International Journal of Research in Advent Technology (E-ISSN: 2321-9637) Special Issue
National Conference "NCPCI-2016", 19 March 2016
Available online at www.ijrat.org

Introduction to Real-Time Processing in Apache Apex

Harsh Pathak[1], Manas Rathi[2], Aniket Parekh[3]
Third Year Students[1,2,3], Department of Computer Engineering,
Vishwakarma Institute of Information Technology, Pune, Maharashtra, India.
harshnpathak@gmail.com[1], manas.rathi@outlook.com[2], someshparekh@gmail.com[3]

Abstract- With the advent the 21st century, data across the World Wide Web was generated in huge quantity. This data was impossible to store on physical devices lest process it to obtain results. Data generated across social networks, wireless sensor networks and other big-data sources made it even more difficult for data to be processed instantaneously. Therefore the concept of stream processing emerged which had an advantage over Batch processing. Apache apex facilitated the real-time processing of unbounded stream of data, efficiently increasing the throughput of output stream. Apache Apex is a platform which provides YARN big data in motion that combines stream and batch processing. Big data processing is done in an extremely scalable, secured and fault tolerant way with high performance and simplicity provisions. In this paper, we are providing a case study of Apache Apex, its functionalities, and extendibility relating to real-world use cases. Thus by stating future applications of the platform, we justify its need.

Index Terms- Apache Apex, Big Data, Hadoop,Stream Processing, Windowing, YARN.

1. INTRODUCTION

As we know big data handling along with real time processing is a necessity today. One of the famous big data handling platforms include Hadoop. Hadoop mainly concentrates on operations using big data. It not only allows storage and processing of big data but also does this in a distributed network over a large scale of clustered computers. Being an open source framework it is designed to scale up from a single node to a large number of computers consisting of individual RAM and storage. [5]

Apache Apex includes key features requested by open source developer community that are not available in current open source technologies. (1) Event Processing guarantees
(2) In-memory performance & scalability
(3) Fault tolerance and state management
(4) Native rolling and tumbling window support
(5) Hadoop-native YARN & HDFS implementation

Figure 1 shows the overall architecture of Apache Apex. Apex is a YARN native platform which facilitates real-time stream processing. Rest API could be integrated along with real world applications.

(1) Physical, Virtual, Cloud: Sources for input data set.
(2) Hadoop: Comprises of YARN and HDFS forming the basis of streaming applications.
(3) Streaming Runtime: In-memory processing of data in motion with windowing. (4) Malhar: Library of open source operators.
(5)Streaming Applications: Includes business logic through DAG (Directed Acyclic Graph)
(6)User Interface: For user interactions. Involves Launch, Dashboard and console.

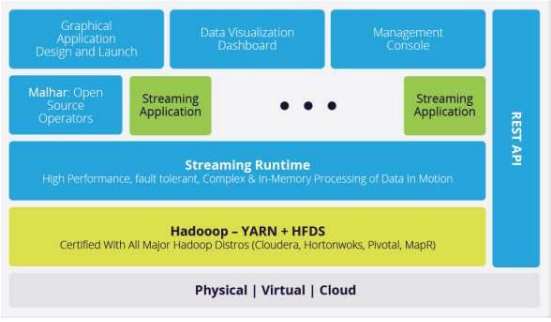

Fig. 1: Architectural framework of Apex. [6]

2. USE CASE

Twitter trends are so viral now a days that the actual need of stream processing is required. The trends for just the city Mumbai is named in normally 1% percent of Indian tweets (approximate). If this holds true for a longer duration then there is a rapid increase to 10% and it becomes a trend. So here is the situation where response throughput must function at very high rate to make the users aware of "what's trending now". Apache Apex fits perfectly in this use-case providing scope for both micro batch stream processing and stream processing.

Another similar use-case is online advertisement, where a particular user's web-page is populated with advertisements in which user's interests to buy/sell are displayed. For example a person searched to buy Puma shoes and within a while when he is checking his Facebook account, the advertisement displaying Puma shoes with best offers is popped up.

Linear Road Benchmark is impelled by the variable tolling system on various observed highways across the globe. This benchmark underlines a variable tolling system for a synchronous and simulated urban expressway system, the tolls for which depend on factors such as traffic congestion and proximity of accidents. Each vehicle comes with a sensor that identifies its location coordinates every 30 seconds.

