Analysis of Data Analytics

Mrs. SowmyaKoneru

Associate Professor, NRI Institute of Technology, Agiripalli, Vijayawada, Andhra Pradesh

Abstract- Today Big Data draws a lot of attention in the IT world. The rapid rise of the Internet and the digital economy has fuelled an exponential growth in demand for data storage and analytics, and IT department are facing tremendous challenge in protecting and analyzing these increased volumes of information. The reason organizations are collecting and storing more data than ever before is because their business depends on it. The type of information being created is no more traditional database-driven data referred to as structured data rather it is data that include documents, images, audio, video, and social media contents known as unstructured data or Big Data. Big Data Analytics is a way of extracting value from these huge volumes of information, and it drives new market opportunities and maximizes customer retention. This paper primarily focuses on discussing the various technologies that work together as a Big Data Analytics system that can help predict future volumes, gain insights, take proactive actions, and give way to better strategic decision-making. Further this paper analyzes the adoption, usage and impact of big data analytics to the business value of an enterprise to improve its competitive advantage using a set of data algorithms for large data sets such as Hadoop and MapReduce.

Keywords- Big Data, Analytics, Hadoop, MapReduce

I. INTRODUCTION

Big Data is an important concept, which is applied to data, which does not conform to the normal structure of the traditional database. Big Data consists of different types of key technologies like Hadoop, HDFS, NoSQL, MapReduce, MongoDB, Cassandra, PIG, HIVE, and HBASE that work together to achieve the end goal like extracting value from data that would be previously considered dead. According to a recent market report published by Transparency Market Research, the total value of big data was estimated at $6.3 billion as of 2012, but by 2018, it's expected to reach the staggering level of $48.3 billion that's almost a 700 percent increase [29]. Forrester Research estimates that organizations effectively utilize less than 5 percent of their available data. This is because the rest is simply too expensive to deal with. Big Data is derived from multiple sources. It involves not just traditional relational data, but all paradigms of unstructured data sources that are growing at a significant rate. For instance, machine-derived data multiplies quickly and contains rich, diverse content that needs to be discovered. Another example, human-derived data from social media is more textual, but the valuable insights are often overloaded with many possible meanings.

Big Data Analytics reflect the challenges of data that are too vast, too unstructured, and too fast moving to be managed by traditional methods. From businesses and research institutions to governments, organizations now routinely generate data of unprecedented scope and complexity. Gleaning meaningful information and competitive advantages from massive amounts of data has become increasingly important to organizations globally. Trying to efficiently extract the meaningful insights from such data sources quickly and easily is challenging. Thus, analytics has become inextricably vital to realize the full value of Big Data to improve their business performance and increase their market share. The tools available to handle the volume, velocity, and variety of big data have improved greatly in recent years. In general, these technologies are not prohibitively expensive, and much of the software is open source. Hadoop, the most commonly used framework, combines commodity hardware with opensource software. It takes incoming streams of data and distributes them onto cheap disks; it also provides tools for analyzing the data. However, these technologies do require a skill set that is new to most IT departments, which will need to work hard to integrate all the relevant internal and external sources of data. Although attention to technology isn't sufficient, it is always a necessary component of a big data strategy. This paper discusses some of the most commonly used big data technologies mostly open source that work together as a big data analytics system for leveraging large quantities of unstructured data to make more informed decisions.

II. LITERATURE REVIEW

Big Data is a data analysis methodology enabled by recent advances in technologies that support high-velocity datacapture, storage and analysis. Data sources extend beyond the traditional corporate database to include emails,mobile device outputs, and sensor-generated data where data is no longer restricted to structured database recordsbut rather unstructured data having no standard formatting [30]. Since Big Data and Analytics is a relatively newand evolving phrase, there is no uniform definition; various stakeholders have provided diverse and sometimescontradictory definitions. One of the first widely quoted definitions of Big Data resulted from the Gartner report of2001. Gartner proposed that, Big Data is defined by three V's volume, velocity, and variety. Gartner expanded itsdefinition in 2012 to include veracity, representing requirements about trust and uncertainty pertaining to data andthe outcome of data analysis. In a 2012 report, IDC defined the 4th V as value—highlighting that Big Dataapplications need to bring incremental value to businesses. Big Data Analytics is all about processing unstructuredinformation from call logs, mobile-banking transactions, online user generated content such as blog posts andtweets, online searches, and images which can be transformed into valuable business information usingcomputational techniques to unveil trends and patterns between datasets.

