Abstract— Recent Technology have led to a large volume of data from different areas (ex. medical, aircraft, internet and banking transactions) from last few years. Big Data is the collection of this field’s information. Big Data contains high volume, velocity and high variety of data. For example Data in GB or PB which are in different form i.e. structured, unstructured and semi-structured data and it requires more fast processing or real-time processing. Such Real time data processing is not easy task to do, Because Big Data is large dataset of various kind of data and Hadoop system can only handle the volume and variety of data but for real-time analysis we need to handle volume, variety and the velocity of data. To solve or to achieve the high velocity of data we are using two popular technologies i.e. Apache Kafka, which is message broker system and Apache Storm, which is stream processing engine.

I. INTRODUCTION

1.1 What is Big Data?

Big Data is a large amount of data set which cannot handle by any conventional database system. And this data is too large, it requires fast processing and it doesn't match to the structure of database architecture [1]. To obtain knowledge from this data, we must select an best alternative way to processing it i.e. real-time processing. The current situation of this era, Big Data has become a viable, complex and expensive approach to handle volume, velocity, variety of massive data [2].

In this data there is important pattern and information, which is previously hidden because we need to more processing to extract that data. To leading firms, like WalMart, Yahoo and Google they are handling or managing this problem in reasonable time and cost. Today’s commodity hardware, open source software and cloud architecture bring Big Data processing effectively by optimizing the resources [3]. The need of Big Data to companies is under into two categories: enabler of new items and Analytical.

Big Data analytic can disclose the hidden previously by data too expensive the processing for example banking transaction and geographical and social data. In past few year the effectively startup of internet are best example of Big Data used as an enabling new product and services for example by gathering large amount of information from users interaction and those of their friends, facebook has been enable highly customized users experience and creating a new way advertising business.

As a knowing all above terms, Big Data can be more complex, like cloud technology because it contains heterogeneous technologies and information. The incoming data to Big Data system are collected from social networks, satellite imaginary, web server logs, banking transactions, scans of government document, MP3 rock music, GPS tracking, airplane tracking, and stock market and so on.

To knowing this entire concept there are three Vs of Big Data i.e. volume, variety and velocity are generally used to categorizing dissimilar feature of Big Data.

Following details of Big Data 3 V’s Challenges:

![Figure 1. Big Data 3 V's Challenges](image)

1.1.1. Volume of Data

The size of data has been growing or increasing day by day. This applies to organization and their employee. For example text file with few KB, MP3 file with few MB and full movie with few GB in size. And more information is added in continuous manner day by day. For organization in last few year all data are generated by internal employee but currently these data can be generated by employee, customer, client, partner and machines, for example billions of smart phone can dump variety of information to the network infrastructure and this kind of data are not introduced in last few year back. Peta Byte dataset are general in these days and Exa Byte is in future days.

1.1.2. Velocity of Data

Previously organizations analyzed the Big Data by using Batch Processing.
In that we are take chunk or group of data, submits that task to the server machine and wait for result of the data from server. This concept will work successfully when the incoming data speed is slower than the batch processing speed and when the result is useful even with the delay. Because of new sources of Big Data like social media and mobile applications, the batch processing systems are not working effectively.

But to solving this problems now Stream Processing is introduced which enables the real time stream processing to the server for working rapidly on Big Data, in nonstop pattern the processing is done and the results are only useful if they arrived in very short time.

1.1.3. Variety of Data

Big Data contains various types of data like structured, unstructured and semi-structured data. Structured data is data which has predefined size and layout for Big Data. For examples of structured data contains numbers, string, dates, RDBMS tables and so on. Unstructured data is data which does not contain a predefined size and layout for Big Data. 20% of data is in structure form and remaining 80% is in unstructured form. For examples of unstructured data contains images, atmospheric data, video, Radar data, social media data and so on.

Semi-structured data is data which lies between structured and unstructured data. It is not formed in complex manner because of this analysis become easy. For examples of unstructured data contains JSON, XML and so on [6].

II. Big Data Analytics

Big Data analytics is the process of investigating of large amount of data sets which containing a different variety of data such as structured, unstructured and semi-structured to reveal or finding these kind of hidden data such as market trends, customer preferences and other useful information related to business[4]. Because of analytics we can find most effective information about business marketing, better customer services, improves operational performance and other business related benefits.

