BigData Analytics using Apache Spark in Social Media

Lokesh Dubey
Indiana University
3209 E 10th St
Bloomington, Indiana 47408
ldubey@indiana.edu

ABSTRACT

Social Media, as organic and diverse it is, is also a vital source of very large amount of data. And it increased even more with the introduction of Smart Phones. As it has been established very well in recent years that Social Media and the data derived from it not only helps with decision making for substantial businesses but also helps considerably for marketing and increasing business revenues. We explore various benefits and techniques of using Big Data technology Apache Spark in unison with enormous Social Media data and how it can overcome the shortcomings of Traditional Analytics Technologies. We illustrate application of Spark with Social Media data with a few Social Media use cases pertaining to product enhancement and marketing.

KEYWORDS

i523, hid309, Apache Spark, Hadoop, Social Media Analytics, Marketing, Cloud Computing

1 DATA

1.1 Traditional Data

Criticality of data has been an accepted fact right from the beginning of computing world. In fact, when the first computer was invented the first few operations and features provided by first generation computers where simple file creations, saving the data and performing calculations on them. Since then types of data and size of data has come a long way along with the advancement in the technology. However, before the introduction of Social Media to the computing world, traditional data typically remained highly structured, static and rigid [11]. A substantial part of traditional data was generated and handled in Banking, Health and Insurance domains. But most of this data stayed extremely monotonous, relational in other words structured and rigid. These data types were always constant, brittle and it was very easy to assess their growth if any in future. Because of which it was easy to forecast what kind of infrastructure and technology needed to be procured.

1.2 Social Media Data

In recent years social media has proliferated at such an exponential rate that the sheer amount of data that is being generated is becoming a challenge for traditional technologies to handle. Initially social media was, as the name suggests, medium for socializing. And the primary focus of social media was upon social interactions of humans on a digital platform which helped with fast progressing life style where the frequency of physical social interactions was reducing day by day. Social media sites, like Facebook[1], Twitter[2] etc., became extremely popular at least within young generation. It started to become an extremely simple way to socialize, catch up with friends, and sharing life events with others merely by login on those sites on internet with the luxury of not moving physically anywhere and save resources and time. And of course internet's vast reach and speed made it a very likable and a viable solution. This, in effect became a huge source of data generation. Every social media user, logging on to a social media site, sharing his own information in form of photos, videos and text, and not just that, a user liking, viewing others photos, videos, status shares, became a huge source of data generation [6]. Computing world was vary of this vital change and looked at this immense amount data as a viable source for gathering different statistics of different demographics [3].

However, with the introduction of Smart Phones the whole paradigm of social media changed [1]. Now, rather than waiting for getting an internet access on a desk to visit social sites, a user had access to all these social sites on his hands. Which essentially provided a way to socialize, share and grasp, all the information from friends and other public information through out the day [14]. This major paradigm shift in social media not only increased the amount of data that was being generated but also provided various other perspectives on how better this data can be used. The data that is being generated by Social Media is used in multitude of domains with a variety of motives [15]. Data sources can be Streaming APIs, where data is being provided almost in real time, simple REST APIs to retrieve data and possibly files archived on file servers to be consumed. Data formats can be comma or any delimiter separated files, JSON[3] files, html etc.

In addition to the wide variety of data sources and formats what can be mined from this data is also very diverse [15] [12]. Commercially, this data can be used to improve on the products by mining for constructive feedback for the productions and the same data can be used in marketing for increasing sales and driving the decision making process. But there are endless possibilities of using this immense amount of data for other analysis. For example, early detection and tracking of diseases and epidemics [19].