Another dimension of the Big Data definition involves technology. Big Data is not only large and complex, but itrequires innovative technology to analyze and process. In 2013, the National Institute of Standard and Technology(NIST) Big Data workgroup proposed the following definition of Big Data that emphasizes application of newtechnology; Big Data exceed the capacity or capability of current or conventional methods and systems, and enablenovel approaches to frontier questions previously inaccessible or impractical using current or conventional methods.Business challenges rarely show up in the appearance of a perfect data problem, and even when data are abundant,practitioners have difficulties to incorporate it into their complex decision-making that adds business value. In 2012,McKinsey & Company conducted a survey of 1,469 executives across various regions, industries and companysizes, in which 49 percent of respondents said that their companies are focusing big data efforts on customerinsights, segmentation and targeting to improve overall performance [10] An even higher number of respondents 60percent said their companies should focus efforts on using data and analytics to generate these insights. Yet, justone-fifth said that their organizations have fully deployed data and analytics to generate insights in one business unitor function, and only 13 percent use data to generate insights across the company. As these survey results show, thequestion is no longer whether big data can help business, but how can business derive maximum results from bigdata.

Predictive Analytics

Predictive Analytics is the use of historical data to forecast on consumer behavior and trends. It is the use ofpast/historical data to predict future trends. This analysis makes use of the statistical models and machine learningalgorithms to identify patterns and learn from historical data. Predictive Analysis can also be defined as aprocess that uses machine learning to analyze data and make predictions.

Sixty seven percent of businesses aim at using predictive analytics to create more strategic marketing campaign infuture, and 68% sight competitive advantage as the prime benefit of predictive analysis. Broadly speaking,predictive analysis can be applied in ecommerce for product recommendation, price management, and predictivesearch. Typically a large e-commerce site offers thousands of product and services for sale. Navigating andsearching for a product out of thousands on a website could be a major setback to consumers. However, with theinvention of recommender system, an E-Commerce site/application can quickly identify/predict products thatclosely suit the consumer's taste.

Using a technology called Collaborative Filtering a database of historical user preferences is created. When a newcustomer access the ecommerce site, the customer is matched with the database of preferences, in order to discover apreference class that closely matches the customer taste. These products are then recommended to the customer.Another technology that is used in ecommerce is the clustering algorithm. Clustering algorithm works by identifyinggroups of users that have similar preferences. These users are then clustered into a single group and are given a unique identifier.

New customers cluster are predicted by calculating the average similarities of the individual members in that cluster.Hence a user could be a partial member of more than one cluster depending of the weight of the user's averageopinion. Advanced analytics is defined as the scientific process of transforming data into insight for makingbetter decisions. As a formal discipline, advanced analytics have grown under the Operational Research domain.There are some fields that have considerable overlap with analytics, and also different accepted classifications forthe types of analytics [2].

III. BIG DATA TECHNOLOGIES

Apache Flume

Apache Flume is a distributed, reliable, and available system for efficiently collecting, aggregating and moving largeamounts of log data from many different sources to a centralized data store. Flume deploys as one or more agents,each contained within its own instance of the Java Virtual Machine (JVM). Agents consist of three pluggablecomponents: sources, sinks, and channels. Flume agents ingest incoming streaming data from one or more sources.Data ingested by a Flume agent is passed to a sink, which is most commonly a distributed file system like Hadoop.Multiple Flume agents can be connected together for more complex workflows by configuring the source of oneagent to be the sink of another. Flume sources listen and consume events. Events can range from newline-terminatedstrings in stdout to HTTP POSTs and RPC calls — it all depends on what sources the agent is configured to use.Flume agents may have more than one source, but at the minimum they require one. Sources require a name and atype; the type then dictates additional configuration parameters.Channels are the mechanism by which Flume agents transfer events from their sources to their sinks. Events writtento the channel by a source are not removed from the channel until a sink removes that event in a transaction.

