Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Big Data Analysis: Ap Spark Perspective

Abdul Ghaffar Shoro[1] and Tariq Rahim Soomro[2]

[1] SZABIST Dubai Campus

Received: 7 December 2014 Accepted: 5 January 2015 Published: 15 January 2015

Abstract

Big Data have gained enormous attention in recent years. Analyzing big data is very common requirement today and such requirements become nightmare when analyzing of bulk data source such as twitter twits are done, it is really a big challenge to analyze the bulk amount of twits to get relevance and different patterns of information on timely manner. This paper will explore the concept of Big Data Analysis and recognize some meaningful information from some sample big data source, such as Twitter twits, using one of industries emerging tool, known as Spark by Apache.

Index terms— big data analysis, twitter, apache spark, apache hadoop, open source.

Introduction n today's computer age, our life has become pretty much dependent on technological gadgets and more or less all aspects of human life, such as personal, social and professional are fully covered with technology. More or less all the above aspects are dealing with some sort of data; due to immense increase in complexity of data due to rapid growth required speed and variety have originated new challenges in the life of data management. This is where Big Data term has given a birth. Accessing, Analyzing, Securing and Storing big data are one of most spoken terms in today's technological world. Big Data analysis is a process of gathering data from different resources and then organizing that data in meaning full way and then analyzing those big sets of data to discover meaningful facts and figures from that data collection. This analysis of data not only helps to determine the hidden facts and figures of information in bulk of big data, but also it provides with categorize the data or rank the data with respect to important of information it provides. In short big data analysis is the process of finding knowledge from bulk variety of data. Twitter as organization itself processes approximately 10k tweets per second before publishing them for public, they analyze all this data with this extreme fast rate, to ensure every tweet is following decency policy and restricted words are filtered out from tweets. All this analyzing process must be done in real time to avoid delays in publishing twits live for public; for example business like Forex Trading analyze social data to predict future public trends. To analyze such huge data it is required to use some kind of analysis tool. This paper focuses on open source tool Apache Spark. Spark is a cluster computing system from Apache with incubator status; this tool is specialized at making data analysis faster, it Author ? ?: Department of Computing, SZABIST Dubai Campus, Dubai, UAE. e-mails: shoroghaffar@gmail.com, tariq@szabist.ac.ae is pretty fast at both running programs as well as writing data. Spark supports in-memory computing, that enables it to query data much faster compared to diskbased engines such as Hadoop, and also it offers a general execution model that can optimize arbitrary operator graph [1]. This paper organized as follows: section 2 focus on literature review exploring the Big Data Analysis & its tools and recognize some meaningful information from some sample big data source, such as Twitter feeds, using one of industries emerging tool, Apache Spark along with justification of using Spark; section 3 will discuss material and method; section 4 will discuss the results of analyzing of big data using Spark; and finally discussion and future work will be highlighted in section 5.

1 II.

2 Literature Review a) Big Data

A very popular description for the exponential growth and availability of huge amount of data with all possible variety is popularly termed as Big Data. This is one of the most spoke about terms in today's automated world

and perhaps big data is becoming of equal importance to business and society as the Internet has been. It is widely believed and proved that more data leads to more accurate analysis, and of course more accurate analysis could lead to more legitimate, timely and confident decision making, as a result, better judgment and decisions more likely means higher operational efficiencies, reduced risk and cost reductions [2]. Big Data researchers visualize big data as follows:

i. Volume-wise This is the one of the most important factors, contributed to emergence of big data. Data volume is multiplying to various factors. Organizations and governments has been recording transactional data for decades, social media continuously pumping steams of unstructured data, automation, sensors data, machineto-machine data and so much more. Formerly, data storage was itself an issue, but thanks to advance and affordable storage devices, today, storage itself is not a big challenge but volume still contributes to other challenges, such as, determining the relevance within massive data volumes as well as collecting valuable information from data using analysis [3].

