Walmart's Sales Data Analysis- A Big Data Analytics Perspective

Manpreet Singh*[§], Bhawick Ghutla[†], Reuben Lilo Jnr[†], Aesaan F S Mohammed[†] and Mahmood A Rashid[†‡§]
* National Training and Productivity Centre, Fiji National University, Samabula, Suva, Fiji
[†]School of Computing, Information and Mathematical Sciences, The University of the South Pacific, Suva, Fiji
[‡]Institute for Integrated and Intelligent Systems, Griffith University, QLD, Australia
[§]Corresponding Authors: manpreet.singh@fnu.ac.fj OR mahmood.rashid@usp.ac.fj

***Abstract*—Information technology in this 21st century is reaching the skies with large-scale of data to be processed and studied to make sense of data where the traditional approach is no more effective. Now, retailers need a 360-degree view of their consumers, without which, they can miss competitive edge of the market. Retailers have to create effective promotions and offers to meet its sales and marketing goals, otherwise they will forgo the major opportunities that the current market offers. Many times it is hard for the retailers to comprehend the market condition since their retail stores are at various geographical locations. Big Data application enables these retail organizations to use prior year's data to better forecast and predict the coming year's sales. It also enables retailers with valuable and analytical insights, especially determining customers with desired products at desired time in a particular store at different geographical locations. In this paper, we analysed the data sets of world's largest retailers, Walmart Store to determine the business drivers and predict which departments are affected by the different scenarios (such as temperature, fuel price and holidays) and their impact on sales at stores' of different locations. We have made use of Scala and Python API of the Spark framework to gain new insights into the consumer behaviours and comprehend Walmart's marketing efforts and their data-driven strategies through visual representation of the analysed data.**

***Keywords*—Big Data Analytics; Hadoop Distributed File Systems; Apache Spark; MapReduce**

I. Introduction

We all are constantly thinking about the future and what is expected to happen in the coming weeks, months and even years, and to be able to do so, a look at the past is mandatory. Business needs to be able to see their progress and the factors affecting their sales [1]. In this technological era of large scale data, businesses need to rethink on the modern approaches to better understand the customers to gain a competitive edge in the market. Data is worthless if it cannot be analysed, interpreted and applied in context [2]. In this work, we have used the Walmart's sales data to create business value by understanding customer intent (sentiment analysis) and business analytics. A picture speaks a thousand words and business analytics would help paint a picture through visualization of data to give the retailers insights on their business. With these insights the businesses can make relevant changes to their strategy for the future to maximize profits and success. Most of the raw data, particularly large scale datasets do not offer value in its unprocessed state. By applying the right set of tools [3], we can pull powerful insights from this stockpile of bits.

The main focus here is to read and analyse the Walmart's available datasets to produce insights and the company's overall overview. The retail stores sell products and gain profit from it. There are a lot of subsidiaries of the stores network which are scattered on various geographical locations. As the network of stores is huge and located at different geographical locations, the company would not fully understand the customer needs and market potentials at these various locations. In this work, we used the gathered store sales datasets of Walmart to understand the factors affecting the sales for example, the unemployment rate, fuel prices, temperature and holidays in the different stores located at different geographical locations so that the resources can be managed wisely to maximize on the returns. These insights can help retailers comprehend market conditions of the various factors affecting sales for example Easter holiday would induce a spike in sales and retailers can better allocate resources (supply of goods and human resources). Thus, customer demands are observed accordingly based on the above factors.

Moreover, the big data application enables retailers to use historical dataset to better observe the supply chain, then a clear picture can be obtained about a particular store whether they are making profit or are under loss. When data is properly analysed, we will start to see the patterns, insights and the big picture of the company. Then the required suitable actions can be applied accordingly. This will help optimize operations and maximize sales and profit. Additionally, these datasets are used to predict/forecast future sales for the coming weeks so that the retailers have a fair picture of what the company's future will be like and it can act as a warning for the company if it is going downhill with its return on investments [4].