Internet of Things is a very fresh and trending domain where Apache Apex can play the vital role. For example A single Journey of passenger flight may generate (approx.) 4TB of data through its sensors.

This coordinates generate statistics about traffic conditions on every segment of expressway for every minute.

Also it requires quick measurements where stream processing comes into picture. Some similar projects like Home Automation, Traffic Control, smart dustbin etc. can be few of the use-cases.

3. STREAM PROCESSING

Stream processing is continuous processing on data as it flows through a system. [1] It allows users to act on events instantaneously via alerts. Stream processing mainly focuses on applications having the following attributes:

(1) Compute intensity
(2) Data parallelism
(3) Data Locality

Figure 2 shows various components of Stream and Batch Processing.

Fig. 2: Batch v/s Stream Processing. [6]

4. HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

This Java based file system stores up to 200 PB reliably and streams those datasets at high bandwidth to user applications. Data after getting processed in Apache Apex, gets stored in HDFS. HDFS succeeds
at remaining economical. [3]

Figure 3 shows how Big-Data from HDFS is getting processed in Apache Apex. For example, there is a large file store in HDFS and is processed line by line.

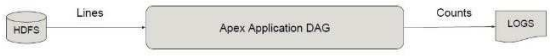

Fig 3: Role of HDFS handling Big Data. [6]

5. APEX ARCHITECTURE

The figure 4 represents the ease of integration with Source, Apex platform and Destination. Source can be a message bus like KAFKA, file system like HDFS or a database like MySQL.

(1) These sources are given as input to the Apex framework.

(2) DAG (Directed Acyclic Graph) takes input from the source and the operator processes it. (3) Output operator is forwarded to the Destination.

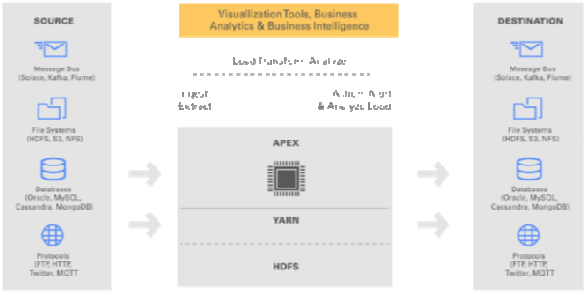

Fig. 4: Architecture showing ease of integration. [6]

Apache Apex (incubating) is a powerful, reliable, versatile and flexible stream processing hybrid platform for Big-Data. At the core of Apache Apex is an (SPE) Stream Processing Engine that enables development of highly performant, fully scalable and completely fault tolerant enterprise grade applications on Apache Apex. [6]

Figure 5 shows how Apex can be easily reused while leveraging investments in Hadoop. It consists of two libraries (1)Apex Malhar (2)Apex Core.

Apex Malhar is a library of operator set containing predefined instances of popular file systems, messages buses and database. This enables various organizations to gain experience over Apex platform and focus on business logic.

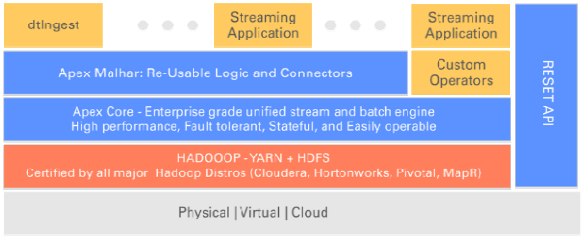

Fig. 5: Reusability while leveraging investments in Hadoop. [6]

5.1 FEATURES

(1) Code Reuse:

Apex enables defining reusable modules. It enables the same business logic to be used for stream as well as batch. Apex is a data-in-motion platform that allows for a unification of processing of never-ending streams of unbounded data (streaming job), or bounded data in
files (batch job)

(2) Stream processing v/s Batch processing big data:

Because big data is tough to envisage in its entirety, the platform must be such that it posits to become the basis for driving big data processing needs, in a batch paradigm, streaming paradigm, or both. Apache Apex is industry's only open-source enterprise-grade engine capable of handling batch data as well as steaming data needs. Apache Apex is groomed to drive the highest value for businesses operating in highly data-intensive environments.

(3) Integration and Use:

The Apex platform comes with support for web services and metrics. This enables ease of use and easy integration with current data pipeline components. DevOps teams can monitor data in action using existing systems and dashboards with minimal changes, thereby easily integrating with the current setup.

