2017 International Conference on Intelligent Computing and Control (I2C2)

Social Media Analytics Based on Big Data

Farzana Shaikh
Department of Information Technology,
M.H. Saboo Siddik College of Engineering,
Byculla, Mumbai-400008, India.
farzanashaikh.117@gmail.com

Firdaus Rangrez
Department of Information Technology,
M.H. Saboo Siddik College of Engineering,
Byculla, Mumbai-400008, India.
firdausrangrez@gmail.com

Afsha Khan
Department of Information Technology,
M.H. Saboo Siddik College of Engineering,
Byculla, Mumbai-400008, India.
afshak574@gmail.com

Uzma Shaikh
Department of Information Technology,
M.H. Saboo Siddik College of Engineering,
Byculla, Mumbai-400008, India.
shaikhuzma86@gmail.com

***Abstract*—In today's competitive world, data driven decision making plays very important role. There is an increased demand to process large amount of data to make future prediction in various domains so as to gain competitive advantages. "Big Data" is broadly recognized to serve this purpose. Twitter is a Social Media Platform where each individual can express his/her own thoughts, opinion or feedback on different topics in form of Tweets. These topics can be from any domain such as Political, Business, Medical, Education and so on. In our system, Tweets are processed in Hadoop Framework to present Analytics of public opinion in demographic form.**

***Keywords*—Big Data, Hadoop Framework, Social Media, Tweets, Dictionary, Analytics.**

I. INTRODUCTION

We are living in an information age where petabytes of data is generated daily. The term “Big Data” deals with high volume of data sets which are unstructured or semi-structured. These data sets are usually generated at very fast rate. Big Data is not only about large amount of data sets but also what we are going to do with it. Big data provides the repository to store, process and analyze these data sets [16]. Social Media Platform Twitter is counted as one of the important sources of Big Data as millions of Tweets are generated each day [1][2][12]. Tweets are the short messages in which one’s opinion or feedback about any topic can be expressed. Number of Twitter users is also increasing day by day. Fig.1 shows the statistics for the same [13]. Twitter can be the rich source for gathering public opinion [10]. Such a mammoth amount of static or real time data is very difficult to process using traditional approaches [16]. “Big Data” serves this purpose. Hadoop is the Big Data processing platform. Proposed System uses Hadoop to process public opinion information i.e. Tweets. By analyzing public opinion information, Individuals or Organizations can work towards betterment of their process/product [9].

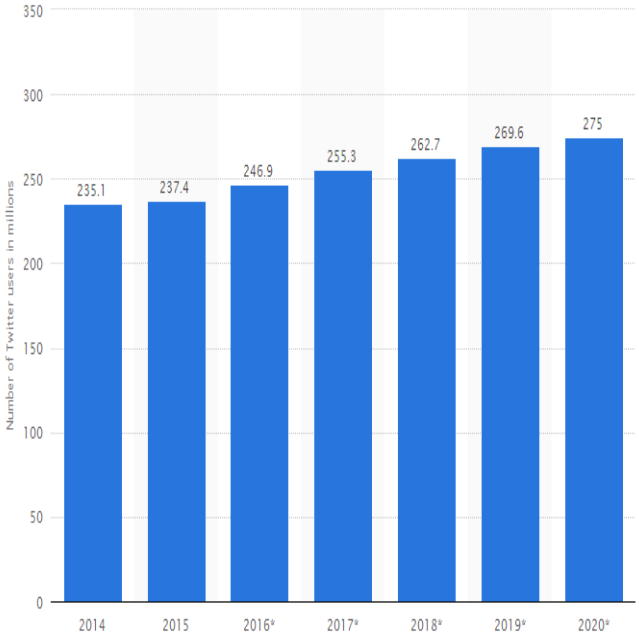

Fig. 1. No. of Twitter users worldwide from 2014 to 2020 (in millions) [13]

II. MOTIVATION

- Millions of tweets are tweeted on Twitter [12]. Twitter is one of the Social Media Platforms from where public opinion information i.e. Tweets, can be gathered [8].
- Tweets are short messages which include unstructured and structured data. Structured data such as Followers and Hashtags will be easier to process.
- It is very difficult to process millions of Tweets using traditional approaches [16].
- In most instances, the public opinion information is all form of unstructured data. Earlier it was very difficult to process the unstructured data. Also processing time was more.