ii. Velocity-wise Volume of data is challenge but the pace at which it is increasing is a serious challenge to be dealt with time and efficiency. The Internet streaming, RFID Big Data Analysis: Ap Spark Perspective tags, automation and sensors, robotics and much more technology facilities, are actually driving the need to deal with huge pieces of data in real time. So velocity of data increase is one of big data challenge with standing in front of every big organization today [4].

iii. Variety-wise Rapidly growing huge volume of data is a big challenge but the variety of data is bigger challenge. Data is growing in variety of formats, structured, unstructured, relational and non-relational, different files systems, videos, images, multimedia, financial data, aviation data and scientific data etc. Now the challenge is to find means to correlate all variety of data timely to get value from this data. Today huge numbers of organizations are striving to get better solutions to this challenge [3].

3 iv. Variability-wise

Rapidly growing data with increasing variety is what makes big data challenging but ups and downs in this trend of big data flow is also a big challenge, social media response to global events drives huge volumes of data and it is required to be analyzed on time before trend changes. Global events impact on financial markets, this overhead increase more while dealing with un-structured data [5].

4 v. Complexity-wise

All above factors make big data a really challenge, huge volumes, continuously multiplying with increasing variety of sources, and with unpredicted trends. Despite all those facts, big data much be processed to connect and correlate and create meaningful relational hierarchies and linkages right on time before this data go out of control. This pretty much explains the complexity involved in big data today [5].

To precise, any big data repository with following characteristics can be termed big data. The following are brief introduction of some of selected big data analysis tools along with brief overview of Apache Spark and finally justification of apache spark with other competitors to distinguish and justify use of Apache Spark.

5 i. Apache Hive

Hive is a data warehousing infrastructure, which runs on top of Hadoop. It provides a language called Hive QL to organize, aggregate and run queries on the data. Hive QL is similar to SQL, using a declarative programming model [7]. This differentiates the language from Pig Latin, which uses a more procedural approach. In Hive QL as in SQL the desired final results are described in one big query. In contrast, using Pig Latin, the query is built up step by step as a sequence of assignment operations.

Apache Hive enables developers specially SQL developers to write queries in Hive Query Language HQL. HQL is similar to standard query language. HQL queries can be broken down by Hive to communicate to MapReduce jobs executed across a Hadoop Cluster.

ii. Apache Pig Pig is a tool or in fact a platform to analyze huge volumes of big data. Substantial parallelization of tasks is a very key feature of Pig programs, which enables them to handle massive data sets [7]. While Pig and Hive are meant to perform similar tasks [8]. The Pig is better suited for the data preparation phase of data processing, while Hive fits the data warehousing and presentation scenario better. The idea is that as data is incrementally collected, it is first cleaned up using the tools provided by Pig and then stored. From that point on Hive is used to run ad-hoc queries analyzing the data. During this work the incremental buildup of a data warehouse is not enabled and both data preparation and querying are performed using Pig. The feasibility of using Pig and Hive in conjunction remains to be tested.

iii. Apache Zebra Apache Zebra is a kind of storage layer for data access at high level abstraction and especially tabular view for data available in Hadoop and relief's users of pig coming up with their own data storage models and retrieval codes. Zebra is a sub-project of Pig which provides a layer of abstraction between Pig Latin and the Hadoop Distributed File System [9]. Zebra allows a Pig programmer to save relations in a table-oriented fashion (as opposed to flat text files, which are, normally used) along with meta-data describing the schema of each relation. The tests can be run using J Unit or a similar Java testing framework [10].

6 iv. Apache H Base

Apache H Base is a data base engine built using Hadoop and modeled after Google's Big Table. It is optimized for real time data access from tables of millions of columns and billions of rows. Among other features, H Base offers support for interfacing with Pig and Hive. The Pig API features a storage function for loading data from an H Base data base, but during this work the data was read from and written to flat HDFS files, because the data amounts were too small to necessitate the use of H Base [11]. [11]. Because Chu kwa is meant mostly for the narrow area of log data processing, not general data analysis, the tools it offers are not as diverse as Pig's and not as well suited for the tasks performed in this work.