(4) Simplicity and Expertise:

Simplicity is a key component of every software. An application is a directed acyclic graph (DAG) of multiple operators. The operator developer only needs to implement the process() call. Plug and play functionality serves processing of incoming events. The single-thread approach and application-level JAVA expertise are the top reasons why Apex enables big data teams to develop applications within weeks.

6. WIRELESS SENSOR NETWORKS

Wireless sensor networks (WSN) are spatially distributed, autonomous sensors that monitor physical or environmental conditions, such as

vibration, fire, light, brightness, temperature, sound, pressure, etc. WSN cooperatively as well as synchronously transfer their stored information through the wireless network to a remote server. [4]

Real-time analytics is the capacity to use, all available task data and resources when the instant processing is needed.

It consists of dynamic analysis and reporting, based on data entered into a system and requires response time less than a minute.

Fig. 6: Data flow of WSN [6]

7. SOCIAL NETWORKING

Social Networking, in today's world, is the prime source of getting connected to other individuals over the internet. According to statistics it is estimated that Facebook alone housing over 1.5 billion users generate data equivalent to 4,166,667 posts every minute. To handle such huge data let aside process it is a colossal task.

Thus Apache Apex can be utilized to handle such traffic that to with real-time conditions.

Figure 9 shows the rate at which data is generated across all famous social networking websites and applications.

Fig. 7: An overview of Data generated Statistics. [8]

8. WINDOWING

Windowing is a time relative concept. Windowing involves intake of small datasets from a large dataset for processing and further analysis. [2]In social networks and well as Wireless sensor networks, the input data set is huge thus the concept of windowing simplifies the process of handling data.

Figure 8 shows how windowing is applied on a realtime data set which is incoming at a high rate. The default time frame is 500 milliseconds.

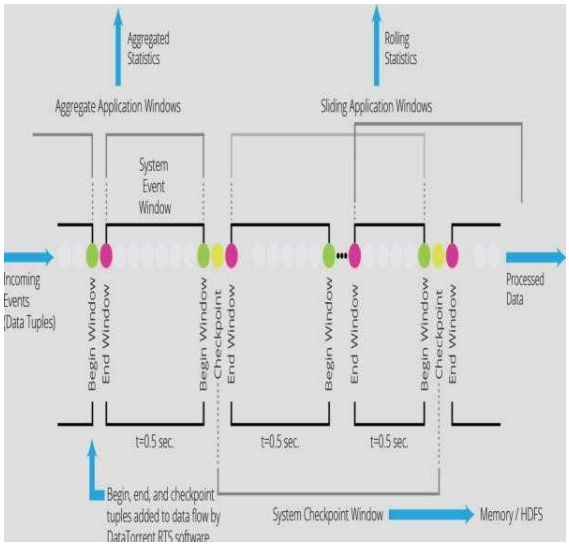

Fig. 8: Windowing in Apache Apex [6]

9. CONCLUSION

In this paper we have studied the basics of Apache Apex, Stream Processing, Hadoop Distributed File System and YARN. This paper also gives readers and idea of the extensive uses of Apache apex during realtime data handling. With the concept of windowing, data handling in smaller frames and transfer from one node to another in HDFS is understood. The paper provides an overlook on real-time data generation, processing and implementation particularly in the field of social networking and wireless sensor networks.

Along with the use cases provided, the paper provides ample future applications for Big Data handling using the concepts of Stream Processing, windowing and DAG with Apache Apex platform.

REFERENCES

[1] Fatos Xhafa, Victor Naranjo, Santicabanle, Leonard Barolli, "A software chain approach to big data stream processing and analytics", CISIS, 2015

[2] Teja MVSR, Ajith Kumar, C. V. Sai Prasanth, "Comparative Analysis of Windowing Techniques in minimizing side lobes in an antenna array", ICCSP, 2014

[3] A. Kala Karun, K. Chitharanjan, "A review on Hadoop – HDFS infrastructure extensions", ICT, 2013

[4] Yong-Sik Choi, "A study on sensor nodes attestation protocol in a Wireless Sensor Network", ICACT, Volume.1, 2010

[5] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler, "The Hadoop Distributed File System", MSST, 2010

[6] DataTorrent 2013 [Online] www.datatorrent.com

[7] The Apache Source Foundation 2016 [Version 2.0] www.apex.incubator.apache.org

[8] DOMO: Business Intelligence, Dashboards, Reporting and Analytics. www.domo.com