- By gaining the review or opinion from public regarding a particular product/process one can make enhancement in their development process and can set marketing strategies accordingly.
- Even Educational domain requires Analytics System to make the decision for the betterment of their organization which in turn helps them to maintain their organizational branding.
- There is a need to respond quickly to the problems faced by public. By responding in a timely manner, organization can gain customer's trust and loyalty. Our system does Real Time Analytics of Twitter data in less time.
- Also our system provides mechanism to measure success of any product or event as it figures out sentiment or attitude of the user towards the given product/event.
- Visualization of huge volume of data is quite challenging. Big Data Visualization is of utmost importance which makes Analytics more readable and understandable.

III. EXISTING SYSTEM

Social Media generates high volume of structured and unstructured data daily [12][3]. Traditional approaches such as Relational Database Management Systems (RDBMS) are not able to process the large amount of unstructured data [16]. RDBMS is used to process small amount of structured data. The existing systems are time consuming as it can process data at a very low speed [4]. There is a limitation on data size also. The processing takes place on few machines so the load is all upon them. Those machines are not fault tolerant by nature. Therefore to overcome above drawbacks our system "Social Media Analytics Based on Big Data" is introduced. Table I. is the comparison of Traditional system and Big Data.

Table I. Traditional System vs. Big Data [4]

Sr. No.	Traditional System	Big Data
1.	It has data size limitation.	It supports virtually unlimited data volume.
2.	It has processing speed limitation.	It has virtually unlimited processing capability.
3.	Traditional system works well with structured data.	Big Data can work with unstructured, semi-structured or structured data.
4.	It is not suitable for adhoc processing.	It is best suited for adhoc system reporting.
5.	It is costly and time consuming.	Data management is easy and cost effective.
6.	Horizontal scaling is costly and usually requires downtime.	It uses commodity hardware to scale up hardware which further reduces cost of cluster.
7.	No inherit fault tolerance is provided.	It is fault tolerant by nature.

IV. PROPOSED SYSTEM

Proposed System will perform analysis of Twitter data i.e. Tweets, on any topic from any domain. Millions of Tweets are generated daily. Tweets are short messages which are used to convey sentiment or opinion that people have about what is going on in the world around them. Real Time Tweets are collected and processed to know the public opinion/sentiment/attitude towards the given topic in any domain such as Political, Business, Medical, Education and so on. Our system will categorize the overall polarity for the retrieved Tweets as Positive or Negative or Neutral. Overall polarity is determined by the aggregate of polarity of all matched Tweet words from dictionary. Analytics is represented in demographic form. Sentiments Polarities can be viewed Location (Country) wise or Timestamp wise in graphical form for the fetched Tweets. Followers Count of the users who have expressed positive sentiments are represented in graphical form for the collected Tweets. Trending Hashtags can also be viewed from the fetched Tweets. Public opinion will help Individuals or Organizations for decision making. Organizations can respond to the problems/reactions of people in very short time. Proposed system is using open source Hadoop Framework which allows huge amount of unstructured, semi-structured or structured data sets to be stored and processed in a distributed manner by applying Map Reduce model [11]. Hadoop Framework is able to process millions of Tweets efficiently. Proposed System makes use of the following:

A. Hadoop Map Reduce Model

Map Reduce Model has "map" function which transforms data sets into key/value pairs. Each of these data sets will then be sorted by their key and will be given to "reduce" function which is used to combine the values of the same key into a single key/value as shown in Fig. 2. [5].

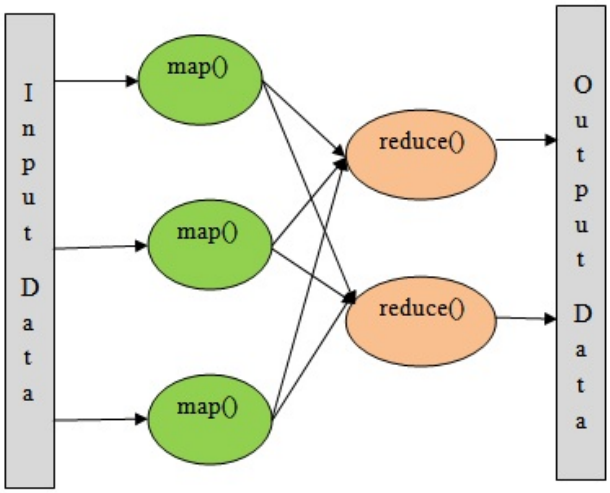

Fig. 2. Map Reduce Functioning [5]

