International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 8, ISSN No. 2455-2143, Pages 213-216
Published Online December 2019 in IJEAST (http://www.ijeast.com)

EMERGING REAL TIME STREAMING ANALYTICS PROCESSING USING HADOOP FRAMEWORK

Sneha. S
M.Tech Scholar,
Dept. Of. Computer Science & Engineering,
BMS Institute of Technology & Management,
Yelahanka, Bengaluru

Anjan K Koundinya
Associate Professor and PG Coordinator,
Dept. Of. Computer Science & Engineering,
BMS Institute of Technology & Management,
Yelahanka, Bengaluru

Abstract— ***Sensors, machines, vehicles, cell phones, web-based social networking systems and other constant sources are creating persistent stream of information. This information is utilized by the organizations to get the advantage from those information. A real time streaming system should address the issues of researchers, developers and records focus activities groups without requiring complex code for incorporation of numerous outsider tools. As there is increment in measure of information that is produced and gathered, statistical analysis wants adaptable, flexible, and high performance tool to analyze and obtain only the necessary data from the large growing data in a required timely manner. Hadoop Distributed File System (HDFS) is one of the file system to store huge measure of information. HDFS can oversee and keep up information in a dispersed manner. Real Time Streaming data can be put away into noSQl databases, for example, Mongo DB and Hive. Enormous information investigation can be performed on information put away on Hadoop distributed file system utilizing Apache Hive, Tez , Storm, Flume and Apache Presto. Hive is a environment which is over Hadoop (Map Reduce), and gives more significant level language to apply to the Hadoop's fundamental part Map Reduce to process the information. The key focal points of this methodology are it can equipped for processing and saving of the enormous measure of information. It additionally can adapt to the a large number of client demands at the same time. It can give the scalability to the machine is increasingly attractive with the guide of including new nodes. Incorporating the Visualization equipment with Big Data projects will give the gigantic picture to the clients to see the bits of knowledge of the Big data. It can offer the analytical reports for giving the big view about the file system..***

Keywords— **Real Time Streaming, HDFS, Hive, Tez, Storm, Flume, Storm, Apache Presto, Mongo DB, Big Data**

I. INTRODUCTION

As we are moving towards digitization, the amount of data being made and amassed is developing and expanding radically. Investigation of this enormously developing data will turn into an difficult task if we utilize the current tools. We expect restructuring to connect the gap between data being created and data that might be analysed accurately.
Enormous big data tools and innovations offer openings and requesting circumstances in having the option to investigate data successfully to understand client needs, increase an upper hand in the commercial center and build up association's business undertaking. Data management architectures have created from the data warehousing model to increasingly convoluted structures that address more noteworthy necessities, comprising of real-time and batch processing, structured and unstructured data, high-speed transactions etc. Investigating huge fact sets calls for colossal process limit that can change long essentially dependent on the amount of input data and the kind of analysis.

1.1 Real-Time Streaming

Internet of Things (IOT) has made a fresh out of the box new age as it is utilized all over the place. A huge measure of detecting devices(sensors) gather/produce different type of certainties after some time for a colossal scope of fields and applications. In view of the idea of the application, these gadgets will bring about gigantic or quick/continuous information streams. Applying examination over such information streams to find new actualities, anticipate future bits of knowledge, and settle on oversee decisions is a basic procedure that makes IoT a value stage for offices and a personal satisfaction improving technology[3]. For example, robots were being used for a considerable length of time in assembling, yet now they have extra sensors so we can perform quality confirmation, never again essentially meeting. For quite a long time, mechanical measures have been typical

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 8, ISSN No. 2455-2143, Pages 213-216
Published Online December 2019 in IJEAST (http://www.ijeast.com)

in numerous enterprises comprising of concoction mixes and utilities. Presently checks are supplanted with the guide of computerized sensors and "smart meters" to give continuous observing and examination. GPS and RFID flag currently exude from cell contraptions and property, running from advanced cells to vehicles to transportation beds, so a greater part of these can be followed progressively and controlled absolutely [1].

1.2 Streaming Analytics

The developing agreement is that examination is the most immediate course to business endeavor worth drawn from new assortments of enormous information, which consolidates spilling information. Existing diagnostic procedures dependent on mining, measurements, predective algorithms, questions, scoring, clustering, etc practice well to device information once it is caught and saved.

More up to date instruments are reengineering these and making new scientific strategies in order to work on data that streams always in addition to on different information on other storage. As devices and file system associated with the Internet of Things produce spilling data on a worldwide scale, organizations need complex systems for distinguishing and utilizing this information in new applications. That may require the combination and investigation of information from internal legacy sources, current client produced information and outside sources from accomplices and data service providers. The data management and analytics tool of a definitive decade might have the option to deal with the degree of information, anyway they truly can't hold up with the speed at which it ought to be handled. A variety of vendors are dashing to satisfy this requesting circumstances with a new generation of information management and analytics tools.From query optimization solutions to integration techniques to new platforms for device interoperability, there is no deficiency of development in this space today.

