Using apache storm for big data

S Surshanov*

IITU, Kazakhstan

**Corresponding author's e-mail: sanzharsurshanov@gmail.com*

Received 10 January 2015, www.cmnt.lv

Abstract

The rapid growth in information technologies has resulted in creating of new concepts, opportunities and challenges. Big Data (BD) is one of them with its tools and techniques. Big data is becoming a metonym for competitive advantage in business. However, majority is not yet familiar on what are or what is meant by "Big Data". In spite its huge benefits, big data has serious challenges and requires some critical thoughts when it comes to analyzing of big data. This study focuses on exploring the meaning of BD, identifying important parts of research on BD and its analysis concepts using sample from Twitter information using the real-time processing software called Apache Storm.

Keywords: Big Data, distributed computing, apache storm, real-time processing.

1 Introduction

Big data has excelled to be one of the most significant technologies that have influenced the way enterprises use information to progress their business models and experiences. Basically it is a combination of data technologies management that has come through some time. Big data enables business to use, store and manage lots of data at the precise time to achieve accuracy and efficiency. The most important feature of BD is that gathered data must be managed in a way so business requirements are kept. There are different ways to collect vast amounts of information to find the hidden patterns, which exists inside that data so it can bring some insights to the business model [1]. Working with big data results require that the base to be set up to back the administration, circulation and flexibility of that data. Therefore, it is essential to put both business and specialized procedure arrangements to utilize this important engineering pattern.

In stream, processing the model of speed is determinant. As a consequence it requires a processing tool, that could operate the new gathered information at a faster pace with guaranteed processing and low latency. Afterwards new instruments are needed to prepare this huge speed of information. "Apache Storm" is the leading real-time processing tool, which delivers the processing of newly gathered information with low latency. Currently the mostly used tool is Hadoop, it is relatively works well, however because Hadoop operates the data in batch it is not most suitable tool for analyzing the cutting edge types of data. Processing the data in real-time is nowadays a usual requirement. This technique is called stream processing; basically it is analysis of real-time data in continues motion. This paper will focus on studying the Big Data analysis and its tools, in particular Apache Storm and will show the comparison of other various tools with Apache Storm.

2 Overview of Big Data and Apache Storm

2.1 ANALYZING OF BIG DATA

"Big Data" has become a catchy term that, for the moment, retains some mystique and persuasive impact in use. The term Big Data is used by different organizations but there is no standard definition of it. On the base of BD is located information; however it's not only significant part of it. Three different attributes such as volume, variety and velocity are combined and linked together to establish Big Data. Real-time information comes from different sources and in various types; for instance it might be content from blog posts, pictures, geolocation, logs and etc. Traditional databases work on homogeneous data structures and cannot process heterogeneous data, therefore are not very compatible to work with BD [2]. The term "real-time" itself can represent two perspectives depending on the context and it is important to be able to recognize them. If it is used related to information, it means transforming the latest available information, operating numbers of data as it is generated. On the other side, if real-time framework is used to detect drifts in Twitter stream, the notion of real-time can be deferred by a couple of seconds. But usually regarding to real-time process's, it means processing the data with very low latency [3]. Stream processing is designed to analyze and act on data, which is generated in real-time, i.e. using "continuous queries" that operate repeatedly over time. This type of processing enables us to analyze the stream i.e. to extract mathematical or statistical information analytics on the runtime within the stream. Stream processing solutions are designed to handle Big Data in real time with a highly scalable, highly available and highly fault tolerant architecture. This enables us to analyze the data in motion [4].

2.2 AVAILABLE TOOLS

Below are listed some open source tools, which are being used for big data analysis:

1. Apache HBase

Apache HBase is a Java based, open-source software,

which enables to store Big Data. It is highly non-relational in nature and provides Google's Bigtable like functionality to store sparse data. HBase is widely used when random and real-time access to Big Data is required and is operates on the top of HDFS [5].

2. Hadoop

The Apache Hadoop project is open source software to process Big Data. The key features of Apache Hadoop are its reliability, scalability and its processing model. It allows processing the large sets of data across clusters of machines using distributed programming paradigm. It operates the information in small batches and uses MapReduce framework to process the data and is called batch processing software [6].

3. Apache Spark

Apache Spark project is open source based for processing fast and large-scale data, which relies on cluster computing system. Like Apache Hadoop it is also designed to operate on batches, but the batch window size is very small. It provides flexibility to develop modules in three different languages Java, Scala and Python. It also provides a rich set of tools that are to process SQL including Spark SQL, for machine learning MLlib, for process graph GraphX, and for stream analysis Spark Streaming [7].