7 vi. Apache Storm

A dependable tool to process unbound streams of data or information. Storm is an ongoing distributed system for computation and it is an open source tool, currently undergoing incubation assessment with Apache. Storm performs the computation on live streams of data in same way traditional Hadoop does for batch processing. Storm was originally aimed at processing twitter streams, and now available as open source and being utilized in many organizations as stream processing tool. Apache spark is quick and reliable, scalable, and makes sure to transform information. It is also not very complex to be deployed and utilized [1].

8 vii. Apache Spark

Apache Spark is a general purpose cluster computing engine which is very fast and reliable. This system provides Application programing interfaces in various programing languages such as Java, Python, Scala. Spark is a cluster computing system from Apache with incubator status, this tool is specialized at making data analysis faster, it is pretty fast at both running programs as well as writing data. Spark supports in-memory computing, that enables it to query data much faster compared to disk-based engines such as Hadoop, and also it offers a general execution model that can optimize arbitrary operator graph. Initially system was developed at UC Berkeley's as research project and very quickly acquired incubator status in Apache in June 2013 [9]. Generally speaking, Spark is advance and highly capable upgrade to Hadoop aimed at enhancing Hadoop ability of cutting edge analysis. Spark engine functions quite advance and different than Hadoop. Spark engine is developed for in-memory processing as well a disk based processing. This inmemory processing capability makes it much faster than any traditional data processing engine. For example project sensors report, logistic regression runtimes in Spark 100 x faster than Hadoop Map Reduce. This system also provides large number of impressive high level tools such as machine learning tool M Lib, structured data processing, Spark SQL, graph processing took Graph X, stream processing engine called Spark Streaming, and Shark for fast interactive question device. As shown in

9 d) When Not to Use Apache Spark

Apache Spark is fasted General purpose big data analytics engine and it is very suitable for any kind of big data analysis. Only following two scenarios, can hinder the suitability of Apache spark [13].

10 Global Journal of Computer Science and Technology (C) Volume XV Issue I Version I

Year 2015

? Low Tolerance to Latency requirements: If big data analysis are required to be performed on data streams and latency is the most crucial point rather anything else. In this case using Apache Storm may produce better results, but again reliability to be kept in mind.

? Shortage of Memory resources: Apache Spark is fasted general purpose engine due to the fact that it maintains all its current operations inside Memory. Hence requires access amount of memory, so in this case when available memory is very limited, Apache Hadoop Map Reduce may help better, considering huge performance gap.

III.

11 Material and Methods

The nature of this paper is to cope with huge amount of data and process / analyze huge volume of data to extract some meaningful information from that data in real time.

The big data is modern day technology term that have changed the way world have looked at data and all of methods and principles towards data. The Data gather of big data is totally different than our traditional ways of data gathering and techniques. Coping with big data specially analyzing in real time has become almost impossible with traditional data warehousing techniques. This limitation have resulted a race of new innovations in data handling and analyzing field. Number of new technologies and tools have emerged and claiming to resolve big data analyzing challenges. So technically speaking, Twitter streaming API is used to access twitter's big data using Apache Spark.

12 a) Research Instrument

? Twitter Stream API: The Streaming APIs provide push deliveries of Tweets and other events, for realtime or low-latency applications. Twitter API is well known source of big data and used worldwide in numerous applications of a number of objectives.

In fact there are some limitation in free Twitter API that should be considered while analyze the results.

13 ? Apache Spark: As an open source computing

framework to analyze the big data. Though apache spark is claiming to be fastest big data analyzing tool in market, but the trust level and validation of results will still be subject to comparison with some existing tools like Apache storm, for example. In this paper the data processing is happening using Twitter streaming API and Apache Spark as shown in Figure-3-1 bellow.

