ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

A Study on Real Time Big Data Analytics

S.Venkata Krishna Kumar, K.S.Ravishankar

Associate Professor, Department of Computer Science, P.S.G College of Arts and Science, Coimbatore, TamilNadu, India

Research Scholar, Department of Computer Science, PSG College of Arts and Science, Coimbatore, TamilNadu, India

ABSTRACT: Big data is the developing technology for handling large volume of different variety data at a high velocity, but these are handled only offline which would help in predicting required solution, but it does not help in on time decision making. because real time results give more accurate and on time decision to solve the problem with at most precision, This problem can be addressed by real time big data analytics, this paper describes about Real Time Big Data Analysis (RTBDA), its types, uses ,its implementations , companies that use and provide as service as well as the tools used.

KEYWORDS: Real Time, Big Data, Predictive Analysis, Smith's five-phase process model

I. INTRODUCTION

The word Real time stands for processing streamed data in motion and analysing it on time, rather than storing the data as it arrives and analyze at some point of time. "Real-time big data isn't just a process for storing petabytes or exabytes of data in a data warehouse, it's about the ability to make better decisions and take meaningful actions at the right time."says Michael Minelli, co-author of *BigData, Big Analytics.*

Big data contains complex and unstructured data which are hard to analyse, Analysing involves drawing user behaviours or data variations and projecting the behavioural pattern that will help in easy decision making, There are two types of big data processing, they are Batch and Real time, and Batch processing involves storing and analysing the historical data in a step by step process while Real time processing involves analysing stream of data and real time prediction. There are two types of Real time and Near Real time analysis, difference between the two is simple that real time queries comes with limitation result must be produced with in limited time period (may be microseconds), that is time is a major constraint where as in near real time, queries executes within seconds or minutes but time is not a major constraint.

II. RELATED WORK

In [1] author explains Big Data Analytics and explains predictive analysis using IBM InfoSphere as example, where as [2] gives an overview about challenges of big data and explains Hadoop as a important batch processing tool for big data processing along with some Real time examples. Paper [3] justifies Real time processing have an edge over traditional batch processing using two tools apache Kafka and storm which explains the Real time big data processing. The smith`s five phase model is briefly explained in [4], where as [5] explains the various challenges are addressed in resolving the Real Time Big Data Analytics.[6] gives variations among the two real time analysis tools storm and spark.[7] and [8] gives the vendors who are successfully providing real time analysis as service.

III. CHALLENGES

Real Time data processing involves very complex challenges, Big Data Analytics that involves processing of the complex and massive datasets, This data sets are different from structured data (which is stored in relational database systems) in terms of five parameters –variety, volume, value, veracity and velocity (5V's in figure 1 with its attributes) are the challenges of big data management are:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

1. *Volume*

Data is ever-growing day by day of all types ever Kilo Byte, Mega Byte, Peta Byte, Yotta Byte, Zetta Byte, Tera Byte of information. The data results into massive files. Excessive volume of information is main issues of storage. This main issue is resolved by reducing storage value**.** Data volumes are expected to grow more than 50 times by 2020.

2. Variety

Data sources (even in the same field or in distinct) are extremely heterogeneous. The files comes in various formats and of any type, it may be unstructured or structured such as text, audio, log files, videos and more. The varieties are endless, and the data enters the network without having been quantified or qualified in any way.

3. *Velocity*

The data comes at high speed. Sometimes one minute is too late so big data is time sensitive. Most organisations data velocity is main challenge. The credit card transactions and social media messages done in millisecond and data generated by this putting in to databases.

4. *Value*

Which addresses the requirement for valuation of enterprise data? It is a most important V in big data. Value is main buzz for big data because it is important for IT infrastructure system, businesses to store large amount of values in database.

5. *Veracity*

The increase in the range of values typical of a large data set. When we dealing with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be dirty data. Big data and analytics technologies work with these types of data.

