SOCIAL DATA ANALYSIS USING BIG-DATA ANALYTIC TECHNOLOGIES-APACHE FLUME, HDFS, HIVE.

Ms.Sulochana Panigrahi [1],
[1] PG Scholar, Department of Computer Science and Engineering,
New Horizon College of Engineering, Bangalore, Karnataka, India
E-Mail: sulochanap01@gmail.com;

Dr.S Mohan Kumar [2],
[2] Associate Professor, Department of Computer Science and Engineering,
New Horizon College of Engineering, Bangalore, Karnataka, India
E-Mail:drsmohankumar@gmail.com

ABSTRACT: Analysis on Social media data is difficult due to language that is used for comments. For this analysis using Hive and its queries to give the sentiment data based up on the groups that defined in the HQL (Hive Query Language) is utmost important and so much in demand. The analysis can be positive, moderate and negative comments. Tools can be used to do the analysis are: Apache Flume, HDFS, Hive. In this paper all related research works and Big-data Tools, technologies and observations from the survey are all presented for better understanding.

Keywords: BIG-DATA, Flume, Hive, HQL, Structured, HDFS, Un-Structured, Semi-Structured, Twitter, Tweets.

INTRODUCTION: The rise of social media in couple of years has changed the general perspective of networking, socialization, and personalization. Use of data from social networks for different purposes, such as election prediction, sentimental analysis, marketing, communication, business, and education, is increasing day by day.

Precise extraction of valuable information from short text messages posted on social media is a collaborative task. People are increasingly using social media to share advice, opinions, news, moods, concerns, facts, and rumors. Sentiment analysis influences users to classify whether the information about the product is satisfactory or not before they acquire it. Marketers and firms use this analysis to understand about their products or services in such a way that it can be offered as per the user’s needs. Existing database environments, designed years ago, lacks the ability to process big data within the specified amount of time.

These types of databases have limitations when dealing with different types of data in real time enterprises. Traditional database cannot help organizations to manage complex and unstructured data generated in several ways. Using big data technologies like Hadoop is the best way to solve the big data challenges. These help industries to handle large of complex and unstructured data from various sources.

BIGDATA:

- **Big data** is a term for data sets that are so complex that traditional data processing applications are inadequate. Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk.

- Challenges include:
  - Analysis,
  - Capture,
  - Data creation,
  - Search,
  - Sharing,
  - Storage,
  - Transfer,
  - Visualization,
  - Querying and Information privacy.

- Big Data as:
  - Velocity – how fast the data is entering the systems
  - Variety – includes all types of structured and unstructured data
  - Volume – the potential data capacity of terabytes to petabytes
  - Complexity – includes everything from transferring operational data to big data platforms and the trouble with managing the data across many geographies and locations.

HDFS: The Hadoop Distributed File System is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications. In a large cluster, thousands of servers both host directly attached storage and execute user application tasks.

APACHE FLUME: It is a distributed, reliable, and available service for efficient collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application.

APACHE HIVE:

It is a data warehousing infrastructure, built on top of Hadoop for providing data summarization, query, and analysis. Apache Hive is now used companies such as Netflix. Amazon maintains a software fork of Apache Hive that is included in Amazon Elastic MapReduce, Amazon Web Service.

Related Research Works:

- ❖ Mahalakshmi R, Suseela(2015) Big-SoSA: Social Sentiment Analysis and Data Visualization on Big Data [1]. It proposes a method of sentiment analysis on twitter by using Hadoop and its ecosystems that process the large volume of data on a Hadoop and the MapReduce function performs the sentiment analysis.
- ❖ Praveen Kumar1, Dr Vijay Singh Rathore (2014) Efficient Capabilities of Processing of Big Data using Hadoop Map Reduce [2] Proposes, several solutions to the Big Data problem have emerged which includes the Map Reduce environment championed by Google which is now available open-source in Hadoop. Hadoops distributed processing,

Map Reduce algorithms and overall architecture are a major step towards achieving the promised benefits of Big Data.