Thisallows Flume sinks to retry writes in the event of a failure in the external repository (such as HDFS or an outgoingnetwork connection). For example, if the network between a Flume agent and a Hadoop cluster goes down, thechannel will keep all events queued until the sink can correctly write to the cluster and close its transactions with thechannel. Sink is an interface implementation that can remove events from a channel and transmit them to the nextagent in the flow, or to the event's final destination and also sinks can remove events from the channel intransactions and write them to output. Transactions close when the event is successfully written, ensuring that allevents are committed to their final destination.

Apache Sqoop

Apache Sqoop is a CLI tool designed to transfer data between Hadoop and relational databases. Sqoop can importdata from an RDBMS such as MySQL or Oracle Database into HDFS and then export the data back after data hasbeen transformed using MapReduce. Sqoop also has the ability to import data into HBase and Hive. Sqoop connectsto an RDBMS through its JDBC connector and relies on the RDBMS to describe the database schema for data to beimported. Both import and export utilize MapReduce, which provides parallel operation as well as fault tolerance.During import, Sqoop reads the table, row by row, into HDFS. Because import is performed in parallel, the output inHDFS is multiple files.

Apache Pig

Apache's Pig is a major project, which is lying on top of Hadoop, and provides higher-level language to useHadoop'sMapReduce library. Pig provides the scripting language to describe operations like the reading, filteringand transforming, joining, and writing data which are exactly the same operations that MapReduce was originallydesigned for. Instead of expressing these operations in thousands of lines of Java code which uses MapReducedirectly, Apache Pig lets the users express them in a language that is not unlike a bash or Perl script.Pig was initially developed at Yahoo Research around 2006 but moved into the Apache Software Foundation in2007. Unlike SQL, Pig does not require that the data must have a schema, so it is well suited to process theunstructured data. But, Pig can still leverage the value of a schema if you want to supply one. PigLatin is relationallycomplete like SQL, which means it is at least as powerful as a relational algebra. Turing completeness requiresconditional constructs, an infinite memory model, and looping constructs.

Apache ZooKeeper

Apache Zoo Keeper is an effort to develop and maintain an open-source server, which enables highly reliabledistributed coordination. It provides a distributed configuration service, a synchronization service and a namingregistry for distributed systems. Distributed applications use ZooKeeper to store and mediate updates to importconfiguration information. ZooKeeper is especially fast with workloads where reads to the data are more commonthan writes. The ideal read/write ratio is about 10:1. ZooKeeper is replicated over a set of hosts (called an ensemble)and the servers are aware of each other and there is no single point of failure.