Apache data science platforms, libraries, and tools are used in this work by testing and implementing the software development tools and environments dealing with Big Data technology. Tools like Hadoop Distributed File Systems (HDFS) [5], Hadoop MapReduce framework [6] and Apache Spark along with Scala, Java and Python high-level programming environments are used to analyse and visualize the data.

II. Related Work

In 2015, Harsoor & Patil [4] worked on forecasting Sales of Walmart Store using big data applications: Hadoop, MapReduce and Hive so that resources are managed efficiently. This paper used the same sales data set that we utilized for analysis, however they forecasted the sales for up coming 39 weeks. Their strategy included the collection of huge Sales data and transferred on HDFS [5] and performed Map Reduce which later due to enormous data size, proved difficult to draw conclusion. Thus Hive processing was done to calculate average sales feature for all 45 stores and 99 departments. Machine learning algorithm, R programming was used for statistic computing. Henceforth, Holt Winters [4] was used for training dataset provided by Walmart and then sales prediction was done. Subsequently the predicted sales were given graphical representation using Tableau interactive data visualization.

In 2013, Katal, Wazid, & Goudar [7] performed thorough studies about handling a Big Data; their issues, challenges, various tools and good practices. Technical challenges like scalability, fault tolerance, data quality and heterogeneous data processing was also mentioned. They have proposed Parallel Programming Model like Distributed file system, MapReduce [6] and Spark as a good tool for Big Data [7].

In 2015, Riyaz& Surekha [8] worked on MapReduce on Hadoop to build a data analytical engine for weather, temperature analysis for National Climate Data Centre. This paper had all the details and results about MapReduce program execution. Their findings concluded that MapReduce with Hadoop [6] is good for weather data analysis and temperature can be analysed efficiently which at the end is important for a lot of industries [8].

In 2017, Chouksey & Chauhan [9]performed weather forecast using MapReduce and Spark in order to formulate earlier weather warnings so that people and businesses are prepared for undesirable weather condition. Weather has greater influence in agriculture sector, sporting, tourism and government planning. Various weather sensors/parameters like wind speed, temperature, humidity, pressure, and other factors was analysed with the technology benchmark comparison for Hadoop MapReduce and Spark. Eventually the performance of Spark for weather analytics is proven to be better in results.

In 2013, Zaslavsky, Perera, & Georgakopoulos [10] recommended the use of Hadoop, Apache Spark and NoSQL Technology to process billions of sensing devices data. They explained Sensing as a service and big data; where storage as well as processing of this huge data is becoming a challenge. This sensing devices are connected to computer networks and thus generates enormous data on daily basis.

In 2016, Sharma, Chauhan, & Kishore, [11] performed comparative study between Hadoop MapReduce and Spark. The paper enclosed chart comparison between these two tools; advantages and disadvantages in big data analysis context. Through this comparative study, they concluded that Spark is much better [12], [13] than MapReduce; however, it also depended on the area of analysis [11].

In 2017, Inoublia, Aridhib, Meznic, & Jungd [14] worked on experimental evaluation and a comparative study of Healthcare scientific applications which decided health status using interconnected sensors over the human body. This included breath, insulin, cardiovascular, glucose, blood and body temperature. They recommended Spark because processing stream of health data, sending and processing iteratively cannot be handled or supported by MapReduce model.

In [7], [9]–[14], the authors have recommended Apache Spark as a better option in terms of faster and having a very intelligent way of processing data in-memory (memory caching), rather than reading it back and again from the disk all the time.

III. Background

Retailers plan to insure success or maximum profit by learning about the factors that affects their sales and in what measure. Big organizations and retailers around the world, such as the one this work is based on, Walmart Stores, Inc., try to maximize the profit by providing maximum customer satisfaction in all geographical locations to maintain the standards of the stores.

Walmart sales data is considered for this work since most of the challenges faced by the company is universal or that all other big retailers are facing similar problems that is to maintain, manage and organize their retail shops data in a way that it provides useful insights on the company as an overall retailer, individual shops or only for the departments in the shops itself. The retailers have to overcome a lot of similar challenges to stay on top of a competitive market [15].

