DOI:10.21276/sjeat

Saudi Journal of Engineering and Technology
Scholars Middle East Publishers
Dubai, United Arab Emirates
Website: http://scholarsmepub.com/

ISSN 2415-6272 (Print)
ISSN 2415-6264 (Online)

Review Article

Beyond Hadoop: The Paradigm Shift of Data From Stationary to Streaming Data for Data Analytics

Affreen Ara[1], Dr Aftab Ara[2]*
[1]Computer Science and Engineering, BPUT University, Rourkela, India
[2]Faculty in School of Management Studies, Ravenshaw University, Cuttack, India

***Corresponding Author:**
Dr Aftab Ara
Email: affi2k7@gmail.com

Abstract: The paradigm shift of data from static to fast flowing data is an important move in the industry, to accommodate growing size of data. The velocity and volume of data are continuing to expand which has started to make its impact in business and other applications of Big Data. The paper describes the paradigm shift of data from static data to streaming data for data analytics beyond Hadoop. It describes how the first generation of Hadoop applications were largely built for batch-oriented paradigm . Streaming data is essentially different from traditional data handling patterns and comes with its own set of challenges and requirements. New applications such as Storm, Flume, Kafka, and other technologies are evolving to bring in an era of real-time analytics Data is generated incessantly from thousands of sources simultaneously and it can be of various type such as log files, mobile and web data, transaction etc. The sections of my paper are Introduction followed by Streaming data, Hadoop, Streaming data analytics, Apache Spark, Comparing Streaming models and Streaming analytics use cases followed by conclusion. The information presented is from secondary source i.e. journal, conference proceedings and websites.
Keywords: paradigm, Hadoop, Apache Spark.

INTRODUCTION

The concept of big data has been there for long time and has the benefit of speed and efficiency .Big data analytics examine huge volume of data and uncover patterns, correlations and insights. Companies have started to make balances between deep analysis of large data sets and faster time to get results. Data analytics helps in changing the way businesses are competing in the market, enabling them to make faster and better decisions based on real time. Real time data analytics is based on timeliness and faster result because everyday computer servers have to process peta-bytes of both structured and unstructured data in real time. Analytics helps organization to reduce cost, make better decision and provide better services to customers.

STREAMING DATA

Streaming data [3] is data that is continuously analyzed and transformed in memory by a computing platform before it is stored on a disk. It is traditionally different from data at rest. Hadoop platform uses data at rest, so primary difference is presence of velocity. In Hadoop cluster data is collected in batches and then processed. Streaming data is required to be processed sequentially and incrementally over time windows. It is used for different types of analytics such as statistical operations, predictive analytics, filtering and sampling techniques. Streaming data is generated from multitude of sources and data records are send simultaneously. There can be wide variety of data sources such as log file, customer profile, web searching history, ecommerce purchase, geospatial services, financial services, web clicks etc. Real time streaming data has same characteristics of big data, such as volume, variety, velocity and veracity. The streaming data can be time-stamped on arrival. Information collected from analysis gives organization insight into customer and business aspects. For example ecommerce sites can track customer by continuously analyzing searches over time and they also recommend new products for future purchase.

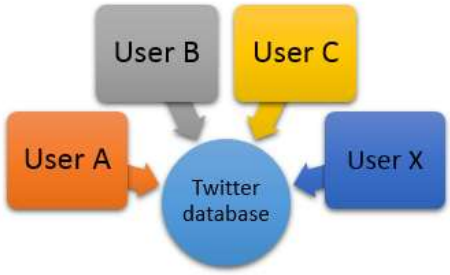

Fig-1: streaming data-continuous streams of data collected from users

Stream processing necessitates management of unpredictable rate of streaming data as the incoming streaming data might involve misplaced data or delays. Many time intensive processes require cautious handling of data stream linked to the system .Streaming data processing is helpful where new and dynamic data is application become more intricate .Over time these application begin to perform advance analytics such as machine learning, predictive analytics, event processing algorithm etc. Stream processing systems are capable of handling high volume of data in real time with a scalable, highly available and generated on repetitive basis .It is used in many industry sectors and big data uses cases. Most companies start with simple application such as collecting system log, reports and gradually these fault tolerant architecture. This enables analysis of data in motion.

