International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

A Review Paper on Big Data Technologies used in Digital Marketing

Harshitha S[1], Shilpa GD[2]

[1]*Department of Telecommunication Engineering, R V College of Engineering, Karnataka, India.*
[2]*Assistant Professor, Department of Telecommunication Engineering, R V College of Engineering, Karnataka, India.*

Abstract - *In the past few years, there has been incredible changes happening in the field of Big Data and Cloud Computing where we are handling tremendous amount of data. This large amount of data produced is not similar in nature. The term big data refers to the big data set which is coming from different sources. This data is obtained in many fields such as Biology, Astrology, Physics, Business etc. Our main focus in this paper is to study the application of big data technology in Business. Out of the huge data set obtained, its our task to find the useful information which can cater the growth of digital marketing. The core purpose is to discuss important aspects of Digital marketing and the different tools used in Big Data technology.*

Key Words: Big Data, Redshift, HBase, Kafka, Nifi, Kylo

1.INTRODUCTION

Digital marketing is a technique which involves promotion of different goods or services using emerging technologies on the Internet which also includes mobile phones, display ads and all different components in a digital media. The main motto of Digital marketing is to turn data into personalized customer experiences which helps in the growth of some of the well known organizations throughout the world. It mainly focuses on targeted digital marketing. Digital marketing enables marketers to create data-driven, omni-channel consumer experiences through a flexible and transparent data platform that provides 360° consumer views.

PII (Personably Identifiable Information) is the information about an individual that can be used to uniquely locate, contact or identify a person. A good digital marketing technique involves being responsible for identifying the customers based on the PII of individuals obtained through online or offline sources. The known data of the customers along with the behavioral activity tracked on different devices of the customer is used to obtain a complete picture of how the user is interacting with the specific brand. But all this specific information about a customer is received as a tremendous amount of data. This huge chuck of data has to be handled using Big data technologies. Big data refers to the large amount of data sets received from different sources which can be structured, semi-structured or unstructured data. It has a complex nature which cannot be handled using the traditional Business Intelligence tools. Business Intelligence is the act of transforming data into useful information for business analysis. This requires advanced algorithm and highly powerful technologies. The data obtained from various sources is stored in a database which cannot be directly visualized. It has to be integrated, processed and later visualized. So, the consolidated data obtained from multiple databases is stored in a Data warehouse. A data mart is a subset of a warehouse which is created in with respect to a particular business line in mind. These data marts helps us to capture insights about the consumer by processing and analyzing the obtained data. They provide us with a clear picture about the customer's journey and obtain the important marketing touchpoints which drives the growth of that particular brand. Necessary business actions are designed accordingly.

As a part of capturing, processing and obtaining important insights about the data, various Big data technologies are employed. In this paper, we will be discussing various Big data tools such as Amazon Redshift, HBase, Apache Kafka, Apache Nifi and Kylo.

2. LITERATURE REVIEW

A CRM Data warehouse acts as a storehouse which fetches the data from different sources within an organization. This data is used for business analysis. Due to exponential increase in the organizational data collected, the concept of in house data warehouse is declining and this in turn is giving rise to Cloud Based Data warehouse. The leading player in market right now in the field of cloud based data warehouse is AWS Redshift [1]. A Data warehouse can be designed using a bottom – down approach, top-down approach or by using a combination of both [2].

AWS Redshift has a number a number of advantages when compared to conventional cloud storage. It is known for its Amazon EC2 for instances, Amazon S3 for backup and Amazon VPC for security. The architecture of Amazon Redshift is discussed in [3]. Redshift is a parallel processing unit and SQL compliant. The cluster in redshift comprises of a leader node and one or more compute nodes. The leader node is the one that accepts connections from client applications. It computes a query plan for execution in compute nodes and passes it on to the compute nodes. The query processing and manipulation of data is done in the compute nodes.

