]

11. OrientDB
This NoSQL database can store up to 150,000 documents per second and can load graphs in just milliseconds. It combines the flexibility of document databases with the power of graph databases, while supporting features such as ACID transactions, fast indexes, native and SQL queries, and JSON import and export. Operating system: OS Independent. [9]

12. Terrastore
Based on Terracotta, Terrastore boasts "advanced scalability and elasticity features without sacrificing consistency." It supports custom data partitioning, event processing, push-down predicates, range queries, map/reduce querying and processing and server-side update functions. Operating System: OS Independent. [9]

13. FlockDB
Best known as Twitter's database, FlockDB was designed to store social graphs (i.e., who is following whom and who is blocking whom). It offers horizontal scaling and very fast reads and writes. Operating System: OS Independent. [9]

14. Hibari
Used by many telecom companies, Hibari is a key-value, big data store with strong consistency, high availability and fast performance. Support is available through Gemini Mobile. Operating System: OS Independent. [9]

15. Riak
Riak humbly claims to be "the most powerful open-source, distributed database you'll ever put into production." Users include Comcast, Yammer, Voxer, Boeing, SEOMoz, Joyent, Kiip.me, DotCloud, Formspring, the Danish Government and many others. Operating System: Linux, OS X. [9]

16. Hypertable
This NoSQL database offers efficiency and fast performance that result in cost savings versus similar databases. The code is 100 percent open source, but paid support is available. Operating System: Linux, OS X. [9]

17. BigData
This distributed database can run on a single system or scale to hundreds or thousands of machines. Features include dynamic sharing, high performance, high concurrency, high availability and more. Commercial support is available. Operating System: OS Independent. [9]

18. Hive
Hadoop's data warehouse, Hive promises easy data summarization, ad-hoc queries and other analysis of big data. For queries, it uses a SQL-like language known as HiveQL. Operating System: OS Independent. [9]

19. InfoBright Community Edition
This scalable data warehouse supports data stores up to 50TB and offers "market-leading" data compression up to 40:1 for improved performance. Commercial products based on the same technology can be found at InfoBright.com. Operating System: Windows, Linux. [9]

20. Infinispan
Infinispan from JBoss describes itself as an "extremely scalable, highly available data grid platform." Java-based, it was designed for multi-core architecture and provides distributed cache capabilities. Operating System: OS Independent. [9]

21. Redis
Sponsored by VMware, Redis offers an in-memory key-value store that can be saved to disk for persistence. It supports many of the most popular programming languages. Operating System: Linux. [9]

Business Intelligence

22. Talend
Talend makes a number of different business intelligence and data warehouse products, including Talend Open Studio for Big Data, which is a set of data integration tools that support Hadoop, HDFS, Hive, Hbase and Pig. The company also sells an enterprise edition and other commercial products and services. Operating System: Windows, Linux, OS X. [9]

23. Jaspersoft
Jaspersoft boasts that it makes "the most flexible, cost effective and widely deployed business intelligence software in the world." The link above primarily discusses the commercial versions of its applications, but you can find the open source versions, including the Big Data Reporting Tool at JasperForge.org. Operating System: OS Independent. [9]

24. Palo BI Suite/Jedox
The open source Palo Suite includes an OLAP Server, Palo Web, Palo ETL Server and Palo for Excel. Jedox offers commercial software based on the same tools. Operating System: OS Independent. [9]

25. Pentaho
Used by more than 10,000 companies, Pentaho offers business and big data analytics tools with data mining, reporting and dashboard capabilities. Seethe Pentaho Community Wiki for easy access to the open source downloads. Operating System: Windows, Linux, OS X. [9]

26. SpagoBI
SpagoBI claims to be "the only entirely open source business intelligence suite." Commercial support, training and services are available. Operating System: OS Independent. [9]

27. Knime
The Konstanz Information Miner, or KNIME, offers user-friendly data integration, processing, analysis, and exploration. In 2010, Gartner named KNIME a "Cool Vendor" in analytics, business intelligence, and performance management. In addition to the open source desktop version, several commercial versions are also available. Operating System: Windows, Linux, OS X. [9]

28. BIRT/Actuate
Short for "Business Intelligence and Reporting Tools," BIRT is an Eclipse-based tool that adds reporting features to Java applications. Actuate is a company that co-founded BIRT and offers a variety of software based on the open source technology. Operating System: OS Independent. [9]

Data Mining

29. RapidMiner / RapidAnalytics
RapidMiner claims to be "the world-leading open-source system for data and text mining." RapidAnalytics is a server version of that product. In addition to the open source versions of

each, enterprise versions and paid support are also available from the same site. Operating System: OS Independent. [9]

30. Mahout
This Apache project offers algorithms for clustering, classification and batch-based collaborative filtering that run on top of Hadoop. The project's goal is to build scalable machine learning libraries. Operating System: OS Independent. [9]

31. Orange
This project hopes to make data mining "fruitful and fun" for both novices and experts. It offers a wide variety of visualizations, plus a toolbox of more than 100 widgets. Operating System: Windows, Linux, OS X. [9]

32. Weka
Short for "Waikato Environment for Knowledge Analysis," Weka offers a set of algorithms for data mining that you can apply directly to data or use in another Java application. It is part of a larger machine learning project, and it's also sponsored by Pentaho. Operating System: Windows, Linux, OS X. [9]

33. jHepWork
Also known as "jWork," this Java-based project provides scientists, engineers and students with an interactive environment for scientific computation, data analysis and data visualization. It's frequently used in data mining, as well as for mathematics and statistical analysis. Operating System: OS Independent. [9]