Retailers have to manage resources wisely to maximize the profit while at the same time minimizing the cost. Retailers fail to gauge market potential at the right time. When there is a sudden spike in sales and the retailers are caught off-guard there might not be enough stock or enough staff to meet the customer needs thus losing potential sales.

With insights to the causes of the spike in sales and the factors affecting it, the retailers can make better resource allocation like getting more employees to the store with more customers or transfer more stock to that store.

Planning of the store can be smarter, providing better human resource management, better supply management [16]. By observing to past helps get an idea of sales in stores and its separate departments and predictions for the future sales can be made. These predictions will be used as a guideline or to mark a trajectory for the future and it will allow the retailers to make relevant changes to the objective of the stores for better success in the future.

A. *Problem studied*

Retailer's first priority is usually to understand their customers to be able to satisfy their needs so that these customers will return to the store for future needs, thus increasing the product demands and adding to the business value. These businesses want this information to plan where and when to invest profitably.

B. *Tools and techniques applied*

The tools and techniques used for this work includes the collection of Huge Walmart sales datasets stored in CSV format. We used Apache Spark with a build version of Hadoop leveraging HDFS [5] as a data storage option. Apache Spark is a framework capable of handling both batch and stream processing on the same application at the same time [7], [9]–[14], [17]. Our development tools include InteliJ Idea Community Edition [18] and iPython Notebook [19], [20]. InteliJ Idea was integrated with Spark instead of using the traditional Spark shell. After we configured our environment, our first task was to load the files as spark dataframes. Dataframe is a distributed collection of data organized into named columns which is equivalent to tables in RDMS [21]. The spark dataframe API was designed to make big data processing simple for a wider audience and also it supports distributed data processing in general purpose programing languages like Scala, Python and Java. Spark supports reading files from popular data types like JSON files, Parquet files, HIVE table, HDFS, cloud storage (S3) or external RDMS [22]; however, the CSV file formats are not natively supported. Thus, we used a separate library instead, called Spark-CSV developed by Databricks [23] to load the datasets. As the files are stored in dataframes, we query the data using spark-SQL component. We then apply MapReduce functions on the datasets using Spark-SQL. After applying some operations on the data such as grouping, sorting etc., we save the files to HDFS as CSV. We then use Ipython Notebook [19], [20] as pySpark shell to read the processed data for graphing. We use Pandas library [20] to visualize the datasets.

IV. Technology Implementation

Walmart has 45 stores in geographically diverse locations, each of the store having 99 departments. The dataset of 3 years contains the weekly sales and the factors affecting sales such as (Temperature, fuel price, unemployment rate, holiday) for each store locations. To analyse the dataset and find relationship between the sales and affecting factors Apache spark and its various libraries was chosen as shown in Figure 1.

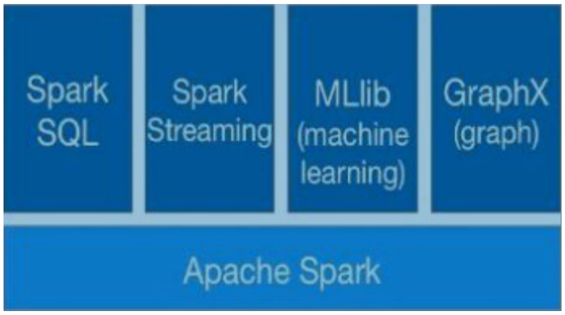


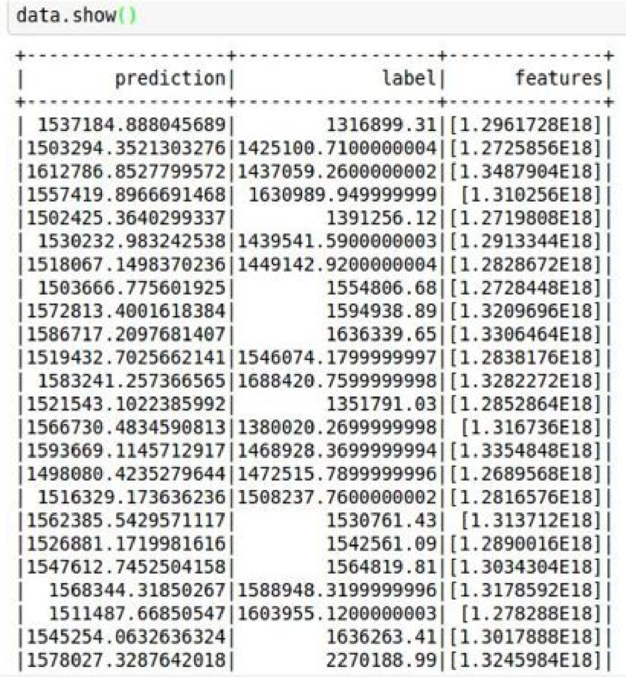
Figure 1: Apache Spark components [13].