One of the goals of Big Data analytics is to enabling better decisions making for business to companies by using various prediction techniques and analytical professional person to analyze the huge amount of information of transaction data [5].

This also contains the data related to server logs, text from emails, sensor data, internet clickstream data and social networking data [7].

2.1. Big Data Analytics using RDBMS

The Relational Database Management System (RDBMS) is a solution for managing small amount of data. It uses Structured Query Language (SQL) to definition and modification of the database. The size of data is increasing continuously in last few years. And now size is increased near Peta Bytes i.e. 1 PB = 1024 TB. To handling this much huge amount of data it becomes more challenging. Need more hardware and more central processing units to processing this huge data.

Figure 2. System with RDBMS.

And out of the incoming data mostly maximum in semi-structured form or unstructured form and this data obtain from audio, video, texts, emails and social media. And when we are thinking about unstructured data so at that the RDBMS cannot handle this unstructured data. Because RDBMS is introduced to organize and manage only structured data like tables record and banking data. And also Big Data has high volume of data and it requires very high velocity of processing. RDBMS can't handle high velocity because it is developed for only steady data operation rather than fast processing. But when we are going to handle Big Data by using RDBMS then it become very costly.

2.2. Big Data Analytics using Hadoop

Real-time processing challenges are very complicated. We all know Big Data is generally categorized according to volume, variety and velocity of the data, and Hadoop framework can only handles the volume and variety only[7]. But with the volume and variety, the real-time processing requires handling fast processing that is high velocity of Big Data. And maintaining the velocity of data in Big Data is very complicated task.
We need first, the system which can able to gather the data generated by real-time stream events which are coming in the rate of millions of message streams per second. And along with this we need parallel processing of this data and when it was gathered. The Real-time Processing have very low latency period of processing so because of this computation and calculations are become very fast and rapid.

2.3. Big Data Analytics using Proposed System

We all know that Hadoop is like Batch Processing system is become popular in last few years for efficient offline data processing for Big Data. Hadoop is very high latency system which can process a large amount of data using parallel processing method called MapReduce. But there are many area where we need fast and real-time processing for better prediction and effective decision making like credit card fraud detection, security predictions, network fault prediction through sensor data and so on [1].

Real-time processing performs analytics in very short time. Because of correlation and prediction of event message streams which are generated in last few seconds. And to achieving Real-Time processing we need stream processing and here we are using two popular open sources software's for stream processing of Big Data for Real-time analytics i.e. Apache Kafka and Apache storm. In this we are collecting various incoming stream data from different sources by using Apache Kafka which is message broker system and then process that data by using apache storm which is stream processing engine [5].

2.3.1 Batch processing vs. Stream processing

Batch processing is very useful when we are processing large amount of data where the incoming data are collected, processed and getting result from that processing which is in the form of batches (groups). But here the processing time is very high. In this processing of data and computational capacity of the system is totally depend on the size of data and output processing required high latency and output is delayed [5].

Now stream processing contains nonstop input and outputs of data. And it is totally emphasizes on velocity of the incoming data, the incoming data can be processed in very small time or in the real-time situation. Stream processing provides ability for instant decision making of continuous data processing.

Figure 4. Batch Processing System

In Figure 4 we are describing the working of batch processing. Here suppose we are finding critical events from the continuous flow of events. Red color events i.e. T2 is a critical event, and in batch processing we are processing the data according to size or the particular time of batching. Here we are making batches of four events, that means when we are getting four record then we are going for processing on that data. And after the processing we find the critical events. But here T2 is critical event which arrives at secondly but we need to wait for forming the batches to processing data and finding these critical events.

Figure 5. Stream Processing System

In Figure 5 we are describing the working of Stream processing. Here we are overcome the problem of batch processing. Because in batch processing we need to wait till the formation of batches and then after the actual processing will start. But in stream processing whenever the events are arrives at that time they will go for processing so here we don't need to wait for any batching and grouping. Because of this the T2 critical event is identified very earlier.
III. SYSTEM ARCHITECTURE OF PROPOSED SYSTEM

Here in this system architecture we are using Producer who is responsible for generating inputs for our system like users or computer program can generated input in continuous flow. And after that this incoming data is queuing in the Kafka message broker queue, Kafka is broker which is interact between producer and consumer of the data.