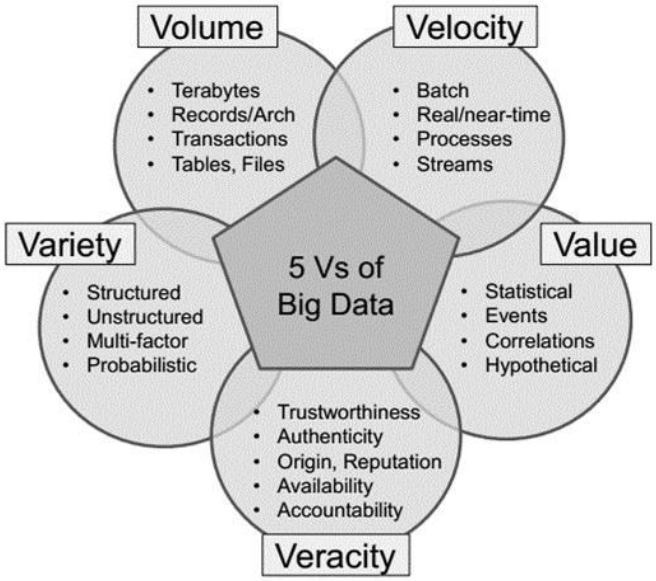

FIGURE 1 BIG DATA PARAMETERS

But Hadoop like systems handles only the Volume and Varity part. but real time big data analytics handles all, Handling all involves three following steps, they are ,at first step the data must be collected from real time events streams coming in at a rate of millions of events per seconds. Then at Second step, data collected need to be parallel processed as quickly as possible then at the Third step, it should perform event correlation using a Complex Event Processing engine to extract the meaningful information from the stream of data in motion, These three steps must be

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

fault tolerant and distributed. The real time system should be a low latency system to process data very fast to enable a near real time response but further faster will unlock Real time Capability with results delivered in less than second.

RTBDA is efficient when the possible factors are improvised, they are

- Relevant Data
- Data Classification methodology
- Data Analysis
- Data filtering(Reducing Noise)
- Data Storage (in case of referring historical data)

IV. HOW RTBDA WORK?

The heart of any prediction system is the Model. There are various Machine Learning algorithms available for different types of prediction systems. Any prediction system will have higher probability of correctness if the Model is built using good training samples. This Model building phase can be done offline. For instance, a credit card fraud prediction system could leverage a model built using previous credit card transaction data over a period of time. Imagine a credit card system for a given credit card provider serving hundreds of thousands of users having millions of transactions data over given period of time; a Hadoop-like system is in need to process them. Once built, this model can be fed to a real time system to find if there is any deviation in the real time stream data. If the deviation is beyond a certain threshold, it can be tagged as an anomaly.

Real-time big abstracts analytics is an accepted action involving assorted accoutrement and systems. Smith says that it's accessible to bisect the action into 5 phases: data distillation, model development, validation and deployment, real-time scoring, and model refresh. At anniversary phase, the agreement "real time" and "big data" are aqueous in meaning. The definitions at anniversary appearance of the action are not carved into stone. Indeed, they are ambience dependent. Like the technology assemblage discussed earlier Smith's five-phase process is devised as a framework for predictive analytics. But it as well works as a accepted framework for real-time big data analytics.

- *Data distillation* — like unrefined oil, abstracts in the data layer is awkward and messy. It lacks the anatomy appropriate for architecture models or assuming analysis. The data distillation phase includes extracting appearance for baggy text, accumulation of disparate data sources, clarification for populations of interest, selecting accordant features and outcomes for modeling, and exporting sets of distilled data to a bounded data mart.
- *Model development* — Processes in this phase cover affection selection, sampling and aggregation; variable transformation; model estimation; model refinement; and model benchmarking. The goal at this phase is creating a predictive model that is powerful, robust, apprehensible and implementable. The key requirements for data scientists at this phase are speed, flexibility, productivity, and reproducibility. These requirements are critical in the context of big data: a data scientist will construct, clarify and analyze dozens of models in seek for an able and robust real-time algorithm.
- *Validation and deployment* — the main goal at this phase is testing the model to accomplish abiding that it works in the absolute world. The validation process involves re-extracting beginning data, running it in adjoining model, and comparing after-effects with outcomes run on data that's been withheld as a validation set. If the model works, it can be deployed into a production environment.
- *Real-time scoring* — in real-time systems, scoring is triggered by accomplishments at the decision layer (by consumers at a website or by an operational arrangement through an API), and the absolute communications are brokered by the integration layer. In the scoring phase, some real-time systems will use the same data that are used in the data layer, but they will not use the same data. At this phase of the process, the deployed scoring rules are "divorced" from the data in the data layer or data mart. Note as well that at this phase, the limitations of Hadoop become apparent. Hadoop today is not decidedly adapted for real-time scoring, although it can be used for "near real-time" applications such as clearing large tables or pre-computing scores. Newer technologies such as Cloudera's Impala are advised to advance Hadoop's real-time capabilities.
- *Model refresh* — Data is consistently changing, so there needs to be abroad to brace the data and brace the model built on the original data. The absolute scripts or programs used to run the abstracts and build the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