4. Yahoo S4

In October 2010, Yahoo released Yahoo S4. In 2011 it joined Apache Foundation Family and it was given the status of Apache Incubator. Yahoo S4 empowers developer to design applications, which can process real-time streams of data. It is inspired by MapReduce model and process the data in distributed fashion. It supports modular programming model i.e. developer can develop plug and play modules in Java. The modules developed in Yahoo S4 can be consolidate to design more advance real-time processing applications [8].

5. Apache Storm

In December 2010, Nathan Marz came up with an idea to develop a stream processing system that can be presented as a single program. This idea resulted to a new project called Storm. Apache Storm empowers developers to build real-time distributed processing systems, which can process the unbounded streams of data very fast. It is also called Hadoop for real-time data. Apache Storm is highly scalable, easy to use, and offers low latency with guaranteed data processing. It provides a very simple architecture to build applications called Topologies. It enables developer to develop their logic virtually in any programming language, which supports communication over a JSON-based protocol over stdin/stdout. Apache Storm becomes the part of Apache Family on 17 September 2014.

2.2.1 Architecture

There are three sets of nodes in a Storm cluster and they are Nimbus node, ZooKeeper nodes and Supervisor nodes. Nimbus is the main server where user code has to be uploaded and Nimbus distributes this code among the worker nodes for execution. Also Nimbus keeps track of the progress of the worker nodes so that it can restart the failed computation or move the tasks to other nodes in case of node failures. The set of worker nodes in the Storm cluster runs a daemon called Supervisor. The coordination between supervisor nodes and the Nimbus happens through the ZooKeeper. The message flow in the system is done using ZeroMQ based transport or Netty based transport. The transport layer is pluggable.

2.2.2 Programming Model

Storm does not try to fit a specific programming model like MapReduce on top of streams. Storm programming model provides distributed stream partition among the processing nodes. Each processing element process the input as it processes the whole stream. Storm programming model consists of Spouts, Bolts, Topologies and Streams. The Spouts and Bolts are arranged in a DAG called a Topology. A user submits a topology to a Storm cluster to execute. Stream is a set of tuples and these tuples can be a user-defined type or a system defined type. Spouts are the stream sources for the topology. Bolts consume events and emit events after processing. Storm topology starts with a set of spouts and the rest of the layers of the topology consist of Bolts. User can write the storm spouts and bolts in different programming languages like python, java or clojure. A storm job is configured using the Java programming language as a topology object and the storm client is used to submit the job to the Nimbus.

2.3 COMPARISON OF APACHE STROM WITH OTHER TOOLS

Below Table 1 will compare big data open source tools with Apache Strom [9]:

TABLE 1 Comparison big data open source tools

Other tools	Developer	Type	Difference
HBase	Apache	Batch	Storm provides real time data processing, while HBase (over HDFS) does not process rather offers low-latency reads of processed data for querying later.
Hadoop	Apache	Batch	The main difference is that Storm can do real-time processing of streams of Tuple's (incoming data) while Hadoop do batch processing with MapReduce jobs.
Spark	UC Berkeley AMPLab	Batch	One way to describe the difference is that Apache Spark is a batch processing framework that is capable of doing micro-batching also called Spark Streaming, while Apache Storm is real-time stream processing frameworks that also perform micro-batching also called Storm-Trident. So architecturally they are very different, but have some similarity on the functional side. With micro-batching, one can achieve higher throughput at the cost of increased latency. With Spark, this is unavoidable and with Storm, one can use the core API (spouts and bolts) to do one-at-a- time processing to avoid the inherent latency overhead imposed by micro-batching. And finally, many enterprises use Storm as a mature tool while Spark Streaming is still new.
S4	Yahoo!	Streaming	The main difference is that, storm gives guaranteed processing with high performance and thread programming support.

2.3 WHY APACHE STORM?

Five key attributes, which make Apache Storm as a first choice tool for processing real-time:

- Easy to use fetch the data in real time
- Fast – benchmarked for processing millions byte data per second per node
- Fault-tolerant – keep the track of all worker nodes, whenever a node dies, Apache Storm restart the process on another node
- Reliability –Guaranteed data processing with at least once semantics
- Scalability – process the data in parallel across a cluster of machines.

Below listed are the main criterion, on basis of which one can decide when to use Apache Storm [11].

- Fault tolerance: High fault tolerance
- Latency: Sub Seconds
- Processing Model: Real-time stream processing model
- Programming language dependency: any programming language
- Reliable: each tuple of data should be processed at least once
- Scalability: high scalability

3 Materials and methods

Big data is cutting edge technology that have changed the way world have looked at data and all of methods and principles towards data. Technically speaking, this paper will be using Twitter streaming API to get access to twitter big data as a big data sample. Storm makes it easy to reliably process unbounded streams of data, doing for real-time processing what Hadoop did for batch processing. This experiment will execute 1 scenario with live data and will collect the statistics, which will be used to analyze the processing tool and to draw some conclusion.