II. REAL WORLD USE INSTANCES FOR STREAMING ANALYTICS

1. Observe and preserve the availability, overall performance, and capability of interconnected infrastructures, which include utility grids, computer networks, and production facilities.
2. Can improve the client experience by understanding the client behavior across multiple channels.
3. Device Telemetry.
4. Infrastructure monitoring.
5. Recognize consistence and security breaks, if it halts correct them immediately.
6. Spot and stop fraud activity, while it is being executed.
7. Inventory management.
8. Web analytics/Content management.

III. REAL TIME STREAMING PLATFORMS

3.1. Apache Spark:

Apache Spark Streaming is a scalable fault-tolerant streaming processing system that natively helps both batch and streaming workloads. Spark Streaming is an expansion of the center Spark API that permits data engineers and data scientists to way real time actualities from different assets together with (yet at this point not confined to) Kafka, Flume, and Amazon Kinesis. This data which is obtained after the processing can be placed on the information file systems, databases, and live dashboards. Its key deliberation is a Discretized Stream or, in snappy, a DStream, which is the data divided into small batches. DStreams are based on RDDs, Spark's core data abstraction. This enables Spark Streaming to consistently join with some other Spark segments like MLlib and Spark SQL.

Spark Streaming is not the same as different structures that either have a preparing engine planned best for processing, or have comparable batch and streaming APIs however collect internally to distinct engines.APIs anyway gather inside to particular engine. Spark's single batch execution engine and brought together programming model for batch processing and streaming purpose has various advantages over existing engines for streaming structures. Specifically, four significant viewpoints are:

- Fast recovery in case of failures.
- Better Load balancing and usage of resources.
- Interactive queries for combining streaming data and statistics data set.
- Integration with advanced processing libraries (SQL, AI, graph preparing)[1].

3.2. Apache Hive:

Hive HCatalog Streaming API:

Traditionally adding new information into Hive calls for gathering a big quantity of statistics onto HDFS and then periodically adding a brand new partition. This is essentially a “batch insertion”. Insertion of latest data into an present partition isn't authorized. Hive Streaming API allows information to be pumped constantly into Hive. The incoming data may be constantly committed in small batches of records into an present Hive partition or table. Once information is dedicated it becomes straight away seen to all Hive queries initiated in the end.

This API is supposed for streaming clients along with Flume and Storm, which constantly generate data. Streaming support is built on top of ACID primarily based insert/replace support in Hive.

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 8, ISSN No. 2455-2143, Pages 213-216
Published Online December 2019 in IJEAST (http://www.ijeast.com)

The Classes and interfaces part of the Hive streaming API are broadly categorized into two sets. The first set provides support for connection and transaction management while the second set provides I/O support. Transactions are managed by the meta store. Writes are performed directly to HDFS.

Streaming to unpartitioned tables is also supported. The API supports Kerberos authentication starting in Hive 0.14 [5].

3.3. Apache Storm:

Storm is a distributed real-time processing system for streaming it performs similar to the Hadoop batch processing system. It can be used for real-time data analysis, machine learning, continuous computation etc. It is not specific to a programming language it is independent of the programming language. It runs on Hadoop YARN and it can be used with Flume to store the processed information on HDFS. Storm is also used by the likes of WebMD, Yelp, and Spotify. Apache Storm programs are designed as to be the directed acyclic graphs(DAG). Storm is designed to perform computation on the unbounded streams, and any programming language can be used for it. It has been benchmarked at processing over 1,000,000 data tuples per second per node basis, and guarantees the processing of the job. Apache Storm can be used with different applications such as distributed machine learning, real time processing of the information and it is highly suitable for applications with large amount of data. Storm can run on YARN and integrate into Hadoop ecosystems, supplying existing implementations a solution for real-time stream processing[2].

3.4 Apache Samza

Samza is a distributed framework for processing stream/flow of data, that is based on Apache Kafka and YARN. It gives an API similar to the mapreduce which is a call back API, and it also has the capability of snapshot management and fault tolerance in a resistant and scalable manner. Samza manages snapshotting and restoration of a stream processor's state.It restores its state to the consistent snapshot state when the processor is restarted. It also handles the huge amount of state. Whenever a node in the cluster fails, it interacts with YARN to safely handlover the task on the failed node to some other node. Samza makes use of Kafka to make sure that the messages are taken care in the order they are posted to the partition, so that the messages posted to the partition are never lost. Samza is partitioned and distributed at each level. Kafka provides ordered, partitioned, replayable, fault-tolerant streams. YARN acts as a disbursed environment for Samza packing containers to run in [1].