The data that is stored in a warehouse undergoes an ETL process which is Extract, Transform and Load. The traditional ETL process is performed on batch data. This process is carried out using three different technologies. NIFI is used for extracting and loading the data. Spark is used for

transformation of data. The key features of NIFI are Guaranteed delivery, data buffering, Visual command and control. Kafka is distributed messaging system which handles high volume of data for real time data feeds. Important features include scalability, fault tolerance, reliability and Performance. Integration of NIFI with Kafka is done to avoid data fault tolerance [4].

Along with the batch data obtained, the real time data analysis plays a vital role in Big data Analysis. The different phases involved in real time data processing are data ingestion, storage, processing, analysis and reporting. The real time data processing tools like Kafka, NIFI, Flume, Spark, HBase are discussed in [5]. HBase is a Column Oriented database system which is placed on top of Hadoop Distributed File System (HDFS). It is a vital component in Hadoop that supports random real time data read/write access in HDFS. It doesn't support SQL. Hence Java is used in HBase.

A framework is discussed in [6] for efficient management of huge datasets obtained in real time data processing in a communication industry where it is used to manage history of call records of the subscriber. The first key component of this framework is big data warehouse and the second key component is Big data ETL which performs transformation and aggregation of the records obtained.

As mentioned earlier Apache Kafka plays an important role in data ingestion in real time. The architecture of Kafka is discussed in [7]. It consists of producers and consumers. The producers are responsible for pushing the data from various devices to Apache Kafka nodes which enables distribution of data as messages. The consumers are responsible for picking up the data from the Kafka brokers and provide it to the processing nodes. Kafka cluster has to manage producers, consumers and brokers by synchronizing the nodes and managing the message queues [8].

Kafka cluster has to manage producers, consumers and brokers by synchronizing the nodes and managing the message queues [8].

3. TOOLS USED IN BIG DATA TECHNOLOGY

Big data is the process of obtaining enormous amount of data from different sources. This huge amount of data obtained cannot be handled with traditional business intelligence tools. Powerful tools are required for capturing and processing of big data. There are different big data tools used in market today. Here, we will be discussing about a few important technologies which plays an important role in Big Data.

3.1 Amazon Redshift

Amazon Redshift is a peta byte scale Data Warehouse service in the cloud. It has the ability to handle large amount of data and perform query execution in no time. It is suitable for OLAP (Online Analytical Processing) systems. It is based on PostGRE SQL 8.0.2. Demand for Redshift is increasing due to Industries shifting from local servers to cloud, its efficiency for deep data analysis, faster and more compressed warehouse with low cost and over 1 million customer base. The important features of redshift are columnar storage, cloud based technology, column compression and Parallel Query execution.

The basic architecture of a Redshift cluster is as shown in Fig.1. Redshift supports various client applications like ETL tools, BI or external databases. It provides various ways for those clients to connect to Redshift. Users may create one or more clusters within Redshift. Each present cluster can host can databases. Most of the projects only need one of the Redshift clusters. For resilience purposes additional clusters may be added if required. Each cluster consists of a leader node which coordinates analytical queries and compute nodes which execute the queries. An internal network with high speed helps in connecting all the cluster nodes together to guarantee high-speed communication. Each node in the cluster is divided into slices, which are essentially data shards. Inside every node is one or more PostgreSQL-based databases. The implementation of Redshift is different from a regular implementation of PostgreSQL which stores the user data.