34. Keel
KEEL stands for "Knowledge Extraction based on Evolutionary Learning," and it aims to help uses assess evolutionary algorithms for data mining problems like regression, classification, clustering and pattern mining. It includes a large collection of existing algorithms that it uses to compare and with new algorithms. Operating System: OS Independent. [9]

35. SPMF
Another Java-based data mining framework, SPMF originally focused on sequential pattern mining, but now also includes tools for association rule mining, sequential rule mining and frequent itemset mining. Currently, it includes 46 different algorithms. Operating System: OS Independent. [9]

36. Rattle
Rattle, the "R Analytical Tool To Learn Easily," makes it easier for non-programmers to use the R language by providing a graphical interface for data mining. It can create data summaries (both visual and statistical), build models, draw graphs, score datasets and more. Operating System: Windows, Linux, OS X. [9]

File Systems

37. Gluster
Sponsored by Red Hat, Gluster offers unified file and object storage for very large datasets. Because it can scale to 72 brontobytes, it can be used to extend the capabilities of Hadoop beyond the limitations of HDFS (see below). Operating System: Linux. [9]

38. Hadoop Distributed File System
Also known as HDFS, this is the primary storage system for Hadoop. It quickly replicates data onto several nodes in a cluster in order to provide reliable, fast performance. Operating System: Windows, Linux, OS X. [9]

Programming Languages

39. Pig/Pig Latin
Another Apache Big Data project, Pig is a data analysis platform that uses a textual language called Pig Latin and produces sequences of Map-Reduce programs. It helps makes it easier to write, understand and maintain programs which conduct data analysis tasks in parallel. Operating System: OS Independent. [9]

40. R
Developed by Bell Laboratories, R is a programming language and an environment for statistical computing and graphics that is similar to S. The environment includes a set of tools that make it easier to manipulate data, perform calculations and generate charts and graphs. Operating System: Windows, Linux, OS X. [9]

41. ECL
ECL ("Enterprise Control Language") is the language for working with HPCC. A complete set of tools, including an IDE and a debugger are included in HPCC, and documentation is available on the HPCC site. Operating System: Linux. [9]

Big Data Search

42. Lucene
The self-proclaimed "de facto standard for search libraries," Lucene offers very fast indexing and searching for very large datasets. In fact, it can index over 95GB/hour when using modern hardware. Operating System: OS Independent. [9]

43. Solr
Solr is an enterprise search platform based on the Lucene tools. It powers the search capabilities for many large sites, including Netflix, AOL, CNET and Zappos. Operating System: OS Independent. [9]

Data Aggregation and Transfer

44. Sqoop
Sqoop transfers data between Hadoop and RDBMSs and data warehouses. As of March of this year, it is now a top-level Apache project. Operating System: OS Independent. [9]

45. Flume
Another Apache project, Flume collects, aggregates and transfers log data from applications to HDFS. It's Java-based, robust and fault-tolerant. Operating System: Windows, Linux, OS X. [9]

46. Chukwa
Built on top of HDFS and MapReduce, Chukwa collects data from large distributed systems. It also includes tools for displaying and analyzing the data it collects. Operating System: Linux, OS X. [9]

Miscellaneous Big Data Tools

47. Terracotta
Terracotta's "Big Memory" technology allows enterprise applications to store and manage big data in server memory, dramatically speeding performance. The company offers both open source and commercial versions of its Terracotta platform, BigMemory, Ehcache and Quartz software. Operating System: OS Independent. [9]

48. Avro
Apache Avro is a data serialization system based on JSON-defined schemas. APIs are available for Java, C, C++ and C#. Operating System: OS Independent. [9]

49. Oozie
This Apache project is designed to coordinate the scheduling of Hadoop jobs. It can trigger jobs at a scheduled time or based on data availability. Operating System: Linux, OS X.[9]

50. Zookeeper
Formerly a Hadoop sub-project, Zookeeper is "a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services." APIs are available for Java and C, with Python, Perl, and REST interfaces planned. Operating System: Linux, Windows (development only), OS X (development only). [9]

RECOMMENDATIONS, FUTURE DIRECTIONS AND TRENDS IN BIG DATA

To successfully identify and implement big data solutions and benefit from the value that big data can bring organizations need to devote time and resources to visioning and planning. This will provide the foundation needed for strong execution. Without this, organizations will not be able to realize the possible benefits of big data and will risk being left behind fellow competitors.[10] Recommendations for organizations looking to leverage big data include :

- Establish a business intelligence division with a focus on big data.
- Decide on an appropriate big data strategy based on the organization's current and target business and technological maturity and objectives.
- Evaluate the various big data initiatives that can be deployed to meet overall enterprise goals and objectives, focusing initially on quick wins.
- Look for a partner that understands the full range of big data technologies and implications, including latest trends, security, internal and external system integration, hosting and development platforms, and application and solution development.
- The future of big data lies in asking the deeper questions and finding out why consumers make the decisions they do.

REFERENCES:

i. http://www-01 .ibm.com/software/data/bigdata
ii. http://en.wikipedia.org/wiki/Big_data
iii. Big Data, A new world of opportunities", NESSI White Paper , 2012
iv. An Oracle White Paper February 2013 Information Management and Big Data A Reference Architecture]
v. http://enterprisearchitects.com/the-5v-s-of-big-data
vi. http://www-01.ibm.com/software/data/infosphere/hadoop/
vii. http://in.teradata.com/partners/IBM/?LangType=16393&LangSelect=true
viii. blog.profitbricks.com/top-45-big-data-tools-for-developers/
ix. http://www.datamation.com/data-center/50-top-open-source-tools-for-big-data-2.html
x. http://www.healthmgttech.com/articles/201308/impacts-of-big-data.php