- Sunil B. Mane, Yashwant Sawant, Saif Kazi (2014) Real Time Sentiment Analysis of Twitter Data Using Hadoop [3] . Proposes and provides a way of sentiment analysis using Hadoop which will process the huge amount of data on a Hadoop cluster(faster in real time).
- Penchalaiah.C, Murali.G, Suresh Babu (2014) Effective Sentiment Analysis on Twitter Data using: Apache Flume and Hive[4]. It says how effectively sentiment analysis is done on the data which is collected from the Twitter using Flume. Twitter contains rich amount of data that can be a structured, semi-structured and un-structured data. Collects the data from the twitter by using BIGDATA eco-system using online streaming tool Flume. Using Hive and its queries to give the sentiment data based up on the groups, that was defined in the HQL (Hive Query Language).
- Manoj Kumar Danthala (2015) Tweet Analysis: Twitter Data processing Using Apache Hadoop [5] . This paper provides a way of analyzing of big data such as twitter data using Apache Hadoop which will process and analyze the tweets on a Hadoop clusters. This also includes visualizing the results into pictorial representations of twitter users and their tweets.
- Manoj Kumar Danthala (2015) Bigdata Analysis: Streaming Twitter Data with Apache Hadoop and Visualizing using Big Insights[6]. It proposes, twitter data, which is the largest social networking area where data is increasing at high rates every day is considered as big data. This data is processed and analyzed using InfoSphere BigInsights tool which bring the power of Hadoop to the enterprise in real time. This also includes the visualizations of analyzing big data charts using big sheets.
- Judith Sherin Tilsha S, Shobha M.S(2015) A Survey on Twitter Data Analysis Techniques to Extract Public Opinion [7] . Using machine learning algorithm ,a feature vector is constructed with the emotion describing words from tweets and are fed to the classifier that classifies the sentiment or opinion. It said that various twitter data analysis techniques that are based on dictionary and that are using the machine learning approaches.
- Munesh Kataria1, Ms.Pooja(2014) Big Data and Hadoop with Components like Flume, Pig, Hive and JAQL [8]. The data from social media using Flume. Flume can take log files as source and after collecting data, it can store it directly to file system like HDFS or GFS. Then, organize this data by using different distributed file system such as Google file system or Hadoop file system. At last, data will be analyzed using map-reducers in Pig, Hive and Jaql. Components like Pig, Hive and Jaql do the analysis on data so that it can be access faster and easily, and query responses also become faster.
- Kushal Sharma, Prashant Singh, Sachin Mote, Sudarshan Patil, Vilas Khedekar 2015 Twitter Sentimental Analysis using Hadoop [9] . It aims to build an algorithm that can accurately classify Tweets as positive or negative with respect to a specific subject. The proposed system uses dictionary based approach to determine the semantic orientation of Tweets. Twitter Sentiment Analysis is used to understand how the public feels about something at a particular moment in time and also tracks how this opinion changes over time. The project uses Hadoop framework for distributed storage and processing of twitter data.

- Mr.Sagar Nadagoud (2015), Market Sentiment Analysis for Popularity of Flipkart [10]. It is taking sentiment analysis, for this it is using Hive and its queries to give the sentiment data based up on the groups that have defined in the HQL (Hive Query Language). Here they had categorized this sentiment analysis into 3 groups like tweets that are having positive, neutral and negative comments.
- Hanane EL MANSSOURI, Soufiane FARRAH. El Houssaine, ZIYATI Mohammed (2015), A Big Data Methodology for Sentiment Analysis of Twitter Data [11]. It is proposing a methodology to collect and store live twitter data and perform sentimental analysis using machine learning techniques and provide some prediction. To store the live data fetched, we are using MongoDB a NoSQL database, the output of the analysis will be trend analysis with different sections that is positive, negative and neutral.

- Mangesh U. Sanap1, Prof.V.S (2015) Survey on Buddy Analytics Based on Social Media [12] .It provides a way of analyzing of big data such as Facebook data using Apache Hadoop which will process and analyze the comments on a Hadoop clusters.

- Ritu Jain1, Mukesh Rawat (2015) Reduce Traffic of Data on Network with Bloom filter [13]. This paper provide the concept to reduce traffic of data on network with efficient manner through probabilistic model of Bloom filter. Bloom filter technique is probabilistic data model for getting data into array so that no need to travel all data in network. With implementation of this filter mapper can reduce the amount of data travel.

- S Anjali , V Aswini , M Abirami (2015) Predictive Analysis With Cricket Tweets Using Big Data [14]. One such field of data analysis is “sports”. People all around the world tweet about cricket matches going on every day. Huge amount of data around the social networking data can be fetched, analyzed and manipulated to predict the chances of team to win. By using big data concept such as map and reducing algorithm the fetched data is analyzed.
- HAN HU, YONGGANG WEN, TAT-SENG CHUA1, AND XUELONG LI (2014), Toward Scalable Systems for Big Data Analytics: A Technology Tutorial [15]. It propose a literature survey and system tutorial for big data analytics platforms, aiming to provide an overall picture for non expert readers and instill a do-it-yourself spirit for advanced audiences to customize their own big-data solutions.

OBSERVATION

Hadoop and its Ecosystems, for getting raw data from the Social Network, we may use Hadoop online streaming tool- using Apache Flume. By utilizing this tool only, we are going to configure everything, which we wanted to get (data) from the Social Network. Mainly we want to set the configuration model and also want to define what information that we want to collect form Social Network. All these will be stored into our HDFS (Hadoop Distributed File System) in our own prescribed format. From this unrefined data we are going to create the table and filter the information that is needed for us and sort them into the Hive Table. And from this, we are going to perform the Sentiment Analysis by using some UDF’s (User Defined Functions) by which we can perform sentiment analysis and also we can decide the list of words that are coming under positive, moderate and negative.