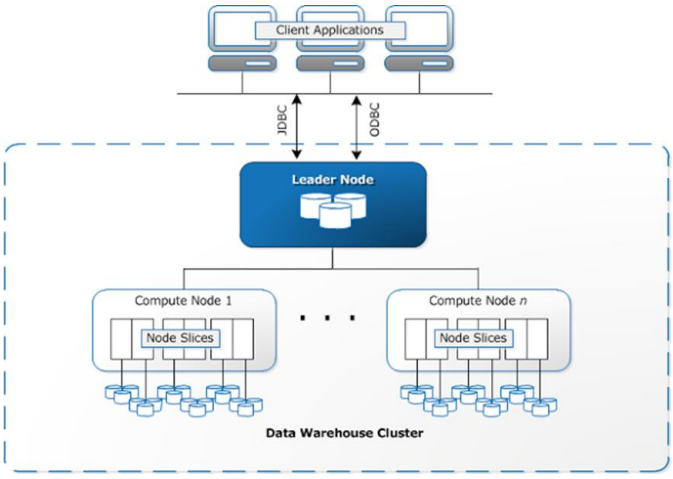

Fig -1: Architecture of Amazon Redshift

3.2 Apache HBase

Apache HBase is a non-relationally distributed open-source database built following Google's Bigtable which is implemented and written in Java. It is developed as a fragment of the Apache Hadoop project from the Apache Software Foundation and runs on top of the Hadoop Distributed File System (HDFS), providing Hadoop with Bigtable-like capabilities. It provides us with a fault-tolerant

way to store large quantities of sparse data, such as finding the non-zero items that represent less than 0.1 percent of a huge collection or finding the 50 largest items in a group of 2 billion records. It is well matched for data processing in real time, or for random read / write access for large volumes of data.
Major enterprises like Netflix, Pinterest, Spotify, Xiaomi, Yahoo etc use HBase for real time database storage system. The advantages of HBase are Random Read/Write access, processes high volume of requests, reliable, flexible as it provides column-based multidimensional map structure, supports variable schema where columns can be added and removed dynamically, enables integration with REST APIs and Java client, provides low latency while accessing data, allows data compression and it is ideal for sparse data.

3.3 Apache Kafka

Kafka is a messaging framework designed to be fast, scalable, and enduring. It is an opensource stream processing platform. Apache Kafka originated in LinkedIn and later became an open-source Apache project in 2011. It aims to provide a low latency, high-throughput platform for managing data feeds in real time. The main benefits of Kafka are Reliability, Scalability, Durability and Stable performance.
The basic components of kafka are Topics, Partitions, Brokers, Zookeeper and Cluster.
A topic is the name of the feed or category in which records are published. Kafka's topics are often multisubscriber, i.e., a topic may have zero, one, or several customers who can subscribe to the data written to them. The Kafka cluster maintains a partition log for each topic.
Brokers are simple systems which are responsible for holding the published data. Kafka brokers don't have a particular condition, so they use Zookeeper to preserve cluster status. Each broker in a cluster can have zero partitions, or more, per topic.
ZooKeeper is the one that is used to handle and organize Kafka brokers. ZooKeeper is primarily used to alert the producers and consumers present of any new broker 's involvement in the Kafka system or any broker's failure in the Kafka system. ZooKeeper notifies supplier and consumer of a broker 's existence or failure on the basis of which producer and consumer make a decision and starts coordinating their tasks with some other broker.
When Kafka has multiple brokers, it is called as a cluster. Without downtimes a Kafka cluster can be extended. The clusters in kafka are used to manage Message data replication and persistence. A structure of Kafka eco system is as shown in Fig.2.

3.4 Apache Nifi

Apache Nifi is a tool which is used in real time for integration of data streams. It is a simple event processing platform which can be used for automating the flow of streams of data between different data sources and software systems. A feature which makes Nifi better than Kafka and Flume is that

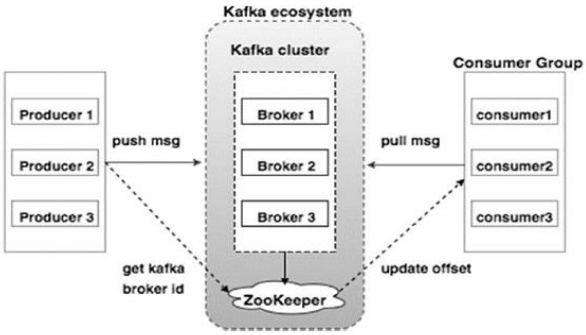

Fig -2: Kafka eco system

it can handle messages with arbitrary sizes. It basically makes use of a web based user interface which has drag-and-drop features and can deliver a real-time control which makes it really easy for the users to manage the flow of data streams between sources of data and the systems.