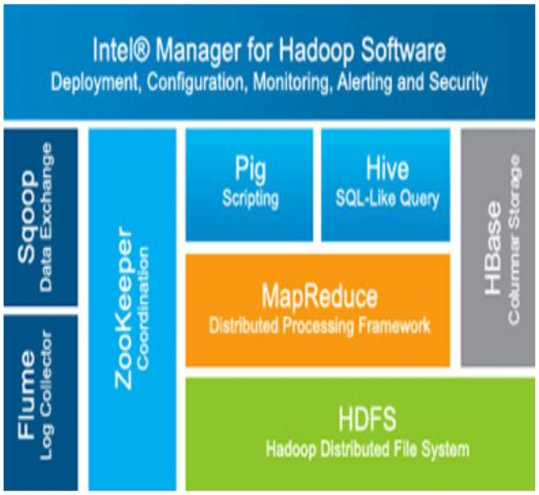

Fig.1: Intel Manager for Hadoop

MongoDB

MongoDB is an open source, document-oriented NoSQL database that has lately attained some space in the dataindustry. It is considered as one of the most popular NoSQL databases, competing today and favors master-slavereplication. The role of master is to perform reads and writes whereas the slave confines to copy the data receivedfrom master, to perform the read operation, and backup the data. The slaves do not participate in write operationsbut may select an alternate master in case of the current master failure. MongoDB uses binary format of JSON-likedocuments underneath and believes in dynamic schemas, unlike the traditional relational databases. The querysystem of MongoDB can return particular fields and query set compass search by fields, range queries, regularexpression search, etc. and may include the user-defined complex JavaScript functions. As hinted already,MongoDB practice flexible schema and the document structure in a grouping, called Collection, may vary andcommon fields of various documents in a collection can have disparate types of the data.

The MongoDB is equipped with the suitable drivers for most of the programming languages, which are used todevelop the customized systems that use MongoDB as their backend player. There is an increasingly demand ofusing MongoDB as pure in-memory database; in such cases, the application dataset will always be small. Though, itis probably are easy for maintenance and can make a database developer happier; this can be a bottle neck forcomplex applications that require tremendous database management capabilities.

IV. BIG DATA FRAMEWORK

Apache Spark

Apache Spark an open source big data processing framework built around speed, ease of use, and sophisticatedanalytics. It was originally developed in 2009 in UC Berkeley's AMP Lab, and open sourced in 2010 as an Apacheproject. Hadoop as a big data processing technology has been around for ten years and has proven to be the solutionof choice for processing large data sets. MapReduce is a great solution for one-pass computations, but not veryefficient for use cases that require multi-pass computations and algorithms. Each step in the data processingworkflow has one Map phase and one Reduce phase and you'll need to convert any use case into MapReduce patternto leverage this solution. Spark takes MapReduce to the next level with less expensive shuffles in the dataprocessing. With capabilities like in-memory data storage and near real-time processing, the performance can beseveral times faster than other big data technologies.Spark also supports lazy evaluation of big data queries, which helps with optimization of the steps in data processingworkflows. It provides a higher-level API to improve developer productivity and a consistent architect model for bigdata solutions. Spark holds intermediate results in memory rather than writing them to disk, which is very usefulespecially when you need to work on the same dataset multiple times. It's designed to be an execution engine thatworks both in-memory and on-disk. Spark operators perform external operations when data does not fit in memory.Spark can be used for processing datasets that larger than the aggregate memory in a cluster. Spark will attempt tostore as much as data in memory and then will spill to disk. It can store part of a data set in memory and theremaining data on the disk. You have to look at your data and use cases to assess the memory requirements. Withthis in-memory data storage, Spark comes with a great performance advantage.

Spark is written in ScaleProgramming Language and runs on the Java Virtual machine. It currently supportsprogramming languages like Scala, java, python, Clojure and R. Other than Spark Core API, there are additionallibraries that are part of the Spark ecosystem and provide additional capabilities in Big Data analytics. SparkStreaming is one among the spark library that can be used for processing the real-time streaming data. This is basedon micro based on micro batch style of computing and processing. Spark SQL provides the capabilities to expose thespark datasets over JDBC API and allow running the SQL like queries on Spark data using traditional BI andvisualization tools. MLlib, GraphX are some other libraries from spark.

V. COMPETITIVE ADVANTAGES

Thomas H. Davenport was perhaps the first to observe in his Harvard Business Review article published in January2006 ("Competing on Analytics") how companies who orientated themselves around fact based managementapproach and compete on their analytical abilities considerably out-performed their peers in the marketplace. Thereality is that it takes continuous improvement to become an analytics-driven organization. In a presentation given atthe Strata New York conference in September 2011, McKinsey & Company showed the eye opening; 10-yearcategory growth rate differences (see Figure 7, below) between businesses that smartly use their big data and thosethat do not.