With Apache spark all the essentials are coupled in a single system in various libraries with the need to only call the libraries needed.

Spark is not the only solution out there. Hadoop MapReduce is also a good choice but it focuses more on batch processing as it was designed in a context where size, scope and data completeness are more important than speed. Spark is 100 times faster [7], [9]–[14], [17] than Hadoop MapReduce, and is one of the best solution out of the box. The comparisons are presented in Figure 2.

Spark SQL was used to map and reduce the dataset to a key-value format to compare. The key is a concatenated value in the format STORE_DATE and the values are the sales, temperature, fuel price, holiday and unemployment rate.

After getting the analysed data in key and value it is easier to graph and see relationships between values of the date and store location using GraphX library provided by Apache Spark using its python API which will be seen in Figure 4, 5, 6. Machine learning library is employed with a simple regression model to predict future sales. The regression model finds relations between variables to see trends. Predictions can be more accurate with multiple variable correlation between temperature, fuel price, holidays, unemployment rate and Store sales can be used to get more accurate predictions (see Figure 3).



```
data.show()
```

```
+------------------+------------------+--------------+
|        prediction|             label|      features|
+------------------+------------------+--------------+
| 1537184.888045689|        1316899.31|[1.2961728E18]|
|1503294.3521303276|1425100.7100000004|[1.2725856E18]|
|1612786.8527799572|1437059.2600000002|[1.3487904E18]|
|1557419.8966691468| 1630989.949999999| [1.310256E18]|
|1502425.3640299337|        1391256.12|[1.2719808E18]|
| 1530232.983242538|1439541.5900000003|[1.2913344E18]|
|1518067.1498370236|1449142.9200000004|[1.2828672E18]|
| 1503666.775601925|        1554806.68|[1.2728448E18]|
|1572813.4001618384|        1594938.89|[1.3209696E18]|
|1586717.2097681407|        1636339.65|[1.3306464E18]|
|1519432.7025662141|1546074.1799999997|[1.2838176E18]|
| 1583241.257366565|1688420.7599999998|[1.3282272E18]|
|1521543.1022385992|        1351791.03|[1.2852864E18]|
|1566730.4834590813|1380020.2699999998| [1.316736E18]|
|1593669.1145712917|1468928.3699999994|[1.3354848E18]|
|1498080.4235279644|1472515.7899999996|[1.2689568E18]|
| 1516329.173636236|1508237.7600000002|[1.2816576E18]|
|1562385.5429571117|        1530761.43| [1.313712E18]|
|1526881.1719981616|        1542561.09|[1.2890016E18]|
|1547612.7452504158|        1564819.81|[1.3034304E18]|
|  1568344.31850267|1588948.3199999996|[1.3178592E18]|
|  1511487.66850547|1603955.1200000003| [1.278288E18]|
|1545254.0632636324|        1636263.41|[1.3017888E18]|
|1578027.3287642018|        2270188.99|[1.3245984E18]|
```

Figure 3: Forecasts of the future sales given by the simple regression model.