models can be re-used to brace the models. Simple exploratory data analysis is as well recommended, along with alternate (weekly, daily, or hourly) model refreshes. The refresh process, as able-bodied as validation and deployment, can be automatic application web-based casework such as RevoDeployR, a allotment of the RevolutionR Enterprise solution.

Data is always changing, so there needs to be away to refresh the data and refresh the model built on the originaldata. The existing scripts or programs used to run the data andbuild the models can be re-used to refresh the models. Simpleexploratory data analysis is also recommended, along with periodic (weekly, daily, or hourly) model refreshes. The refresh process, as well as validation and deployment, can be automated using web-based services such as RevoDeployR, a part of the RevolutionR Enterprise solution.

Refreshing the model based on reinvesting the data and re-running the scripts will only work for a limited time, since the underlying data — and even the underlying structure of the data — will eventually change so much that the model will no longer be valid. Important variables can become non-significant, non-significant variables can become important and new data sources are continuously emerging. If the model accuracy measure begins drifting, go back to phase 2 and re-examine the data. If necessary, go back to phase 1 and rebuild the model from scratch.

V. WHEN RTBDA USED?

Real Time Big Data Analytics is useful in Credit card fraud analytics, network fault prediction from sensor data, security threat prediction, stock analysis, recommendation systems, Emergency situation analysis, Defense purposes and Situation aware Systems, even in traffic data analysis. It is employable in all real time situations were critical data need to be analysed.

VI. WHO USES RTBDA?

Google Analytics

Google Analytics is a freemium web analytics service offered by Google to track and report website traffic of client's website. After acquiring Urchin, Google launched this service in November 2005. Google Analytics includes Google Analytics Content Experiments, Google Analytics Cohort analysis and Google Analytics e-commerce. Google Analytics Content Experiments helps in standard and multi variant testing, Google Analytics Cohort analysis feature helps to understand the behavior of component groups of users apart from the given user population. Google Analytics e-commerce helps in analyzing revenue, transaction and e commerce related things. It is very much beneficial to marketers and analysts for successful implementation of Marketing Strategy

DRDO NETRA

NETRA (NEtwork TRaffic Analysis) is a developed by India's Centre for Artificial Intelligence and Robotics (CAIR), a Defense Research and Development Organization (DRDO) laboratory, and is used by the Intelligence Bureau, India's domestic intelligence agency, and the Research and Analysis Wing (RAW),NETRA can analyze voice traffic passing through software such as Skype and Google Talk, and intercept messages with keywords such as 'attack', 'bomb', 'blast' or 'kill' in real-time from the enormous number of tweets, status updates, emails, instant messaging transcripts, internet calls, blogs, forums and even images generated on the internet to obtain the desired intelligence. This system with RAW analyses large amount of international data which crosses through the internet networks in India.

Amazon Kinesis

Amazon Kinesis is a fully managed, cloud-based service for real-time data processing over large, distributed data streams. Amazon Kinesis can continuously capture and store terabytes of data per hour from hundreds and thousands of sources such as website clickstreams, financial transactions, social media feeds, IT logs, and location-tracking events. With Amazon Kinesis Client Library (KCL), you can build Amazon Kinesis Applications and use streaming data to power real-time dashboards, generate alerts, and implement dynamic pricing and advertising, and more. You can also emit data from Amazon Kinesis to other AWS services such as Amazon Simple Storage Service (Amazon S3), Amazon Redshift, Amazon Elastic Map Reduce (Amazon EMR), and AWS Lambda. Amazon Kinesis allows for real-time data

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

processing of continuously collected data as it is generated and which enable earlier response to critical information about your business and operations.