Amazon uses Big Data to monitor, track and secure 1.5 billion items in its inventory that are laying around 200fulfillment centers around the world, and then relies on predictive analytics for its 'anticipatory shipping' to predictwhen a customer will purchase a product, and pre-ship it to a depot close to the final destination. Wal-Mart handles more than a million customer transactions each hour, imports information into databases to contain more than2.5 petabytes and asked their suppliers to tag shipments with radio frequency identification (RFID) systems thatcan generate 100 to 1000 times the data of conventional bar code systems. UPS deployment of telematics in theirfreight segment helped in their global redesign of logistical networks. Amazon is a big data giant and the largestonline retail store. The company pioneered e-commerce in many different ways, but one of its biggest successes wasthe personalized recommendation system, which was built from the big data it gathers from its millions of customers' transactions.

The U.S. federal government collects more than 370,000 raw and geospatial datasets from 172 agencies and subagencies. It leverages that data to provide a portal to 230 citizen-developed apps, with the aim of increasing publicaccess to information not deemed private or classified. Professional social network LinkedIn uses data from its morethan 100 million users to build new social products based on users' own definitions of their skill sets. Silver SpringNetworks deploys smart, two-way power grids for its utility customers that utilize digital technology to deliver morereliable energy to consumers from multiple sources and allow homeowners to send information back to utilities tohelp manage energy use and maximize efficiency. Jeffrey Brenner and the Camden Coalition mapped a city's crimetrends to identify problems with its healthcare system, revealing services that were both medically ineffective andexpensive.

VI. CONCLUSION

Today's technology landscape is changing fast. Organizations of all shapes and sizes are being pressured to be datadrivenand to do more with less. Even though big data technologies are still in a nascent stage, relatively speaking,the impact of the 3V's of big data, which now is 5v's cannot be ignored. The time is now for organizations to beginplanning for and building out their Hadoop-based data lake. Organizations with the right infrastructures, talent andvision in place are well equipped to take their big data strategies to the next level and transform their businesses.They can use big data to unveil new patterns and trends, gain additional insights and begin to find answers topressing business issues. The deeper organizations dig into big data and the more equipped they are to act uponwhat's learned, the more likely they are to reveal answers that can add value to the top line of the business. This iswhere the returns on big data investments multiply and the transformation begins. Harnessing big data insightdelivers more than cost cutting or productivity improvement but it definitely reveals new business opportunities.Data-driven decisions always tend to be better decisions.

VII. REFERENCES

[1]. Apache Software Foundation. (2010). Apache ZooKeeper. Retrieved April 5, 2015 fromhttps://zookeeper.apache.org

[2]. Chae, B., Sheu, C., Yang, C. and Olson, D. (2014). The impact of advanced analytics and data accuracy onoperational performance: A contingent resource based theory (RBT) perspective, Decision Support Systems, 59,119-126.

[3]. Chambers, C., Raniwala, A., Adams, S., Henry, R., Bradshaw, R., and Weizenbaum, N. (2010). Flume Java:Easy, Efficient Data-Parallel Pipelines. Google, Inc. Retrieved April 1, 2015 fromhttp://pages.cs.wisc.edu/~akella/CS838/F12/838-CloudPapers/FlumeJava.pdf.

[4]. Cisco Systems. Cisco UCS Common Platform Architecture Version 2 (CPA v2) for Big Data withComprehensive Data Protection using Intel Distribution for Apache Hadoop. Retrieved March 15, 2015, fromhttp://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/Cisco_UCS_CPA_for_Big_Data_with_Intel.html

[5]. DATASTAX Corporation. (2013, October). Big Data: Beyond the Hype - Why Big data Matters to you [Whitepaper]. Retrieved March 15, 2015 from https://www.datastax.com/wp-content/uploads/2011/10/WP-DataStax-BigData.pdf

[6]. Davenport, T &Patil, D. (2012). Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review,90, 70-76.

[7]. Dhawan, S &Rathee, S. (2013). Big Data Analytics using Hadoop Components like Pig and Hive. AmericanInternational Journal of Research in Science, Technology, Engineering & Mathematics, 88, 13-131. Retrievedfrom http://iasir.net/AIJRSTEMpapers/AIJRSTEM13-131.pdf.