B. Hadoop Distributed File System (HDFS)

The Hadoop Distributed File System (HDFS) is a distributed file system which provides storage of very large amount of data. HDFS is fault tolerant and scalable. It allows fast access to data sets across Hadoop clusters. These clusters consist of commodity hardware. To store such high volume of data, HDFS breaks entire data into smaller blocks and stores those blocks across multiple machines of a cluster. HDFS also replicates these blocks so as to achieve reliability. A Hadoop cluster consists of collection of nodes called as Name Node and Data Node. Name node manages namespace and mapping of blocks to Data Nodes. Data Nodes stores Data itself [6]. HDFS architecture is shown in Fig. 3.

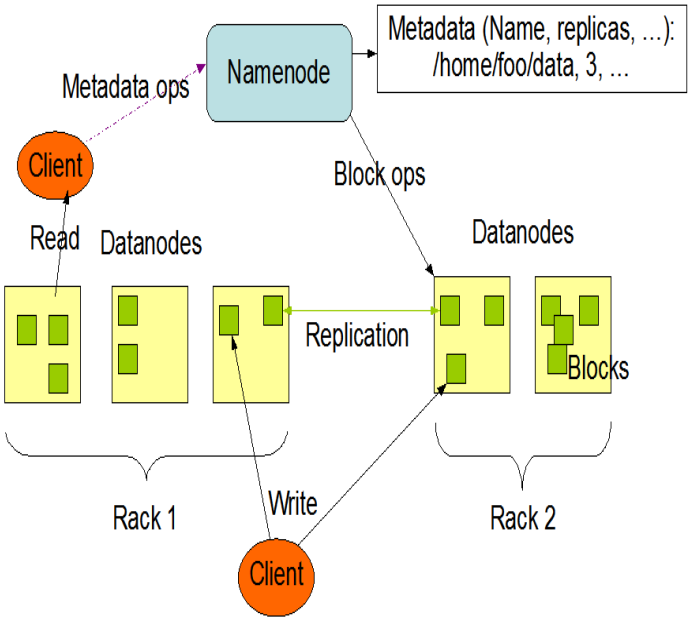

Fig. 3. HDFS Architecture [6]

C. Hive and Hive Query Language (HiveQL)

Apache Hive is an open source software which is built on top of Apache Hadoop to make querying and analysis easy. It manages large data sets stored in distributed storage by querying using Hive Query Language [14].

D. Twitter Streaming API's

The real time Tweets are required for our system which will be obtained by Twitter's streaming API's. Twitter Streaming API's allows accessing Tweets data stream in very short time [7]. Twitter application will be created in order to get various keys. Those keys can be incorporated in our system to get faster retrieval of Tweets. These keys are required: Consumer Key, Consumer Secret, OAuth Access Token, OAuth Access Token Secret.

E. SentiWordNet Dictionary

In our system, polarity of Tweets is calculated by using SentiWordNet Dictionary. SentiWordNet Dictionary has a vast collection of words. It is most widely used Dictionary for opinion mining. It assigns sentiment score either positive or negative or neutral to each word [15].