VII. RTBD TOOLS: SPARK OR STROM

Storm, a distributed computation framework for event stream processing, began life as a project of BackType, a marketing intelligence company bought by Twitter in 2011. Twitter soon open-sourced the project and put it on GitHub, but Storm ultimately moved to the Apache Incubator and became an Apache top-level project in September 2014. Storm has sometimes been referred to as the Hadoop of real-time processing. The Storm documentation appears to agree: "Storm makes it easy to reliably process unbounded streams of data, doing for real time processing what Hadoop did for batch processing."

Apache Storm is a free and open source distributed real-time computation system. Storm makes it easy to reliably process unbounded streams of data, doing for real-time processing what Hadoop did for batch processing. Storm is simple, can be used with any programming language. Storm has many use cases: real time analytics, online machine learning, continuous computation, distributed RPC, ETL, and more. Storm is fast: a benchmark clocked it at over a million tuples processed per second per node. It is scalable, fault-tolerant, guarantees all data will be processed, and is easy to set up and operate. Storm integrates easily with the all queuing and database technologies. A Storm topology consumes streams of data and processes those streams in arbitrarily complex ways, repartitioning the streams between each stage of the computation however needed.

Spark suited to real-time distributed computation, this was started out as a project of AMPLab at the University of California at Berkeley before joining the Apache Incubator and ultimately graduating as a top-level project in February 2014. Spark supports stream-oriented processing, but it's more of a general-purpose distributed computing platform. As such, Spark can be seen as a potential replacement for the MapReduce functions of Hadoop, while Spark has the ability to run on top of an existing Hadoop cluster, relying on YARN for resource scheduling. In addition to Hadoop YARN, Spark can layer on top of Mesos for scheduling or run as a stand-alone cluster using its built-in scheduler.

Apache Spark is a fast and large-scale data processing engine. Spark has an advanced DAG execution engine that supports cyclic data flow and in-memory computing which enables to run programs up to 100x faster than Hadoop`s MapReduce in memory, or 10x faster on disk. Spark runs on Hadoop, Mesos, standalone, or in the cloud. It can access diverse data sources including HDFS, Cassandra, HBase, and S3. DataFrames and SQL provide a common way to access a variety of data sources, including Hive, Avro, Parquet, ORC, JSON, and JDBC. Spark SQL includes a cost-based optimizer, columnar storage and code generation to make queries fast. At the same time, it scales to thousands of nodes and multi hour queries using the Spark engine, which provides full mid-query fault tolerance. Spark SQL lets query structured data inside Spark programs, using either SQL or a familiar DataFrame API. It is also usable in Java, Scala, Python and R languages. Spark Streaming module helps in Real time processing.

VIII. CONCLUSION

In this article, an overview of Real Time Big Data's concept, tools, techniques, applications, advantages and challenges have been reviewed. The results have given away that regardless of the fact that accessible information, tools and techniques available in the literature, there are numerous focuses to be viewed as, discussed, analyzed, developed, and improved, and so on. Although this paper obviously has not resolved the complete subject about this substantial topic, emphatically it has provided some useful discussion and we can conclude that Strom is possibly one of the best solutions to the Real Time Big Data Analytics.

REFERENCES

1. Kiranmai munagapati, D.Usha Nandhini, “Real Time Data Analytics” ,International Journal of Applied Engineering Research ,ISSN 0973-4562 Volume 10, Number 3 (2015) pp. 7209-7214
2. D. Rajasekar , C. Dhanamani , S. K. Sandhya “A Survey on Big Data Concepts and Tools” IJETAE ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 2, February 2015

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

3. Manidipa Mitra, Dibyendu Bhattacharya, "Analytics On Big Fast Data Using Real Time Stream Data Processing Architecture"
4. Mike Barlow "Real-Time Big Data Analytics: Emerging Architecture" Copyright © 2013 O'Reilly Media. ISBN: 978-1-449-36421-2
5. http://davebeulke.com/big-data-impacts-data-management-the-five-vs-of-big-data/
6. Andrew C. Oliver "Storm or Spark: Choose your real-time weapon" www.infoworld.com
7. Mikael Ricknäs "Amazon adds Kinesis real-time data analysis service" IDG News Service
8. https://cloud.google.com/solutions/articles#bigdata
9. Big Data steps to Real-Time Analysis www.insightssuccess.com