2 BIG DATA PROCESSING

2.1 Traditional Analytics Methodologies, Challenges

Data and specifically Big Data has been around for some time. However, the data has almost always have been structured. There

[1] https://www.facebook.com/
[2] https://twitter.com/
[3] http://www.json.org/

have been a lot of work done in the field of data warehousing and there are some other traditional appliance based warehousing systems like Netezza[4], Teradata[5] which are also used for a lot of analytics. These systems however have their own limitations and if not all, they do not perform well on the contemporary Social Media data [16] when the objective is to handle complete data in real time. There are some explicit and implicit problems using these traditional technologies and methodologies with Social Media data. Explicit problems are the type of data. There are multiple data sources, formats and types in social media which are difficult to be incorporated in these traditional systems. For example, the data sources could be a stream of twitter, unstructured live chat data from a chat server, various formats of data like JSON, comma separated. These data types can very well be integrated within these systems as well but there's a huge cost to massage and transform the data to be made usable by these traditional systems. Other than the explicit challenges there are some implicit challenges which are faced when trying to ingest and processing data for which the size and its frequency is not fixed. In traditional technologies like Netezza, Teradata we have to understand our data first not only on the structure but also on the size of the data before hand so that appropriate capacity on the appliance can be procured. But with Social Media data, which can be of any type, format, size, its difficult to scale the traditional systems this quickly [11]. Because of these challenges the traditional analytics systems are not completely obsolete as there are still a lot of other data sources other than social media but for Social Media specifically when our concern mostly tackling this immense amount of data its better to move towards a technology which can handle any sources of data, formats and types of data which can be achieved very easily with a technology based off Cloud Computing. With these challenges, the traditional data methodologies face some limitations, which are summarized by Krishnan with the following sentence 'Lack of scalability due to processing complexities coupled with inherent data issues and limitations of the underlying hardware, application software, and other infrastructure' [13].

2.2 Cloud Computing

As explained in traditional analytics methodologies and traditional data before one of the major challenges in handling the ever growing and dynamic data was being able to foresee the amount of data that needs to be processed and to be able to estimate the amount of hardware/infrastructure to be procured. Both of these problems couldn't be solved by traditional warehousing and on premise or even off premise labs with high performance infrastructure. Because these systems are not scalable to the needs of big data. As far as the infrastructure for Big Data is concerned introduction of Cloud Computing was a ground breaking advancement which opened up the doors for numerous possibilities [13]. With on demand computing and on domain scale up, scale down features which were provided by a 3rd party Infrastructure as a Service (IaaS) service providers it was extremely easy to manage the dynamic data. With Virtualization, in Cloud Computing, big Infrastructure providers take care of all of the infrastructure needs and provide on demand service to provide high configuration and high performance virtual machines on demand, which can also be backed up passively in form of snapshots and can be recovered back to avoid any kind of data or infrastructure loss. These features are seldom available in traditional on premise infrastructure and if it is then it comes with a very high cost. From cost point of view as well these machines can be purchased on hourly billing rates and the user only pays for the time the machine was used. Other than compute (Memory and CPU) advancements were made on making storage highly scalable, fast and manageable like the vms in form of SAN[6]Storage with very high IOPS and Object Storage[7] for providing highly reliable and easy and remotely accessible data storage for huge data archival or even for using the same storage for Big data I/O even over network [10]. In last decade, Cloud computing grew much more than just being IaaS providers and various other providers used IaaS underneath and started providing Platform As As Service and eventually Software As A Service. It is explained later in more detail but Cloud Computing has progressed enough to even provide MapReduce and Hadoop platforms as a service.

2.3 Hadoop

Biggest breakthrough in the field of Big data were the two research paper released by Google Inc 'The Google File System' [8] and 'MapReduce: Simplified Data Processing on Large Clusters' [7]. This was the next stage of progression from traditional analytics methodologies explained in previous sections. Similar principles of Google File System and MapReduce were developed into open source tools Apache Hadoop Distributed File System (HDFS) and Apache MapReduce and they were collectively called Apache Hadoop. Both of these tools were designed to work on commodity hardware and to work in unison on a cluster of machine to provide a distributed filesystem which supported MapReduce principle of breaking the work in smaller pieces to be done in parallel on individual cluster machines (Map) and then join the work together to provide a final result (Reduce). Gradually, lot of other opensource tools were developed to work with HDFS and MapReduce to handle different types and formats of data. Tools like Apache Hive[8], Apache HBase[9] were developed and were widely used for providing a relational access point to structured and non structured data respectively. There are numerous other tools which were developed other than these to provide a wide spectrum of flexibility to Hadoop platform to deal with nearly any type, format or data source. Namely, Apache Pig[10], Apache Flume[11], Apache Kafka[12], Apache Sqoop[13].