Streaming data source example

- Financial trading companies track changes in stock in real time to compute possible risk and automatically rebalance portfolios to avoid disaster.
- Twitter social network is source of real time streams of message at maximum speed of 50 second. Sentiment analysis is used by corporates to determine opinion about product or any subject.
- Online gaming companies such as GameAnalytics collect streaming data about user's player-game interactions into their system in real time. After analyzing data about player they offers incentives and give rich experiences to engross its player's.

Difference between batch and streaming data.

Batch processing can be used to compute random queries over different sets of data. It is used to calculate results that are derived from all the data it incorporate .It can also do in depth analysis of big data sets. MapReduce-based systems, like Amazon EMR, are examples of platforms that support batch jobs. Stream processing requires consumption of sequence of data. The [4] system incrementally updates each metrics, reports, and summary statistics in response to each incoming data record. Spark is batch processing system but Streaming Spark is for streaming data. The following table shows difference between batch and streaming data.

Table 1: It shows difference between batch and streaming data

Property	Batch data	Streaming data
Data Size	Batches of data (Large size)	Individual records or micro batches
Performance	Latency measure in minutes or more	Latency measure in milliseconds or seconds
Analyses	Complex computations	Simple problems
Data Scope	All data available at once	Computes function of data element or small window of data

HADOOP

Hadoop was created by Doug Cutting,the architecture of Lucane. It has its origin in Apache Nutch, an open source web search engine which itself was the part of Lucane project. Hadoop is an open source software which supports storage and execution of vast amount of data. The core of Hadoop is the distributed file system and its programming framework runs distributed MapReduce programs. In today's world most of programming community use Hadoop for its ability to store, manage ,analyze unstructured and structured data in lesser time .Hadoop projects include Hadoop Yarn, Hadoop Common, MapReduce. There are many other Hadoop collaborated projects such Cassandra, Avro, Hive and many more.

Hadoop is a collection of frameworks that includes many sub components. The Hadoop Distributed File System makes it unique as it provides scalable, fault-tolerant, cost-efficient storage for big data storage. It was designed to accommodate large clusters of commodity servers counting up to hundreds of petabytes and thousands of servers. Open-source technologies need to evolve organically rather than

being purposefully designed to meet requirement of changing demand of industry. The overall capabilities of Hadoop continue to be shaped by this evolving core.

HDFS

Hadoop Distributed File System (HDFS) [5] is the core component of Hadoop framework. Data in a cluster is fragmented into smaller pieces (called blocks) and distributed throughout the cluster. By using this method the map and reduce functions can be executed on smaller subsets of larger data sets. This process provides the scalability that is needed for big data processing. MapReduce is the original framework for scripting enormously parallel applications that process huge amounts of data stored in HDFS.

An example of HDFS

A file contains the phone numbers for each person in that country. The people with a last name starting with A might be stored on server 1, B on server 2, and so on. In a Hadoop context, pieces of this phonebook would be stored across the cluster. In order to rebuild the entire phonebook, the program needs the blocks from every server in the cluster .When components fail, HDFS duplicates these smaller fragments onto two additional servers by default to achieve availability. .This problem of redundancy in Hadoop can be increased or decreased on a per-file basis or for a whole environment.

Hadoop Limitations.

Hadoop limitations are given below:

- **No support for real time data:** Hadoop[6] can be optimized to analyse both structured and non-structured data as it was designed for batch processing system. It cannot support fast data analysis on streaming data.
- **Multiple copies of already present data:** HDFS results in producing multiple copies of the data. At minimum, there are generally three copies of the data but while maintaining data locality more copies of the data are required which is already oversized..
- **Limited SQL support:** They have very limited SQL support as they lack such basic SQL functions such as sub queries, 'group by' analytics, etc.
- **Inefficient performance:** HDFS does not have concept of a query optimizer, so it does not have cost-based plan for execution. Hadoop clusters are generally significantly larger than would be required for a similar database due to this reason.
- **Challenging framework:** The MapReduce framework is disreputably problematic to leverage for more than simple transformational logic. Open source components which attempt to streamline this also use proprietary languages which is challenging.