![System Architecture of Proposed System](image)

After this we have two branches one is batch processing and another is for stream processing. In batch processing we are collecting the whole incoming data of the day or this branch is work on daily basis. Means the analyzer or analytics can be done at the end of the day. So we use hadoop for batch processing and hive for managing data over HDFS storage. And in Stream processing we are selecting those data or events which are critical i.e. predefined critical. Strom is used for stream processing of Big Data, so whenever predefine critical data or events are arrives then storm will process to that data and generated analysis report within few seconds. This is analysis data which are generated in last few seconds. So this is fast as compare to batch processing.

For Examples the incoming data is in the form of events no. as shown in Figure 6 and from this input data storm will collect the critical data (Red colored Events) and generate reports immediately and hadoop will collect all the incoming events or data and generate report at the end of the day.

3.1. Apache Kafka

Apache Kafka is a open source distributed messaging system which is invented by LinkedIn for LinkedIn activity stream processing. There are many issues of traditional messaging system overcome by this. Every messaging system has 3 parts i.e. message producers, message consumer and message broker [5]. The message broker is responsible for generating inputs to message broker, message broker interact with producer and consumer who can use the incoming generated data by producer through Kafka. And this inter process communication uses asynchronous queue for processing. We all know traditional messaging system support point-to-point communication, in that there only one producer produce data and send that data to queue and only one consumer can access that message at a time. But in Kafka there are multiple producer who produced data and they can will send the messages to Kafka queue. And queue sent that messages to multiple consumers [8].

3.2. Apache Storm

Apache Storm is a also distributed processing engine which is used for stream processing. Storm is applicable in many area such as real-time analysis, online machine learning, Distributed RPC, Continuous computing and more[5]. Storm is very fast, and it processed million of tuples per seconds at per node. And it is fault tolerant, scalable and guaranteed data processing. Storm is similar to hadoop but basic difference is that in hadoop we are executes the Map Reduce jobs and in storm we are executes topologies. Topologies are nothing but the combination of Bolts and spouts. Spouts is like gate to the incoming data where from they are enter data are in the stream form i.e. Stream is unbounded series of tuples. And Bolts can do processing on those incoming streams. Thus multiple topologies can use for obtaining effective parallel processing [9].

IV. OBJECTIVE OF THIS PROPOSED SYSTEM

We are here using this two approaches of processing i.e. batch and stream processing we need to analysis the data on daily basis and the seconds or minutes basis.

We are living in fastest century where we don't need delay for any processing even when we are surfing internet that time we don't want suffers late webpage's loading. So for this we can use this proposed system to troubleshooting the problems by identifying problems as soon as possible.
Figure 7. Problem Troubleshooting with hadoop.

For examples in Figure 7, here is problem occurred at 12:15:30 pm this a time of event occurrence to system. Then after within few seconds i.e. 12.15.32 pm the analyzer who is responsible for analysis the incoming data can find problem. And after that they will inform to Troubleshooting Team (TS Team) then TS team allocating troubleshooters to that problem according to near positions at 12:15:34 pm and troubleshooters visit to the area where problems are occurred within a few minutes i.e. at 12.20 pm. Here within 5-6 seconds the problem is resolved by using stream processing which based on minute or seconds basis analysis. In Traditional system whenever problems occurred to any system our analysis team can find that problem at the end of the day because they using batch processing.

V. CONCLUSION

Hence we take overviews of Big Data, Real-time Processing and techniques of analysis of big data.

But any system we need fast processing for that results of processing is play important role. As we are see in the handling of Big Data using many ways i.e. RDBMS, Hadoop and Proposed system. And we are studied difference between batch and stream processing.

But every system has drawbacks and advantages so we are developing best one for better performance. Here we are proposed an most popular architecture of stream processing by using Storm and Kafka, which is very efficient and scalable real-time processing system, it can handles large scale of input stream and processing them in very short period.

REFERENCES

[1] Toward Scalable Systems for Big Data Analytics: A Technology Tutorial, Han Hu1, Yonggang Wen2, Tat-Seng Chua1,And Xuelong Li3.2014
[4] Big Data analytics how an agile appoarch helps you quickly realise the value of bgi data, 2016 ThoughtWorks, Inc