V. RESULTS AND DISCUSSION

The following are the results of our paper:

1) Retailers need to plan and evaluate according to the market driving factors which are, and not limited to, the temperature, unemployment rate, fuel prices holidays, human resources, geographical location and many more.
2) Effective and efficient supply chain, inventory, human resource management is needed to avoid losing competitive edge in the market, especially planning sales at different locations.
3) We analysed largest tycoon retailer, Walmart's sales dataset to gain valuable and analytical insights, especially determining customer behaviours at a desired time in a particular store at different geographical locations.
4) There was 45 Walmart stores with different department (approximately 99), weekly sales, temperature, etc. located in different regions dataset[1].
5) We have used Big Data Technology: MapReduce with Hadoop, Apache Spark combined big data fundamentals in high level API's for Scala, Python and Java

[1] Dataset can be retrieved from following link [22]: https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data

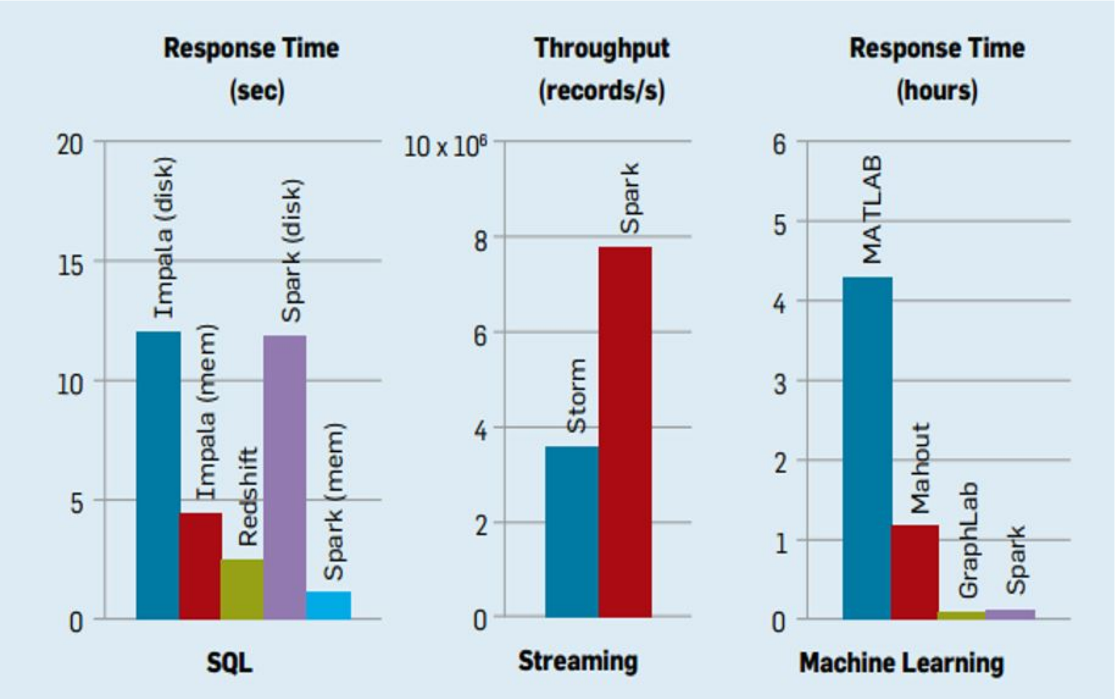


Figure 2: Comparing Spark's performance with several widely used specialized systems [5].

to analyse this tremendous Weekly Sales dataset and outline a pattern and meaning to it.

6) Hadoop MapReduce is good for batch processing [4], [24] whereby intermediate data between Mapper and Reducer is stored on disk which clearly depicted to us that there is latency causing slower performance and result generation.

7) Spark does in-memory computing [7], [9]–[14], [17] whereby intermediate data is stored in the memory to avoid latency which is in former, MapReduce. Spark is mostly used for stream data processing, graphing, machine learning and iterative computing.

8) Hence, for our experiment Spark had much better performance and job execution to show pattern and Sales analysis in terms of market condition like weather, temperature, fuel pricing, holiday and many more.

9) Finally, we used Sparks with its python API, (Pandas – python library for graphing) for graphical visualization.

The analysis achieved from Walmart's data for the 421571 tuple needed to be visualized for better insights and understanding for improved decision making and acquire advantage in terms of resource allocations.