BEYOND HADOOP

Real time Analytics

Real time data denotes the ability to process data on its arrival. Real time data analytics work well with predefined algorithm and SQL queries. Present day tools can synthesize multiple stream of flowing data and perform complex operation on them .Today there are many open source frameworks and tools options available for real-time analytics. The amalgamation of Spark Streaming, Kafka, and Cassandra has given exceptional results and is a good start for building real-time data pipelines. Kafka is a distributed streaming platform which uses publish-subscribe messaging system to capture and publish streams of data. Spark Streaming is an extension of the core Spark API that allows user to consume and process data in real-time from disparate event streams. Cassandra provides a scalable and resilient operational database for real-time analytics. Amazon Kinesis is a streaming platform for collecting and processing streams of data in real time.

The Need for Streaming Analytics

Organization[7] collect data from users to extract value data in real-time from variety of sources such as IoT devices, social networks, and online transactions. The data extracted from user is constantly monitored and acted upon quickly. There is an urgent need for large scale, real-time stream processing which is more evident than ever before ;for example when a user make a purchase online all the associated data (e.g. date, time, items, price) of the order needs to be stored .The data is used by online retailers to analyse and make quick decisions based on the customer's behaviour. Retailers continuously monitor user search history for future product recommendation to customer.

SPARK STREAMING

Spark Streaming application can consume both static and streaming data from variety of sources and they process data using Spark SQL and Data Frames. The system applies machine learning techniques from Machine Library MLlib, and send the results to data storage.

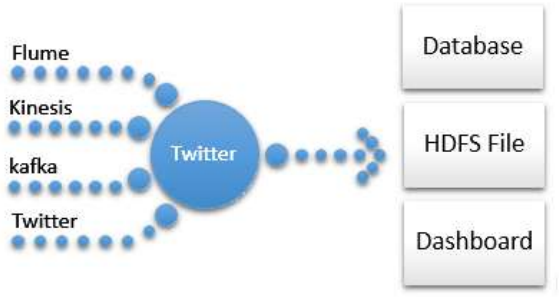

Fig-2: Architecture of Streaming Spark

Spark Streaming was added to Apache Spark in 2013, an extension of the core Spark API .The model works as follows.

- It process real-time data from various sources like Kafka, Flume, and Amazon Kinesis.
- Data is consumed from sources such as Kafka, Flume and more.
- The collected data is processed with algorithms with complex functions.
- The processed data in flushed out of system onto database, file system and dashboard.

Spark [9] is gaining mainstream in real world application such as Uber, Piniterest and video sharing platform Conviva. Other use case are maching learning, interactive analysis ,fog computing. There are many tools within Hadoop ecosystem, these tools are used in collaboration with other software tools to work and meet desired use case.

COMPARING STREAMING MODELS

Table 2: Different streaming models[11]

Criteria	Apache Gearpump	Streaming Spark	Apache Storm	Apache Flink	Twitter Heron
Streaming	Continuous, Checkpoint per batch	Micro-batch, Checkpoint per batch	Continuous, Ack per record	Continuous ,checkpoint per batch	continuous,Ack per batch
Throughput	low	low	high	low	high
Overload	high	high	low	high	low
Delivery guarantee	Exactly once	Exactly once	At least once -	Exactly once,	At least once
Native State operator	Gearpump: persistState	Spark 1.5: updateStateByKey Spark 1.6: mapWithState	Key value state	Flink Java API: ValueState ListState ReduceState Flink Scala API: mapWithState	Heron: User Maintain
Runtime model	Multi tasks on single thread	Multi Tasks of Single application on Single Process Single task on single thread	Multi tasks on single thread	Multi Tasks of Multi Applications on Single Process	Single Task on Single Process
Window Support	Sliding Window	Sliding Window	Sliding ,Count Window	Sliding ,Count and Session Window	Sliding Window
Memory management	JVM	JVM,Self manage on/off heap	JVM	JVM	JVM