14 Results

This section illustrates and analysis the data collected for the experiment purpose by Apache Spark using twitter streaming API. The amount of data processed for each scenario, processing time and results are given in tabular as well as graphical format. Following scenarios were executed for experiment purpose on live streams of twits on twitter. 1. Top ten words collected during a particular period of time. (10 minutes) 2. Top ten languages collected during a particular period of time. (10 minutes) 3. Number of times a particular "word" being used in twits, twitted in a particular period of time.

Scenario 1: Top ten words collected in last 10 minutes

15 Statistics:

? The total number of tweets analyzed during this time=23865 ? The total number of unique words =77548 ? The total number of words=160989 ? Total time duration=10 minutes (600 seconds).

? See Table 4-1 for top ten words in tabular form.

? See Figure 4 ? See Table 4-3 for number of twits posted using word "mtvstars" in tabular form ? See Figure 4-3 for number of twits posted using word "mtvstars" shown graphically in charts Discussion & Future Work

As not many organizations share their big data sources. So study was limited to twitter free feed API and all limitations of this API, such as amount of data per request and performance etc. and that directly impact the results presented. Also a common laptop was used to analyze tweets as compare to dedicated Server. As a result of this study, following Scenarios were considered and analyzed and their results were presented in previous section. 1. Top ten words twitted during last specific period of time. 2. Top ten languages used to twit during specific period of time. 3. A list of twitted items matching a given search keyword.

Considering the above mentioned limitations, Apache Spark was able to analyze streamed tweets with very minor latency of few seconds. Which proves that, despite being big general purpose, Interactive and flexible big data processing engine, Spark is very competitive in terms of stream processing as well. During the process of analyzing big data using spark, couple of improvement areas were identified as of utmost importance should be persuaded as future work. Firstly, like most open source tools, Apache Spark is not the easiest tool to work with. Especially deploying and configuring apache spark for custom requirements. A flexible, user friendly configuration and programming utility for apache spark will be a great addition to apache spark developer community. Secondly, analyzed data representation is poor, there is a very strong need to have powerful data representation tool to provide powerful reporting and KPI generation directly from Spark results, and having this utility in multiple languages will be a great added value. [1]

Figure 1: I

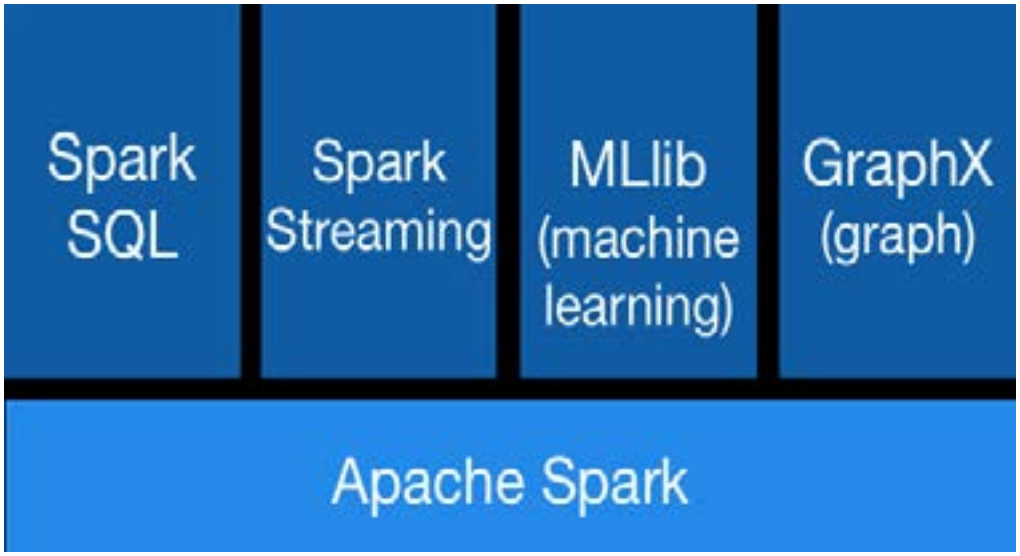


Figure 2:

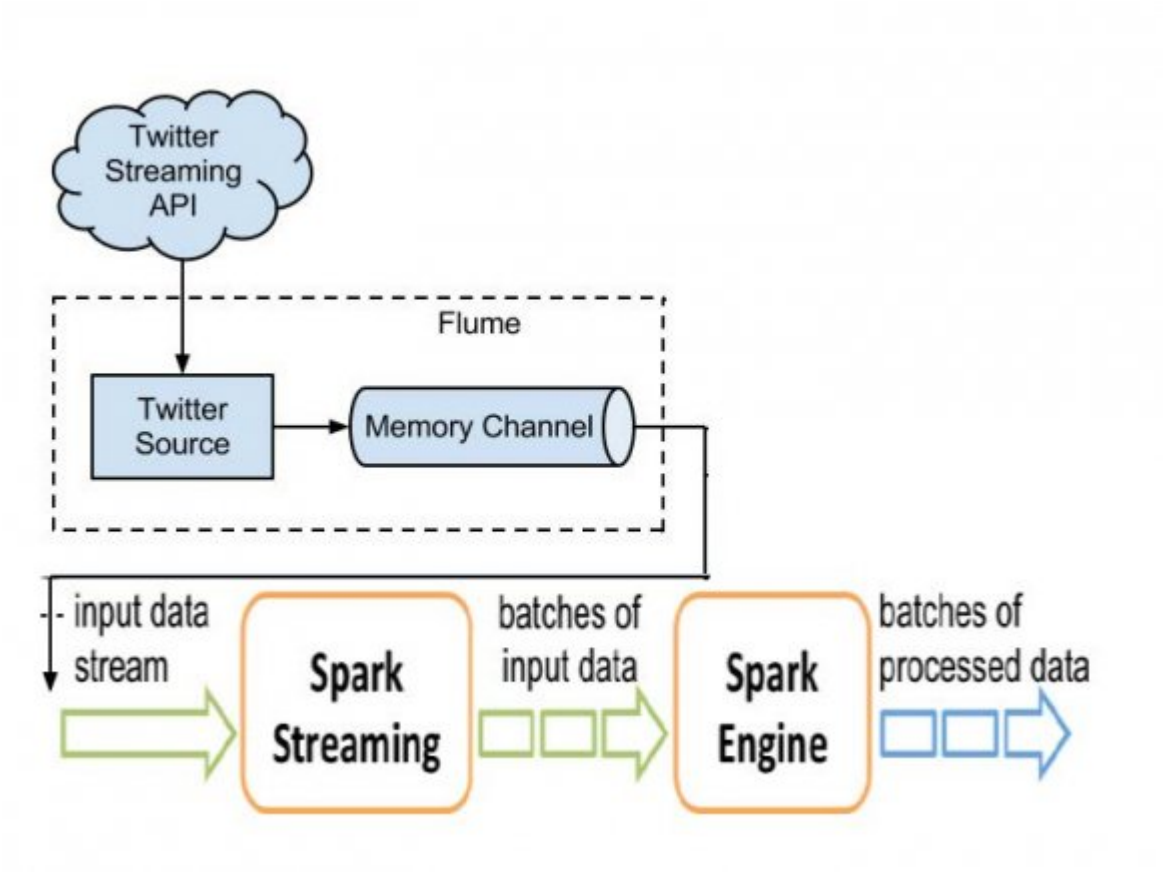


Figure 3: Global