In figure 4, we have data visualized to comprehend the pattern of weekly sales for all 45 stores across different locations observing the years from 2010 to 2012, fuel price and Temperature respectively.

According to Data visualization in Figure 3, we have observed sales at beginning of all the three years. The first quarter for each year, i.e. January-March, the Sales is low (decreasing) for entire Walmart stores at different locations. However, as we approach second quarter (April – June), Sales intensifies upwards for 2010, 2011, and 2012. Similarly, in third quarter (July-September) all the 45 stores all around has declining Sales values. Eventually, in final quarter (October- December) we noticed spike in Sales across all 45 stores as we approach end of the year for 2010, 2011 and 2012.

From the Figure 5, the following observations has been organized in Table I. Therefore, more sales occur when fuel price is at reasonable range of $2.90 to $3.80 per liter. From the Figure 6, the following observations has been organized in Table II. Therefore, more sales occur when temperature is at reasonable $\approx 21^0$ to $\approx 60^0$ in Fahrenheit scale which is neither too cold or too hot, more of normal temperature.

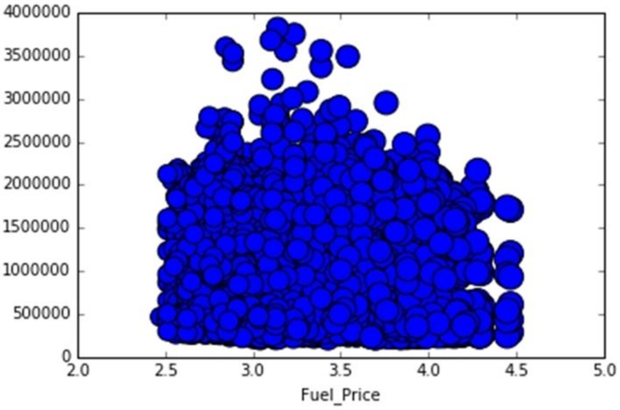


Figure 5: Fuel price effect on all weekly sales: - summarized information of the figure is outlined in Table I.

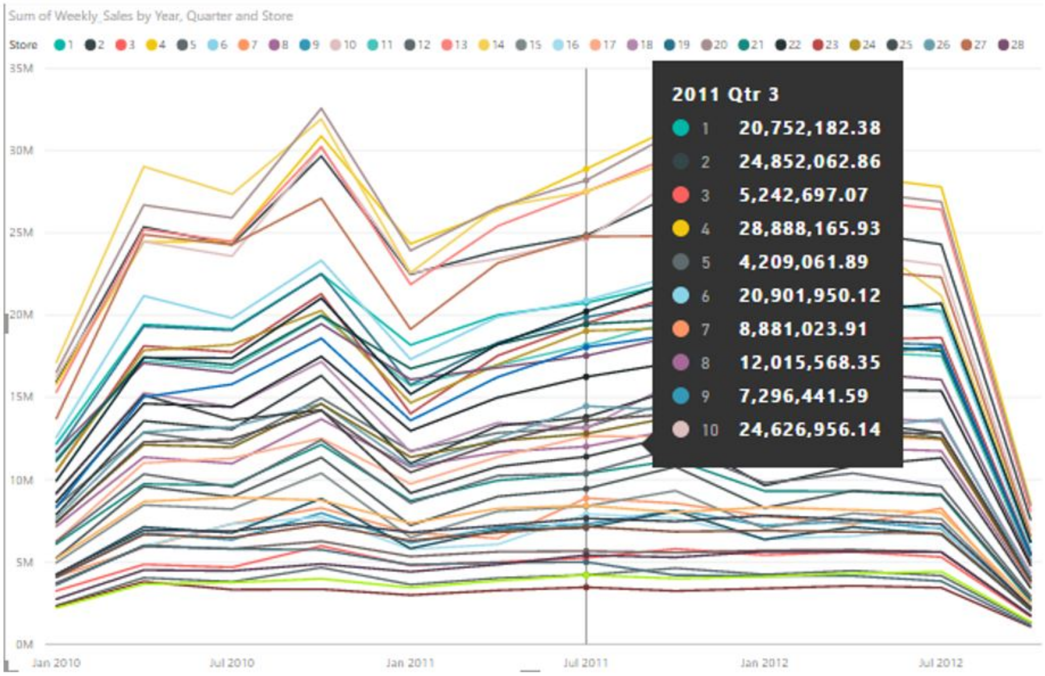


Figure 4: Quarterly Sales Graph from year 2010–2012.