3.5 Flume

Flume is a system for efficiently collecting, aggregating, and moving big amounts of event data. It is a distributed , reliable and available system. Flume contains many number of agents where each of the agent executes in a separate Java Virtual Machine (JVM). Each agent consists of three components which are pluggable, named source, sink and channel. Source collects the data that is incoming as events, sink writes events out, and channels provide connection between source and sink. In case of failure or shutdown channel stores the event in the data buffer in each agent. Flume contains file-based and memory based channels. The data aggregation is logged by the Flume. It can be used for transporting large amount of data which is obtained by social media, network traffic and emails, these data can also be customized[1].

3.6 Scribe

Streaming data which is logged can be aggregated by using Scribe. It can scale to large number of nodes ,it is very robust and fault tolerant. Scribe contains a central server which receives the aggregated message from each node which runs the scribe. If the central scribe server goes down scribe will write all the messages to the local disk of that agent and write backs to the central server after its recovery . After the message is available at the central scribe server it writes it to the destination file or else it writes to the another level of scribe servers[1].

3.7 HStreaming

HStreaming is developed on top of Hadoop and mapreduce it is an analytical framework. Data Acquisition and Data Analytics are the two components of HStreaming. The real time streams and data are collected by data acquisition, it also has ETL capabilities , the analytics component is used for analyzing different types of data such as structured and unstructured data in a real time manner. The data from different databases can be analyzed by using the connectors, which provides connection between SQL and NoSQL databeses.HStreaming provides both enterprise and community editions[1].

3.8 Amazon Kinesis

Real time processing of streaming data can be done on cloud using Kinesis. It is a amazon service. It is also implemented with amazon services using connectors, such as S3, Redshift, and DynamoDB, for a complete Big Data architecture. Kinesis also provides a library which can be used for building applications and use stream data for dashboards, dynamic pricing or alerts it includes Kinesis Client Library (KCL) [1].

IV. CONCLUSION

There exists different solutions to find solution to the big data analytic requirements. To provide a complete solution big data architecture has many tools, this can help to meet the stringent business necessities within the most cost-optimized, performance, and resilient manner possible. The end result obtained is a versatile, massive information architecture this is able to scale in conjunction with your business on the worldwide infrastructure.

V. REFERENCES

[1]. Jayanthi. D, Dr.Sumathi.G.(2016).“A Framework for Real-time Streaming Analytics using Machine Learning Approach”, Proceedings of National Conference on Communication and Informatics-2016.

[2]. Surekha.D, Swamy.G,Venkatramaphanikumar.(2018). “Real Time Streaming Data Storage and Processing using Storm and Analytics with Hive”, 978-1-5386-4225-2/18/$31.00 ©2018 IEEE.

[3]. Mohammadi Mehdi,” Deep Learning for IoT Big Data and Streaming Analytics: A Survey”, IEEE COMMUNICATIONS SURVEYS & TUTORIALS , VOL. X, NO. X, XXXXX 201X.

[4]. André Leon Sampaio Gradvohl, Hermes Senger, Luciana Arantes, Pierre Sens.(2014).”Comparing Distributed Online Stream Processing Systems Considering Fault Tolerance Issues,Journal of Emerging Technologies in Web Intelligence, Vol 6, No 2 (2014), 174-179, May 2014,doi:10.4304/jetwi.6.2.174-179.

[5]. Rahnama A.H.A.(2014). “Distributed real-time sentiment analysis for big data social streams”, IEEE International Conference on Control, Decision and Information Technologies (CoDIT), (Nov 2014) page(s):789-794,doi:10.1109/CoDIT.2014.6996998”

[6]. Gianmarco De Francisci Morale.(2013). “SAMOA: A Platform for Mining Big Data Streams”, 22nd International Conference on WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil. ACM 978-1-4503-2038-2/13/05.

[7]. Mohit Maske, Dr. Prakash Prasad.(2015). “A Real Time Processing and Streaming of Wireless Network Data using Storm “, International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE),vol 5,Issue 1, Jan 2015.

[8]. DunrenChe, MejdlSafran, and Zhiyong Peng.(2013). “From Big Data to Big Data Mining: Challenges, Issues, and Opportunities”, DASFAA Workshops 2013, LNCS 7827, pp. 1–15, 2013.

[9] Gruenheid, Anja, Edward Omiecinski and LeoMark(2011).”Query optimization using column statistics in hive Proceedings of the 15th Symposium on International Database Engineering & Applications-IDEAS 11IDEAS11”, 2011.

[10] http:// www.adhocshare.tk