4 Results and conclusions

This section illustrates and analysis the data collected for the experiment purpose using twitter streaming API. The study was aimed at analyzing the twitter big data streams using state of art Apache Storm open source tools to recognize particular patterns from huge amount of data. Following scenarios were executed for experiment purpose on live streams of tweets on twitter:

- Top ten words collected during a particular period of time.
- Number of times a particular "word" was being used in tweets, tweeted in a particular period of time.

Scenario-1: Top ten words collected in last 10 minutes

Statistics:

- Total time duration=10 minutes (603 seconds).
- The total number of tweets analysed during this time=67271
- The total number of words=482874
- See Table II for top ten words in tabular form.

TABLE 2 Top ten words in last 10 minutes

No.	Word	Frequency
1	Jessie	15585
2	Lady	18543
3	Gaga	18552
4	Rey	23664
5	Lana	23677
6	Del	23690
7	Swift	23881
8	Taylor	24284
9	Coldplay	25330
10	Mtvstars	62726

This study explored for companies to understand the Big Data and its notions. It reviewed for the companies to choose between traditional databases and the big data tools. It is also empowering the developer to understand the use of Storm to analyze and process big data. This study was conducted under some experimental limitations in terms of infrastructure and data. In terms of data approximately 1% of the total tweets were available with Twitter free API. In terms of hardware configuration the experiment was not performed on dedicated Server, rather this study was conducted using laptop ASUS K53SV having corei7 - 2670QM CPU @ 2.20GHz × 8, 4 GB RAM, graphics Intel® Sandybridge Mobile, OS type 64-bit and Linux Ubuntu 14.4. The following scenario was performed:

- Top ten words twitted during last 10 minutes.

The above three mentioned scenario was performed successfully, proving Apache Storm can process real-time streams with very low latency. All the tweets were queued as they were received without any delay and calculations were performed on the tweets using bolts. The programming model was easy to build on own topologies. The execution of the topology can be drawn as a directed graph. Even though Apache is built on Clojure, the topologies were created in Java, so programming can be done in multiple languages. During this experiment, some areas of future development were identified. To install and configure Apache Storm is not easy task. No direct setup is available to install pre-requisites and configure the tool. All the steps have to be performed manually and there is no comprehensive guide available. So for future releases a user friendly installer and configuration module will be of great use for developers. Although Apache Spark provides some key performance indicator's (KPI's) to measure the performance and reliability but it is not enough to call it user friendly. There is no reporting module either. For future releases addition of a reporting module will make the tool the leading open source tool for real-time processing.

References

[1] J. M. a. M. C. Tim McGuire, August 2012 [Online] http://iveybusinessjournal.com/topics/strategy/why-big-data-is-the-new-competitive-advantage#.VKv0wSuUe9E.
[2] "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute 2011
[3] Perroud B 2013 A hybrid approach to enabling real-time queries to end-users *Software Developer's Journal*
[4] Wähner K 10 Sep 2014. [Online] http://www.infoq.com/articles/stream-processing-hadoop.
[5] "Apache HBase," Apache, 22 December 2014. [Online] http://hbase.apache.org/. [Accessed 06 January 2015].
[6] 1 Dec 2014. [Online]. Available: http://hadoop.apache.org/

[7] 4 Dec 2014. [Online]. Available: https://spark.apache.org/.
[8] 4 Dec 2014. [Online]. Available: http://incubator.apache.org/s4/.
[9] Damji J S "Discover HDP 2.1: Apache Storm for Stream Data Processing in Hadoop," 23 June 2014. [Online] http://hortonworks.com/blog/discover-hdp-2-1-apache-storm-stream-data-processing-hadoop/ [Accessed 06 January 2015]
[10] "Apache Storm," Hortonworks, [Online] http://hortonworks.com/hadoop/storm/. [Accessed 06 January 2015].
[11] "Apache Storm," Apache Software Foundation, 2014. [Online] https://storm.apache.org/about/integrates.html. [Accessed 01January 2015]

Author

Sanzhar Surshanov, 1991, Almaty, Kazakhstan

University studies: currently studying at International Information Technology University, Almaty, 2015.
Scientific interest: grid computing, real-time processing.
Publications: Thesis, "Developing a system using Cloud Computing", The 12th International Scientific Conference, Information Technologies and Management Institute (ISMA_IT&M2014), Riga, Latvia.