STREAMING ANALYTICS USE CASE

- **Continous Risk Monitoring**. It is used to identify risky behavior by real time monitoring of user activity. It can also identify potential users who can put company business in jeopardy .It provides audit reports about suspected users to its clients after calculating user risk to the organization. After identifying potential users, they are examined for their actions and risky user filter is created. Those suspected users list is kept for further inspection. Therefore by application of behavior analytics security team can identify which user can violate security policies and put organization at risk.
- **Fraud detection**: Information age reports that pattern analysis of user average balance, check bounces and user attributes helps bank to identify

potential fraudsters. Every customer has its own credit and debit card, account information .Analytics can identify consumer usage pattern with help of deep analytics against its fraud indicators. Banks fraud detection can issue alert on an unusual account activity such as high transaction until accounts are cross checked with user.

- **Real time wildfire monitoring:** Data about wildfires is collected from numerous sources which also includes real time reporting from satellite imaging and historical data. Analyst collect and analyze data using predictive analytics and models. . Data modelling tool and fire simulators are used to predict future fires. The system gives real time situation updates on fire .Remote weather stations record weather every day .The data collected is used for making correlation to field moisture. When weather analysis shows dry and hot trend then alert is generated.

CONCLUSION

Hadoop is an ecosystem of many open source software projects that changed the world of data analytics .Present day system speed have evolved from traditional to near real time processing .The term real time means zero latency and 24/7 hours availability when required .Analyst analyze real time data that is stored somewhere after some event .In today world machines and sensors are processing data and transmitting data to cloud ,analyzing in real time. Big Data has become more than Hadoop as companies are embracing new streaming technologies such as Spark, Cassandra and Storm. Open source projects such as Spark attract skill developers. There is an imperative need to satisfy growing size of data . Streaming Analytics will continue to evolve to enhance computation, storage and memory processing speed. There is a huge mitigating factor that Hadoop vendors should embrace latest technologies that complement Hadoop.If enterprise want to survive longer they need to look beyond Hadoop.

REFERENCES

1. Ellis, B. (2012). *Real time Analytics-Techniques to visualize data*" (1st edition). [On-line]. Vol. (issue). from: www.wiley.com.
2. Agneeswaran, V. (2012). *Big Analytics beyond Hadoop Engineering* –Real time application with Storm, Spark and Hadoop alternatives (1st edition). [On-line]. from *ptgmedia.pearsoncmg.com/images/9780133837940/samplepages/0133837947.pdf*
3. Anderson, S. (2016). *Beyond ETL Real time streaming architecture*. from https://vision.cloudera.com/beyond-etl-real-time-streaming-architectures
4. *What is streaming data*. from:https://aws.amazon.com/streaming-data/.
5. About Hadoop distributed file system. from https://www-01.ibm.com/software/data/infosphere/hadoop/hdfs/
6. *Hadoop limitations for Big data analytics*. from https://www.em360tech.com/wp-content/files_mf/1360922634PARACCEL1.pdf [2012]
7. Anderson, S. (2016). *Beyond ETL Real time streaming architecture*: from:www.iceengg.edu/staff.html, https://vision.cloudera.com/beyond-etl-real-time-streaming-architectures
8. Spark Streaming. available:http://spark.apache.org/streaming/
9. Amster, A. (2016). *Top Apache Spark Use case*. from https://www.qubole.com/blog/big-data/apache-spark-use-cases.
10. Kafka, A.(2012). *distribued streaming platform*. from: https://kafka.apache.org .
11. Wang, H., Mao, W., & Zhang, T. (2016). Streaming Report Function Comparison and Performance Evaluation.