CONCLUSION:

Big data is mainly collection of data sets, so large and complex in nature, it is very difficult to handle them using on-hand database management tools. The main challenges with Big databases include creation, duration, storage, sharing, search, analysis and visualization. So to manage these databases need, “highly parallel software’s. First of all, data is acquired from different sources such as social media, traditional enterprise data or sensor data etc.Flume can be used to acquire data from social media/ network. Then, this data can be organized using distributed file systems such as Google File System or Hadoop File System. These file systems are very efficient when number of reads are very high as compared to writes. At last, data can be analyzed using map reducer, so that queries can be run on this data easily and efficiently.

REFERENCES:

[1] Mahalakshmi R, Suseela S PG Scholar, Big-SoSA:Social Sentiment Analysis and Data Visualization on Big Data, International Journal of Advanced Research in Computer and Communication Engineering ,VOL.4, April 2015

[2] Praveen Kumar, Dr Vijay Singh Rathore,Efficient Capabilities of Processing of Big Data using Hadoop Map Reduce,International Journal of Advanced Research in Computer and Communication Engineering , Vol .3, ISSUE6, June 2014

[3] Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde , Real Time Sentiment Analysis of Twitter Data Using Hadoop, International Journal of Computer Science and Information Technologies. Vol.5, ISSUE 3, 2014.

[4] Penchalaiah.C,Murali.G ,Suresh Babu, Effective Sentiment Analysis on Twitter Data using: Apache Flume and Hive, IJISET - International Journal of Innovative Science, Engineering & Technology, Vol.1, ISSUE 8, ISSUE 8, October 2014.

[5] Manoj Kumar Danthala ,Tweet Analysis: Twitter Data processing Using Apache Hadoop, International Journal Of Core Engineering & Management (IJCEM), Volume 1 , Issue 11 , February 2015 .

[6] Manoj Kumar Danthala , Dr. Siddhartha Ghosh , Bigdata Analysis: Streaming Twitter Data with Apache Hadoop and V isualizing using BigInsights, International Journal of Engineering Research & Technology (IJERT) Vol. 4 Issue 05, May-2015.

[7] Judith Sherin Tilsha S*, Shobha ,A Survey on Twitter Data Analysis Techniques to Extract Public Opinion , International Journal of Advanced Research in Computer Science and Software Engineering Research Paper , Volume 5, Issue 11 , November 2015 .

[8] Munesh Kataria1, Ms. Pooja Mittal,Big Data and Hadoop with Components like Flume, Pig, Hive and Jaql, International Journal of Computer Science and Mobile Computing A Monthly Journal of Computer Science and Information Technology . IJCSMC, Vol. 3, Issue 7, July 2014.

[9] Kushal Sharma, Prashant Singh, Sachin Mote, Sudarshan Patil,Twitter Sentimental Analysis using Hadoop, International Journal of Computer Application , Volume 2, Issue 5, January 2015 .

[10] Mr.Sagar Nadagoud , Channa basaveshwara ,Market Sentiment Analysis for Popularity of Flipkart, International Journal of Advanced Research in Computer Engineering &Technology(IJARCET) , Volume 4, Issue 5, May2015.

[11] Hanane EL MANSSOURI* Soufiane FARRAH El Houssaine ZIYATI Mohammed ,A Big Data Methodology for Sentiment Analysis of Twitter Data , International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization), Vol. 3, Issue 7, July 2015.

[12] Mangesh U. Sana, Prof.V.S.Phad, SKNCOE, Pune, India, Survey on Buddy Analytics Based on Social Media, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 12, December 2015 .

[13] Ritu Jain, Mukesh Rawat,Reduce Traffic of Data on Network with Bloom filter, Volume1, Issue-5, August 2015.

[14] S Anjali , V Aswini , M Abirami,Predictive Analysis With Cricket Tweets Using Big Data, International Journal of Scientific &Engineering Research, Volume 6, Issue 10, October 2015.

[15] HAN HU, YONGGANG WEN, TAT-SENG CHUA, AND XUELONG LI, Toward Scalable Systems for Big Data Analytics: A Technology Tutorial, IEE ACCESS, Volume 1, Issue10, June 24 2014.

[16] Revathi Y1, Dr. S Mohan Kumar 2 ,Review on Importance and Advancement in Detecting Sensitive Data Leakage in Public Network ,International Journal of Engineering Research and General Science Volume 4, Issue 2,March-April,2016.

[17] Revathi Y1, Dr. S Mohan Kumar 2 ,A Survey on Detecting the Leakage of Sensitive Data in Public Network,International Journal of Emerging Technology andAdvanced Engineering.

[18] Mr.Dilish Babu.J1 ,Dr. S Mohan Kumar 2,A Survey on secure communication in public network during disaster