2.4 Challenges with contemporary technologies

Hadoop MapReduce and HDFS were considerably used, with other required Hadooop tools based on need, and are still utilized substantially for a wide variety of big data processing, transformation

[4]https://www-01.ibm.com/software/data/netezza/
[5]http://www.teradata.com/
[6]https://www.snia.org/education/storage_networking_primer/san/what_san
[7]https://www.ibm.com/cloud-computing/learn-more/what-is-object-storage/
[8]https://hive.apache.org/
[9]https://hbase.apache.org/
[10]https://pig.apache.org/
[11]https://flume.apache.org/
[12]https://kafka.apache.org/
[13]http://sqoop.apache.org/

and analytics. As explained previously Social Media domain is so dynamic and growing that the amount of data being generated [6] grew so large that MapReduce started to appear as it has reached to maximum of performance it can provide and there was a need for an alternative [9]. On the other hand, however, HDFS still remains a very important pillar in this domain. Continuous advancements to increase the performance of HDFS are still being made to increase the I/O performance of the data like storing data in Apache Parquet[14], Apache Avro[15], Apache ORC[16] file formats to serialize the data and to increase the performance while reading bulk of data. In addition to that different file compression formats like normal gzip, snappy etc. are also used these days to compress the files while writing them on HDFS to impart a shorter footprints of file sizes which in turn increases the performance on writing and reading files to and from HDFS [2].

MapReduce has certain challenges when the amount of data grows too big. The fundamental problem with MapReduce is that in principle it creates multiple stages for any type of query of data transformation and all of the data output of these intermediate stages is stored on HDFS and then it is read back from HDFS for subsequent stages. Because, MapReduce works on these files directly from HDFS in principle it spends a lot of time doing I/O on HDFS and eventually the disk. This performance is good for a certain amount of data but as we established that Social Media data is huge and ever growing, MapReduce is not a viable solution because of low performance. There are various use cases on social media data analytics where the results are expected to be retrieved very quickly. For example, There are use cases where a 3D model is generated to visualize the social media data usage and it demands a very high performance throughput from the system on a very huge size of data sets [18]. Many social media data sources are not really static data and are streams of data like Live Chat data or Live Google My Business Reviews where if the analysis is to do live reporting of the data as it is being generated MapReduce may not be the optimal choice.

2.5 Apache Spark and its benefits

Apache Spark[17] was an open source tool developed keeping these shortcomings of MapReduce in mind [9]. Spark on one hand works on the similar concept of MapReduce but the data for different Stages of the execution is not stored on HDFS or actual disks. Spark attempts to store as much data as possible in the Memory of the distributed cluster. Because Memory (RAM) are must more faster than any kinds of disks SATA, SSD etc. the performance of Spark is much faster than MapReduce jobs [9]. Spark necessarily doesn't require a Cluster of machine and can work on single nodes as well. However, the real throughput and performance of Spark data processing, transformation and analysis jobs is when running it on distributed system. Spark provides fundamental data structures like Resilient Distributed DataSets (RDDs), DataFrames and DataSets which can work on highly distributed systems and also provide immense amount of APIs to make the data processing quicker and easier to Develop [5]. Spark doesn't have a specific requirement to be used on a Hadoop Cluster but in the interest of this work we'll focus only on applications of Spark where HDFS provides the distributed file system to work hand in hand with Distributed Data of Spark Data types. In addition to that, if required, Spark can also work with other data types directly like Object Storage, File Data as well. Spark, can also work with Mesos or in standalone mode on Cloud.