F. Block Diagram & it's Working

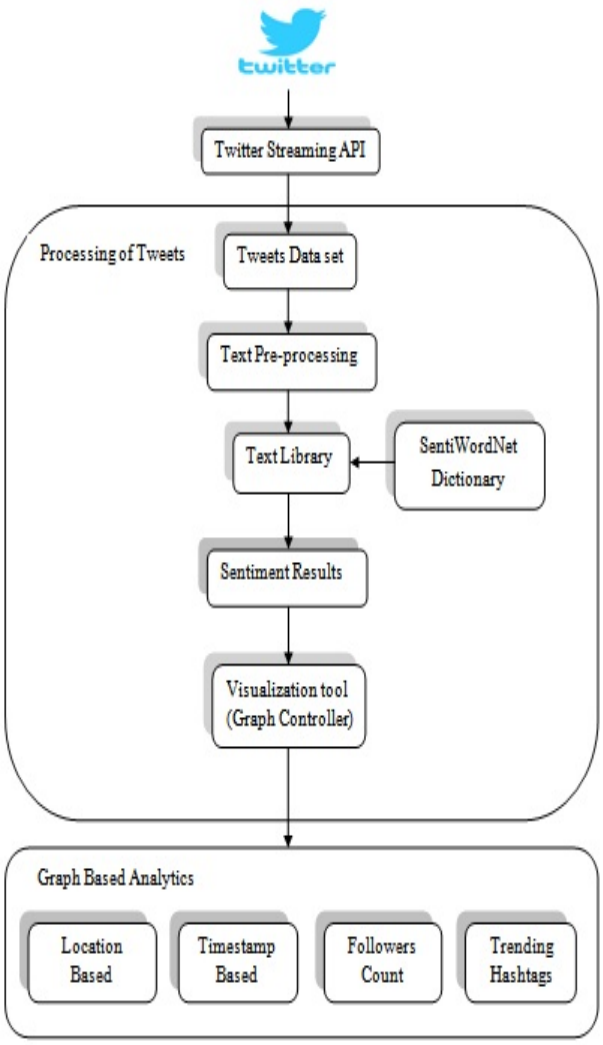

Fig. 4. Block Diagram of Social Media Analytics Based on Big Data

Fig. 4 shows Block Diagram of proposed system. Its working is given below:

Step 1: Tweets related to any topic from any domain can be retrieved from the search panel provided in our system.

Step 2: The Tweets Data Set will be collected from Twitter using Twitter Streaming API's.

Step 3: The retrieved Tweets will then be given to the Hadoop Distributed File System (HDFS).

Step 4: The Raw Tweets which is in unstructured format (JSON format) will be fetched from HDFS and stored in structured format in Hive using Hive Query Language. Storing in structured format requires Text Pre-processing.

Step 5: For analyzing each Tweet, we need to break it into individual words called Tokens. This process is called Tokenization. These tokens are stored in Hive using HiveQL.

```
create table text_l1 as select id, words from raw_tweets lateral view explode(sentences(lower(text))) dummy as words;
```

Example:

Tweet:
going to watch baahubali 2

Explode each tweet - words on multiple rows
going
to
watch
baahubali
2

Step 6: Retrieve only words.

```
create table words_l2 as select id, word from text_l1 lateral view explode(words) dummy as word ;
```

going
to
watch
baahubali

Step 7: SentiWordNet Dictionary is used for polarity calculation. Matching tokens from dictionary will be assigned polarity as Positive/Negative/Neutral.

Step 8: Perform Join Operation with Dictionary to get the polarity for each word present in a Tweet.

```
create table words_l3 as select id, words_l2.word, case d.polarity when 'negative' then -1 when 'positive' then 1 else 0 end as polarity from words_l2 left outer join dictionary d on words_l2.word = d.word;
```

Step 9: Overall polarity is determined by the aggregate of polarity of all the matching tokens.

```
create table tweets_sentiment as select id, case when sum( polarity ) > 0 then 'positive' when sum( polarity ) < 0 then 'negative' else 'neutral' end as sentiment from words_l3 group by id;
```

Step 10: Finally, Analytics will be represented in graphical form using Graph Controller. Location (Country) Based, Timestamp Based, Followers Count Based Polarities will be shown in various graphs for the fetched Tweets. Trending Hashtags are also displayed from the fetched Tweets.

V. RESULTS

Fig.5. Retrieved Tweets for the searched topic

Our system provides search panel for fetching the desired Tweets. Tweets for “srk” are retrieved as shown in Fig. 5.

Fig.6. Sentiments Based on Location (Country)

Sentiments of people (Positive/Negative/Neutral) towards the given topic based on location (Country) are represented in graphical form. Fig. 6 shows sentiments of people in U.S. for the fetched Tweets on “Hillary Clinton”.