Table I: Fuel price effect on all weekly sales.

Fuel ($/Gal)	Total Sales
2.5 – 2.8	Sales ranging from ≈$500000 – ≈$3M
2.9 – 3.8	Sales ranging from≈$500000 – ≈$4M
3.9 – 4.5	Sales ranging from ≈$500000 – ≈$25M

Table II: Temperature effect on total weekly sales.

Temp (^{0}F)	Total Sales
0 – 20	Sales ranging from ≈$100000 – ≈$2M
21 – 60	Sales ranging from≈$100000 – ≈$4M
61 – 100	Sales ranging from ≈$500000 – ≈$3M

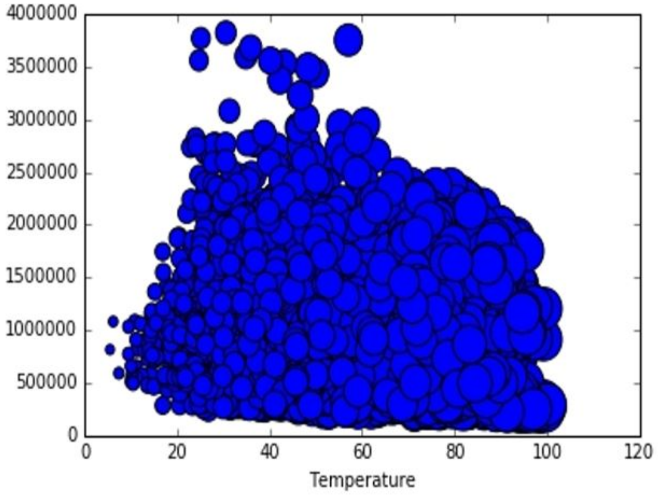


Figure 6: Temperature effect on total weekly sales:- summarized information of the figure 6 is outlined in Table II.

VI. CONCLUSION

In conclusion, Wal-Mart is the number one retailer in the USA and it also operates in many other countries all around the world and is moving into new countries as years pass by. There, are other companies who are constantly rising as well and would give Walmart a tough competition in the future if Walmart does not stay to the top of their game. In order to do so, they will need to understand their business trends, the customer needs and manage the resources wisely. In this era when the technologies are reaching out to new levels, Big Data is taking over the traditional method of managing and analyzing data. These technologies are constantly used to understand complex datasets in a matter of time with beautiful visual representations. Through observing the history of the company's datasets, clearer ideas on the sales for the previous years was realized which will be very helpful to the company on its own. Additionally, seasonality trend and randomness and future forecasts will help to analyse sale drops which the companies can avoid by using a more focused and efficient tactics to minimize the sale drop and maximize the profit and remain in competition.

REFERENCES

[1] M. Franco-Santos and M. Bourne, “The impact of performance targets on behaviour: a close look at sales force contexts,” *Research executive summaries series*, vol. 5, 2009.

[2] D. Silverman, *Interpreting Qualitative Data: Methods for Analyzing Talk, Text and Interaction 3rd Ed.* Text and Interaction, Sage Publications Ltd: Methods for Analyzing Talk, 2006.

[3] UBM. (2003) Big Data analytics: Descriptive vs. predictive vs. prescriptive. [Accessed 17 September 2017]. [Online]. Available: www.informationweek.com/about-us/d/d-id/705542

[4] A. S. Harsoor and A. Patil, “Forecast of sales of walmart store using Big Data application,” *International Journal of Research in Engineering and Technology*, vol. 4, p. 6, June 2015.

[5] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. Mccauley, M. J. Franklin, S. Shenker, and I. Stoica, “Fast and interactive analytics over Hadoop data with Spark,” *Usenix - The Advanced Computing Systems Association*, 2012.