3.5 Kylo

Kylo is an open source software used in enterprises. It is a ready to use software which acts as a data lake management software platform. It helps in ingestion of data and preparation of data with integrated metadata management, governance, security and best practices inspired by Think Big's 150+ big data implementation projects.
A Data Lake refers to the methodology which involves a large data repository based on low-cost technologies. It contains raw unstructured or multi-structured data which is an important asset to the firm. It is built in order to handle large and fast arriving volumes of data using which important insights can be derived. The Kylo data lake architecture is as shown in Fig.3.
Kylo is built on Apache Hadoop and Spark. Kylo was initially developed by Teradata company which offers various features for data setwards, data analysts, data scientists etc. It is compatible with EMR, Hadoop distribution and supports software like Hortonworks DataFlow and Apache NiFi.

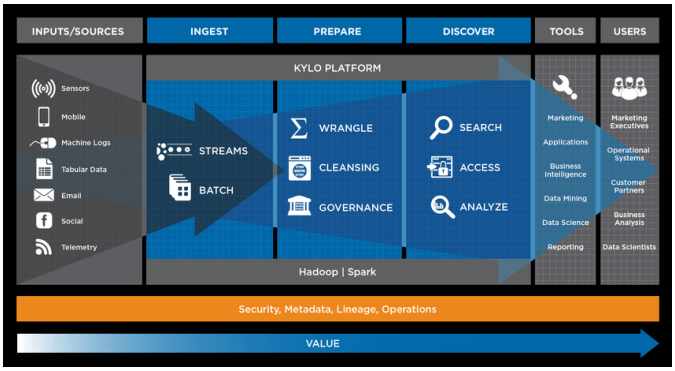

Fig -3: Kylo Data lake architecture

4. CONCLUSIONS

In the recent years data is being produced at a rapid rate in different fields. In this paper, we have surveyed on how to process and store huge amount of data obtained in digital marketing using tools like Amazon Redshift, HBase, Apache Kafka, Apache Nifi and Kylo. These tools play a vital role in managing huge amount of data in a suitable way. The main problems encountered in big data are the huge storage capacity required and processing of the obtained data. The tools available currently do not address all the problems of Big Data analytics. But comparatively there is a fair growth in increase in the ease of handling diverse data sets in a reduced amount of time. This paves a way for future improvements and developments of Big Data analytics tools.

REFERENCES

[1] P. Dutta., "Business Analytics using Microsoft Power BI and AWS Redshift," International Journal of Trend in Scientific Research and Development, vol. 3, Issue-2, pp. 984–986, Feb. 2019.

[2] S. Rizzi, A. Abelló, J. Lechtenbörger, and J. Trujillo., "Research in data warehouse modeling and design," *in Proceedings of ACM international workshop on Data warehousing and OLAP - DOLAP*, 2006.

[3] A. Gupta et al., "Amazon Redshift and the Case for Simpler Data Warehouses," *in Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15*, 2015.

[4] Saroja Chatti., "Using Spark, Kafka and NIFI for Future Generation of ETL in IT Industry", Journal of Innovation in Information Technology, Dec. 2019.

[5] F. Gurcan and M. Berigel., "Real-Time Processing of Big Data Streams: Lifecycle, Tools, Tasks, and Challenges," *in Proceedings of International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2018.

[6] Abbas Raza Ali, "Real-Time Big Data Warehousing and Analysis Framework," *in Proceedings of IEEE International Conference on Big Data Analysis*, 2018.

[7] P. Le Noac'h, A. Costan, and L. Bouge., "A performance evaluation of Apache Kafka in support of big data streaming applications," *in Proceedings of IEEE International Conference on Big Data (Big Data)*, 2017.

[8] O.C. Marcu et al., "Towards a unified storage and ingestion architecture for stream processing," *in Proceedings of IEEE International Conference on Big Data (Big Data)*, 2017.

[9] Dimitar Trajanov et al.,"Dark Data in Internet of Things (IoT): Challenges and Opportunities", *in Proceedings of 7th Small Systems Simulation Symposium*, 2018.

[10] Mr. Dishek Mankad, Mr. Preyash Dholakia., "The Study on Data Warehouse Design and Usage", International Journal of Scientific and Research Publications, vol. 3, Issue-3, March 2013.