There are many other features that make Spark an extremely viable solution for Social Media Analytics. Hadoop, on one hand, resolved a lot of issues with having different formats of data and types of data but there are still a lot of other analysis which require data to be learned on the fly etc. Spark provides a lot of libraries and APIs which can directly handle these different sources of data. Spark Streaming provides APIs to read data from streams of data like Twitter Stream etc, Spark SQL[18] provides APIs to run SQL like queries on data retrieved, Spark Machine Learning[19] library provides APIs to create models on the data to make prediction analysis and finally Spark GraphX[20] library provides APIs for graph data and for graph parallel computation.

3 USE CASE

At this point we have established the limitations of Traditional Analytics technologies and methodologies which are limited to Traditional Data analytics needs, whereas, for the ever growing and extremely dynamic data of Social Media we need much more than Traditional Methodologies. Even the contemporary tools which are widely used in Social Media lack performance and supported features which can fit all kind of data analysis needs of Social Media [13]. Two substantial usages of Social Media data other than many are collecting data to find insights on how the product itself can be improved or to find how the product is doing in the market and to advertise it better.

3.1 Product enhancements

A use case of the first category is reviews. Yelp[21] and Google My Business[22] are crowd sourcing sites which helps getting reviews from all the users of Yelp and Google about various businesses. A substantial part of these businesses are restaurants, where users can provide their feedback of all of these restaurants in form of textual information as reviews. And can also provide ratings in stars to the restaurants. This data has a great potential of providing great insights of what the restaurants can improve upon. We do know that there's a lot of research and technologies available for Natural Language Processing (NLP) and Sentiment analysis. But the problem here is not how to find insights, that is the data science part of the problem. The problem is data engineering and the sheer amount of data that is being generated. With Spark this data can be ingested to high performance clusters directly via Apache Flume and Kafka to Spark Streaming APIs. By applying the Lambda architecture [17] spark can provide a continuous ingestion of data at real time and it can processed, transformed (possibly NLP) and can be aggregated to generate reports in real time for different

[14]https://parquet.apache.org/
[15]https://avro.apache.org/
[16]https://orc.apache.org/
[17]https://spark.apache.org/
[18]https://spark.apache.org/sql/
[19]https://spark.apache.org/mllib/
[20]https://spark.apache.org/graphx/
[21]https://www.yelp.com/sf
[22]https://www.google.com/business/

businesses. This is not possible with any of the traditional analytics technologies or even contemporary Hadoop MapReduce.

3.2 Descision Making for Marketing

Another use case for the second category is marketing. Many Social Media Sites are being used to market products these days in form of advertising. It could be a sponsored post in someone's timeline (Facebook, Instagram) or it could simply be an ad which shows up on ad space on your webpage or in the social media application. This advertising depends highly on conversion rate of any user i.e. the user actually clicks or visits the site or product being advertised. It is highly possible that user might not be interested in that kind of product at all. There are some lower level analytics done in the browsers themselves these days where cache of the browsing history of any user can be utilized to show an ad of a product which the user was looking at sometime back. This particular advertising is called Behavioral Retargeting [20]. But that's very straight forward problem to solve and there are many 3rd pary providers like Adroll, Retargeter who provide these services. The advertising can be improved to a very larger extent if the social media interactions of the users like what kind of video the user liked, what photos user is more interested in, what kind of demographics and geography the user has affinity to [4]. Numerous such statistics, if processed and mined, a good machine learning model can be created using machine libraries of spark to get this data in real time via Spark Streaming APIs and after processing, analyzing data with lambda architecture, final reports can be generated or if required actions can be triggered in real time to choose what category of the ads for a particular user has a high chance of getting a conversion. This again is something where considering the amount of data and the very high throughput expectancy its not possible to achieve this with traditional analytics technologies [13].

4 FUTURE WORK

After this work it can be said with at most ease that Apache Spark is one of the best available technology for Social Media Analytics and as we've have established its viability in some use cases as well, a good meaningful next step on this work would be to implement a spark project on a virtualized environment and integrate it with a Social Media data source. This can help quantify the performance and other aspects of application of Spark in Social Media and Big Data.