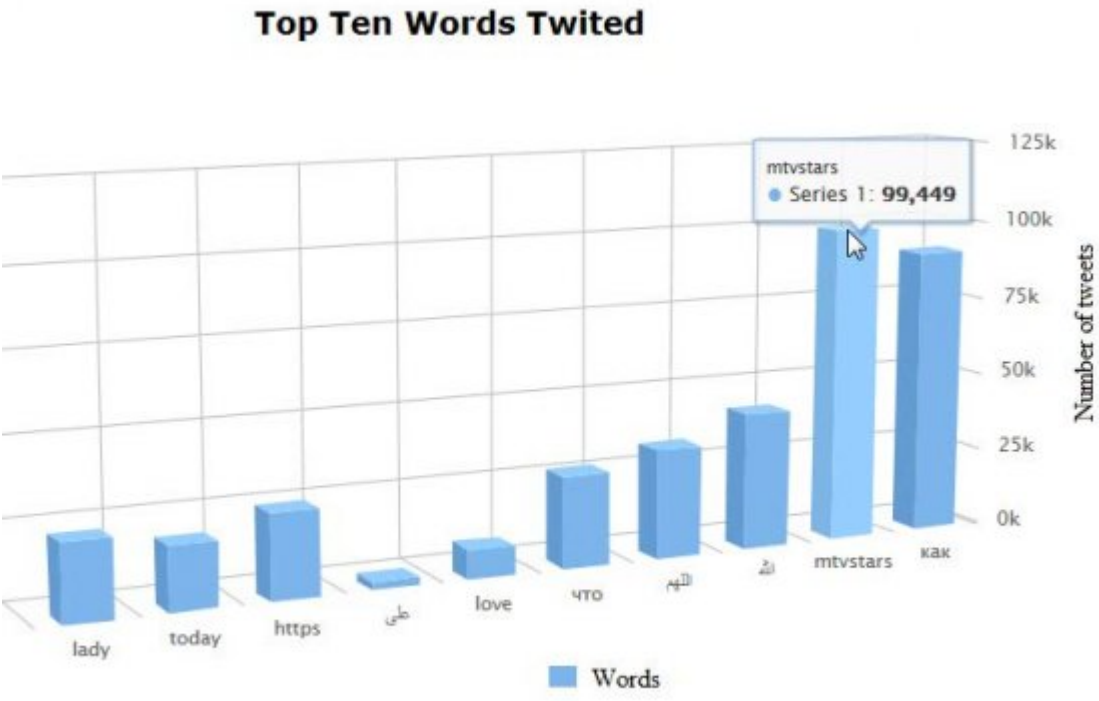
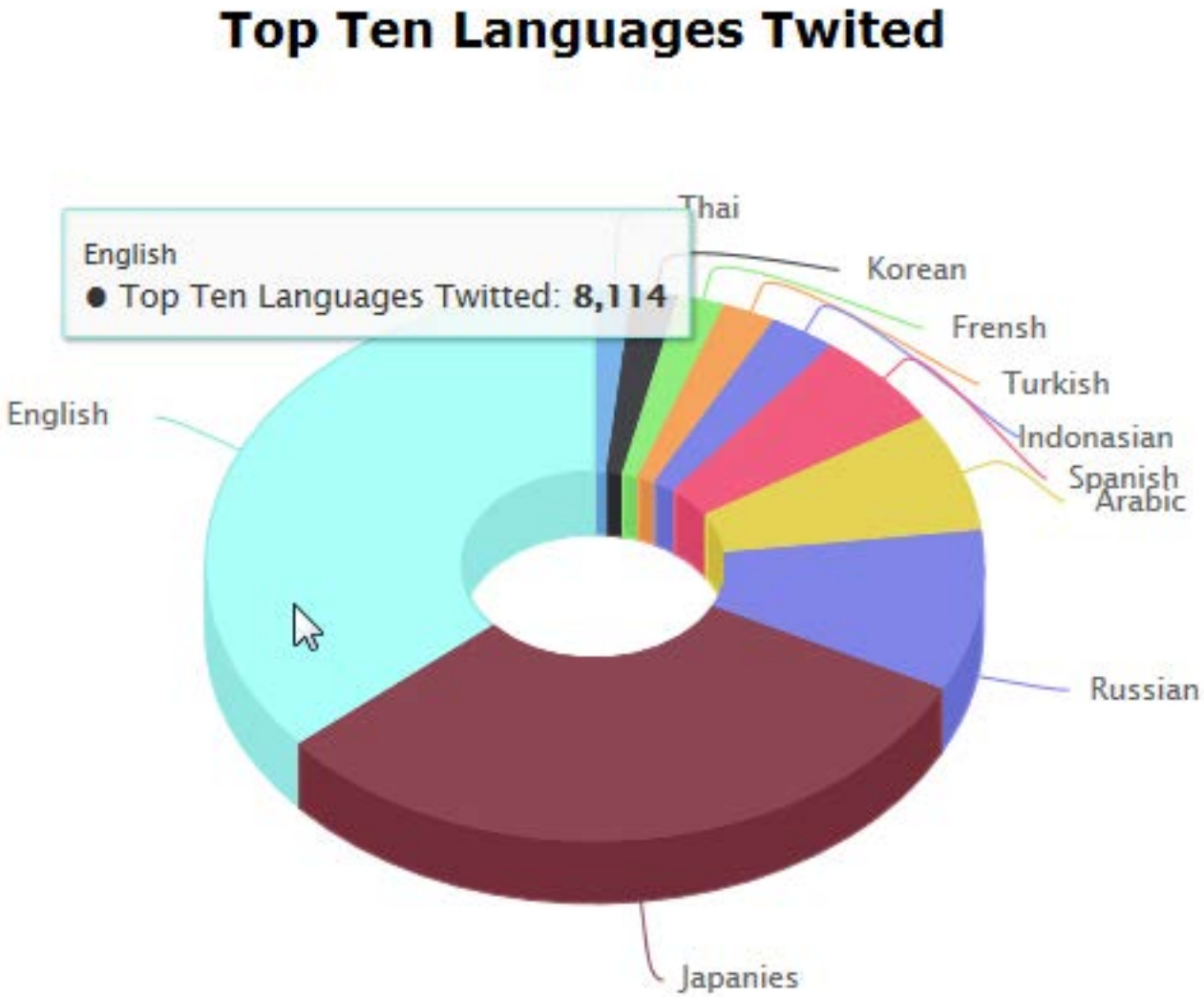