[6] J. Dean and S. Ghemawat, *MapReduce: simplified data processing on large clusters*. Association for Computing Machinery, 2008.

[7] A. Katal, M. Wazid, and R. H. Goudar, *Big Data: Issues, Challenges, Tools and Good Practices*, 2013.

[8] P. A. Riyaz and V. Surekha, *Leveraging MapReduce With Hadoop for Weather Data Analytics*. OSR Journal of Computer Engineering (IOSR, 2015.

[9] P. Chouksey and A. S. Chauhan, *A Review of Weather Data Analytics using Big Data*. International Journal of Advanced Research in Computer and Communication Engineering, 2017.

[10] A. Zaslavsky, C. Perera, and D. Georgakopoulos, “Sensing as a service and Big Data,” in *Proceedings of the International Conference on Advances in Cloud Computing (ACC)*, I. Bangalore, Ed., July 2012.

[11] M. Sharma, V. Chauhan, and K. Kishore, “A review: MapReduce and Spark for Big Data analysis,” in *5th International Conference on Recent Innovations in Science*. 5: Engineering and Management, June 2016.

[12] H. Pandey, “Is Spark really 100 times faster on stream or its hype?” vol. 2, Sept 2016, [Online]. Available: [Accessed 2017]. [Online]. Available: www.quora.com/Is-{Spark}-really-100-times-faster-on-stream-or-its-hype

[13] T. A. S. Foundation, “Lightning-fast cluster computing,” *[Online]. Available: [Accessed Sept*, vol. 2017, 2017. [Online]. Available: {Spark}.apache.org/

[14] W. Inoublia, S. Aridhib, H. Meznic, and A. Jungd, *An Experimental Survey on Big Data Frameworks*, 2017.

[15] W. C. Kim and R. A. Mauborgne, “Blue ocean strategy, expanded edition: How to create uncontested market space and make the competition irrelevant,” vol. 2015.

[16] D. Läubli, G. Schlögl, and P. Silën, “Mckinsey & company,” *[Online]. Available: [Accessed Sept 2017*. [Online]. Available: www.mckinsey.com/industries/retail/our-insights/smarter-schedules-better-budgets-how-to-improve-store-operations

[17] J. Ellingwood, “Hadoop, storm, samza, Spark, and flink: Big Data frameworks compared,” *[Available Online][Accessed Sept 2017]*. [Online]. Available: www.digitalocean.com/community/tutorials/{Hadoop}-storm-samza-{Spark}-and-flink-big-data-frameworks-compared

[18] JetBrains. (2011) Intellij idea, the most intelligent java ide. [Online]. Available: www.resources.jetbrains.com/storage/products/intellij-idea/docs/Comparisons_IntelliJIDEA.pdf

[19] S. Duvvuri and B. Singhal, *Spark for Data Science*, ser. Analyze your data and delve deep into the world of machine learning with the latest Spark version. Packt Publishing Ltd, 2016.

[20] W. Mckinney, “Python for data analysis: Data wrangling with pandas, numpy, and ipython, o’reilly,” vol. 2012.

[21] A. Spark, “Spark sql and dataframe guide,” *[Online]. Available: [Accessed Sept 2017*. [Online]. Available: www.{Spark}.apache.org/docs/1.5.2/sql-programming-guide.html

[22] R. Xin, M. Armbrust, and D. Liu, “Introducing dataframes in apache Spark for large scale data science,” *[Online]. Available: [Accessed August 2017]*, February 2015. [Online]. Available: https://databricks.com/blog/2015/02/17/introducing-dataframes-in-{Spark}-for-large-scale-data-science.html

[23] S. Venkataraman, Z. Yang, D. Liu, E. Liang, H. Falaki, X. Meng, R. Xin, A. Ghodsi, M. Franklin, I. Stoica, and M. Zaharia. (2016) Sparkr: Scaling r programs with Spark.

[24] V. Nivargi, “Big Data: From batch processing to interactive analysis,” *[Online]. Available: [Accessed Sept 2017]*, 2013. [Online]. Available: www.clearstorydata.com/2013/01/evolving_big_data/