5 CONCLUSION

After exploring all types of data available, traditional and contemporary, specifically Social Media, we established the enormity, wide variety and growth rate of Social Media Data. We also examined the shortcomings of the traditional technologies and even the contemporary big data methodologies and how they are not a best fit for the analytical and data processing needs for Social Media data. After looking closely at the wide set of features and custom solutions that Apache Spark can provide we were successfully able to showcase how Spark can be a best bit for all the data processing and analytics needs of Social Media data. We also discussed the application of Spark on Social Media Data with a few example use cases. The use cases we discussed are much broader and are a simple overview of how Spark can be utilized best with the contemporary data analysis needs with the highly volatile and exponentially growing social media data of various types, sources and formats.

REFERENCES

[1] Abir S. Al-Harrasi and Ali H. Al-Badi. 2014. The Impact Of Social Networking: A Study Of The Influence Of Smartphones On College Students. *Contemporary Issues in Education Research (CIER)* 7, 2 (2014), 129–136. https://doi.org/10.19030/cier.v7i2.8483

[2] Vaddeman B. 2016. *Beginning Apache Pig* (1st. ed.). Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-2337-6_15

[3] Bogdan Batrinca and Philip C. Treleaven. 2015. Social media analytics: a survey of techniques, tools and platforms. *Springer London* 30 (2015), 89fi?!116. https://doi.org/10.1007/s00146-014-0549-4

[4] Ricardo Limongi Frana Coelho, Denise Santos de Oliveira, and Marcos Inácio Severo de Almeida. 2016. Does social media matter for post typology? Impact of post content on Facebook and Instagram metrics. *Online Information Review* 40 (2016), 458–471. https://doi.org/10.1108/OIR-06-2015-0176

[5] Jules Damji. 2016. A Tale of Three Apache Spark APIs: RDDs, DataFrames, and Datasets. (2016). https://databricks.com/blog/2016/07/14/a-tale-of-three-apache-spark-apis-rdds-dataframes-and-datasets.html accessed 2017.

[6] Sarah Dawley. 2016. A Long List of Facebook Statisticsfi!?And What They Mean For Your Business. (2016). https://blog.hootsuite.com/facebook-statistics/ accessed 2017.

[7] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: Simplified Data Processing on Large Clusters. *Commun. ACM* 51, 1 (Jan. 2008), 107–113. https://doi.org/10.1145/1327452.1327492

[8] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. 2003. The Google File System. *SIGOPS Oper. Syst. Rev.* 37, 5 (Oct. 2003), 29–43. https://doi.org/10.1145/1165389.945450

[9] Satish Gopalani and Rohan Arora. 2015. Article: Comparing Apache Spark and Map Reduce with Performance Analysis using K-Means. *International Journal of Computer Applications* 113, 1 (March 2015), 8–11. Full text available.

[10] INTEL. 2012. Cloud Computing Research for IT Strategic Planning. (2012). https://www.intel.com/content/dam/www/public/us/en/documents/reports/next-generation-cloud-networking-storage-peer-research-report.pdf accessed 2017.

[11] George J. Trujillo Jr., Charles Kim, Steven Jones, Rommel Garcia, and Justin Murray. 2015. *Virtualizing Hadoop* (1st. ed.). VMware Press. http://www.pearsonitcertification.com/articles/article.aspx?p=2427073&seqNum=2

[12] Andreas M. Kaplan and Michael Haenlein. 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons* 53, 1 (2010), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

[13] K. Krishnan. 2013. *Data Warehousing in the Age of Big Data* (1st. ed.). Elsevier Science. https://books.google.com/books?id=8ngws8f_lNsC

[14] Amanda Lenhart. 2015. Teens, Social Media & Technology Overview 2015. (2015). http://www.pewinternet.org/2015/04/09/teens-social-media-technology-2015/ accessed 2017.

[15] NCSU.EDU. 2014. Social Media Data Research and Use. (2014). https://www.lib.ncsu.edu/social-media-archives-toolkit/research-and-use/research accessed 2017.