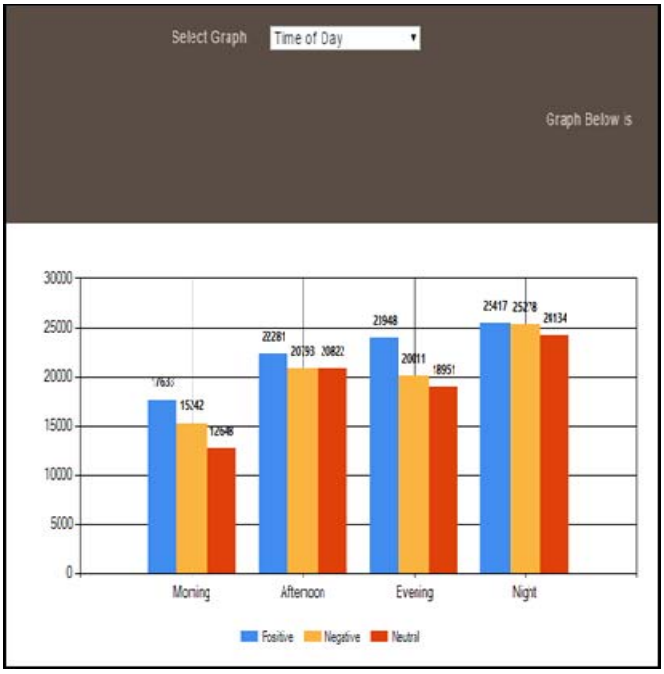

Fig. 7. Sentiments Based on Time of the Day

Fig. 7 shows sentiments of people across the world according to Time of the day (Morning, Afternoon, Evening, Night) for the fetched Tweets on “Hillary Clinton”.

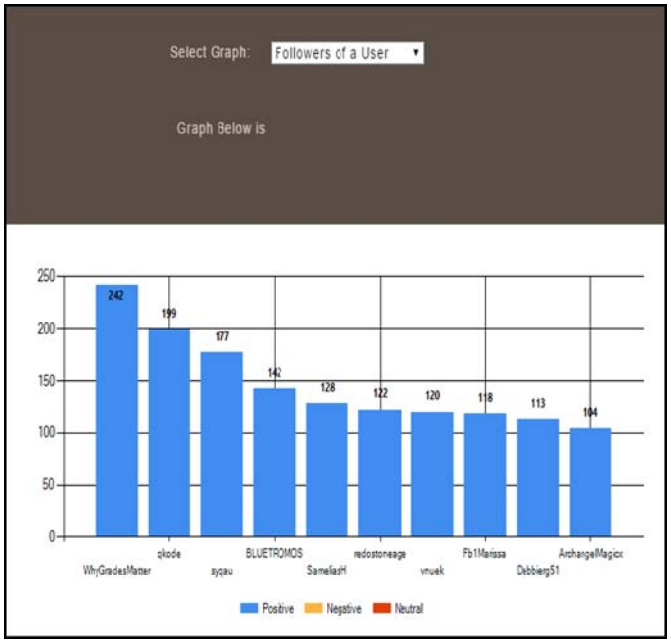

Fig. 8. Sentiments Based on Followers Count

Followers Count of the users across the world who have expressed positive opinion can be shown. Fig. 8 depicts Followers Count of positive sentiment users of the fetched Tweets on “Hillary Clinton”.

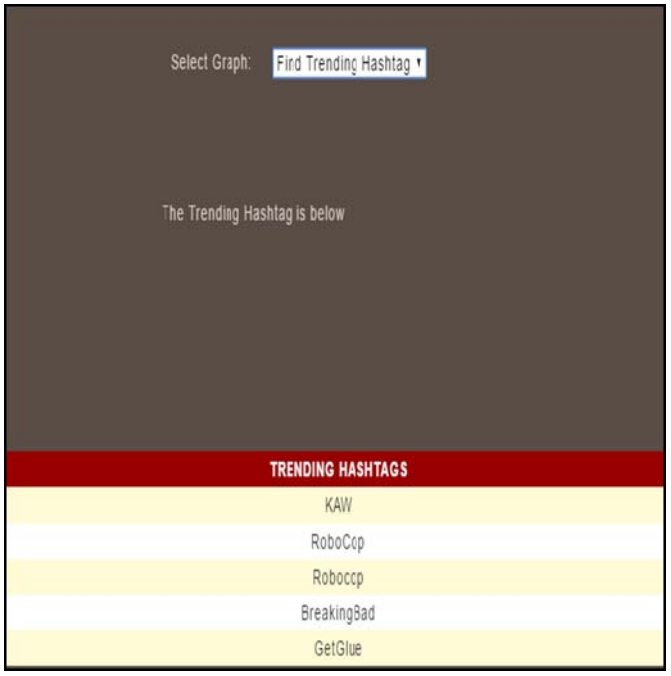

Fig. 9. Trending Hashtags

Trending Hashtags from the fetched Tweets on “Hillary Clinton” are shown in Fig. 9.