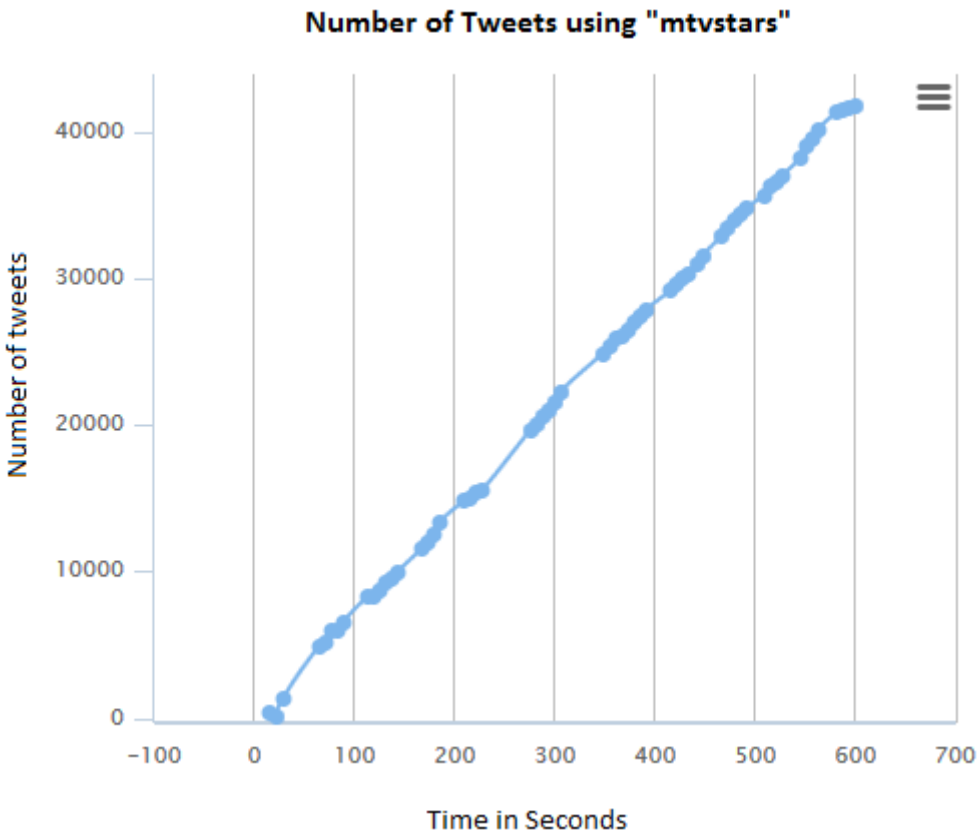