[16] Abderrazak Sebaa, Fatima Chikh, Amina Nouicer, and Abdelkamel Tari. 2017. Research in Big Data Warehousing using Hadoop. *Journal of Information Systems Engineering & Management* 2, 10 (2017). https://doi.org/10.20897/jisem.201710

[17] Gwen Shapira. 2014. Building Lambda Architecture with Spark Streaming. (2014). https://blog.cloudera.com/blog/2014/08/building-lambda-architecture-with-spark-streaming/ accessed 2017.

[18] Zachary Weber and Vijay Gadepally. 2014. Using 3D Printing to Visualize Social Media Big Data. *CoRR* abs/1409.7724 (2014). http://arxiv.org/abs/1409.7724

[19] Yusheng Xie, Zhengzhang Chen, Yu Cheng, Kunpeng Zhang, Ankit Agrawal, Wei-Keng Liao, and Alok Choudhary. 2013. Detecting and Tracking Disease Outbreaks by Mining Social Media Data. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI '13)*. AAAI Press, Beijing, China, 2958–2960. http://dl.acm.org/citation.cfm?id=2540128.2540556

[20] Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. 2009. How Much Can Behavioral Targeting Help Online Advertising?. In *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*. ACM, New York, NY, USA, 261–270. https://doi.org/10.1145/1526709.1526745

6 BIBTEX ISSUES

Warning–numpages field, but no articleno or eid field, in mapreducegoogle

Warning–numpages field, but no articleno or eid field, in ghemawatgoogle

Warning–empty address in georgevirthadoop

Warning–empty address in krishnan

Warning–page numbers missing in both pages and numpages fields in sebabigdata

Warning–page numbers missing in both pages and numpages fields in 3dvisual

Warning–numpages field, but no articleno or eid field, in detectingoutbreaks

Warning–numpages field, but no articleno or eid field, in bretargeting

(There were 8 warnings)

7 ISSUES

no title

you do not use “quotes” properly

you need to *emphasize* and not “quote”

this is cool

Have you written the report in the speciffied format?

Have you included an acknowledgement section?

Have you included the paper in the submission system (In our class it is git)?

Have you speciffied proper identiffication in the submission system. THis is typically a form or ASCII text that needs to be ffilled out (In our case it is a README.md ffile that includes a homework ID, names of the authors, and e-mails)?

Have you included all images in native and PDF format in the submission system?

Have you added the bibliography ffile that you managed (In our case jabref to make it simple for you)?

In case you used word have you also provided the jabref?

In case of a class and if you do a multi-author paper, have you added an appendix describing who did what in the paper?

Have you spellchecked the paper?

Are you usingaandtheproperly?

Have you made sure you do not plagiarize?

Is the title properly capitalized?

Have you not used phrases such as shown in the Figure below, but instead used as shown in Figure 3 when referring to the 3rd ffigure?

Have you capitalized fiFigure 3fi, fiTable 1fi, ... ?

Have you removed any ffigure that is not referred explicitly in the text (As shown in Figure ..)

Are the ffigure captions bellow the ffigures and not on top. (Do not include the titles of the ffigures in the ffigure itself but instead use the caption or that information?

When using tables have you put the table caption on top?

Make the ffigures large enough so we can read the details. If needed make the ffigure over two columns?

Do not worry about the ffigure placement if they are at a different location than you think. Figures are allowed to ffoat. If you want you can place all ffigures at the end of the report?

Are all ffigures and tables at the end?

In case you copied a ffigure from another paper you need to ask for copyright permission. IN case of a class paper youmustinclude a reference to the original in the caption.

Do not use the word fiIfi instead use we even if you are the sole author?

Do not use the phrase fiIn this paper/report we showfi instead use fiWe showfi. It is not important if this is a paper or a report and does not need to be mentioned.

Do not artifficially inffate your paper if you are bellow the page limit and have nothing to say anymore.

If your paper limit is 12 pages but you want to hand in 120 pages, please check ffirst ;-)

Donotusethecharacters & # % _ put a bakslash berfore them

If you want to say and do not use & but use the word and.

Latex uses double single open quotes and double single closed quotes for quotes. Have you made sure you replaced them?

Pasting and copying from the Web often results in non ascii characters to be used in your text, please remove them and replace accordingly.