VI. CONCLUSION

The main aim of proposed system is to analyze the Tweets within minimum time and demonstrate the results which are processed Tweets depicting the opinion of the people across the world including their Location (Based on Countries), Time of the day (Morning, Afternoon, Evening, Night), Followers Count in graphical forms. Trending Hashtags can also be viewed. Individual or Organization can take necessary action after analyzing sentiments of people towards any topic from any domain. System can give more accurate results if it is executed repeatedly with the better set of Tweets.

VII. FUTURE WORK

The system is implemented on a single machine; it can also be implemented on cluster of machines. Processing of Tweets in English language is within the scope of the system. In future, it can be implemented for other languages too. Recommendation algorithms can be applied to find out the reasons behind each sentiment polarity (Positive/ Negative/Neutral).

REFERENCES

[1] "Big Data & Analytics", [Online] Available: https://www.cmswire.com/big-data/cmswires-hits-of-2015-big-data-analytics/, [Accessed on Nov 15 2016]

[2] "Impact of Big Data on Social Media",[Online] Available: https://tech.co/impact-big-data-social-media-marketing-strategies-2016-01, [Accessed on Nov 15 2016]

[3] "Understanding Big Data World", [Online] Available: https://www.pearsonitcertification.com/articles/article.aspx?p=2427073&seqNum=2, [Accessed on Dec 4 2016]

[4] "Hadoop Ingestion System", [Online]Available: https://ngvtech.in/droidhub/hadoop-ingestion-systems/, [Accessed on Jan 10 2017]

[5] "How Hadoop MapReduce Work", [Online] Available: https://dzone.com/articles/how-hadoop-mapreduce-works, [Accessed on Aug 20 2016]

[6] "HDFS Architecture Guide", [Online] Available: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html/, [Accessed on Aug 20 2016]

[7] "Streaming APIs", [Online] Available: https://dev.twitter.com/streaming/overview/, [Accessed on Jan 4 2017]

[8] Songtao Shang; Minyong Shi; Wenqian Shang; Zhiguo Hong, "Research on public opinion based on Big Data," in *Computer and Information Science (ICIS), 2015 IEEE/ACIS 14th International Conference* , pp.559-562, June 28 2015.

[9] Segev, A.; Chihoon Jung; Sukhwan Jung, "Analysis of Technology Trends Based on Big Data," in *Big Data (BigData Congress), 2013 IEEE International Congress on* , vol., no., pp.419-420, June 2013.

[10] Wenbo Wang; Lu Chen; Thirunarayan, K.; Sheth, A.P., "Harnessing Twitter "Big Data" for Automatic Emotion Identification," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)* , pp.587-592, 3-5 Sept. 2012.

[11] "Apache Hadoop", [Online] Available: https://hadoop.apache.org/, [Accessed on Feb 4 2017]

[12] "Twitter usage Statistics", [Online] Available: https:// www.internetlivestats.com/twitter-statistics/, [Accessed on March 15 2017]

[13] "Number of Twitter users Worldwide from 2014 to 2020 in millions",[Online] Available: https://www.statista.com/statistics/303681/twitter-users-worldwide/, [Accessed on April 25 2017]

[14] "Apache Hive", [Online] Available: https://hive.apache.org/, [Accessed on April 20 2017]

[15] "SentiWordNet", [online] Available:https://sentiwordnet.isti.cnr.it/,[Accessed on Jan 4 2017]

[16] "BigData",[online]Available:https://en.wikipedia.org/wiki/Big_data/,[Accessed on July 25 2016]

[17] "Twitter Logo", [online] Available: https://brand.twitter.com/logo/, [Accessed on Feb 20 2017]

[18] "Hortonworks Data Platform", [online] Available: http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0-Win/bk_QuickStart_HDPWin/bk_QuickStart_HDPWin-20150721.pdf/, [Accessed on Aug 13 2016]