Figure 4:



21

Figure 5: Figure- 2 - 1 :



1

Figure 6: Figure-3- 1 :

4

1 : Top ten words in last 10 minutes

S. No.	Word	Frequency
1	Lady	24005
2	Today	20056
3	https	26558
4	?????	2619
5	Love	86288
6	???	29002
7	??? ????? ???	34406
8	2014	43101
9	Mtvstars	99449
10	???	90619

Figure 7: Table 4 -

4

2 : top ten languages in last 10 minutes

S. No.	Language	Frequency
1	Thai	359
2	Korean	426
3	French	435
4	Turkish	491
5	Indonesian	621
6	Spanish	1258
7	Arabic	1560
8	Russian	2109
9	Japanese	6957
10	English	8114

Figure 8: Table 4 -

4

Twits frequency	Time duration in seconds	Twits frequency	Time duration in seconds	Twits frequency	Time duration in seconds
405	15	15051	215	29158	415
100	22	15401	221	29589	421
1444	29	15557	227	30017	427
2031	35	16281	233	30374	433
2876	41	16689	240	30939	442
3570	47	17104	246	31601	448

Figure 9: Table 4 -

[Marketwired ()] , Marketwired . http://www.marketwired.com/press-release/apache-spark-beats-the-world-record-forfastest-processing-of-big-data-1956518.htm 2014.

[Donkin (2014)] , R B Donkin . http://people.apache.org/~rdonkin/hadooptalk/hadoop.html May 2014.

[Hadoop et al. (2014)] , Welcome Hadoop , To Apache , Hadoop . http://hadoop.apache.org May 2014.

[Mlib (2014)] *Apache Spark performance*, Spark Mlib . https://spark.apache.org/mllib October 2014.

[Big Data: what I is and why it mater ()] *Big Data: what I is and why it mater*, http://www.sas.com/en_us/insights/big-data/what-is-big-data.html 2014.

[Goldberg ()] *Cloud Security Alliance Lists 10 Big data security Challenges*, Michael Goldberg . http://data-informed.com/cloud-security-alliance-lists-10-big-data-security-challenges/ 2012.

[Cloudera et al. (2014)] *Community effort driving standardization of Apache Spark through expanded role in Hadoop Project*, Databricks Cloudera , Ibm , R Map , Open Sourcestandards . http://finance.yahoo.com/news/communityeffortdrivingstandardizationapache162000526.html July 1 2014.

[Basu (2014)] 'e-papers/big-data-real time health care-analyticswhite paper'. Abhi Basu . http://www.intel.com/content/dam/www/public/uen/documents/whit *Real-Time Healthcare Analytics on ApacheHadoopusingSparkandShark*, December 2014.

[Hoover ()] Mark Hoover . http://washingtontechnology.com/articles/2013/02/28/big-data-challenges.aspx *Do you know big data's top*, 2013. 9.

[Hurst ()] Steve Hurst . http://www.scmagazine.com/top-10-security-challenges-for-2013/article/281519/ *To 10 Security Challenges for 2013*, 2013.

[Securosis ()] Securosis . https://securosis.com/assets/library/reports/SecuringBigData_FINAL.pdf *Securing Big Data: Security Recommendations for Hadoop and No SQL Environment*, 2012.

[Stella ()] *Spark for Data Science: A Case Study*, Casey Stella . http://hortonworks.com/blog/spark-data-science-case-study/ 2014.