MIT 805 Assignment 2 Visualizing H&M Personalized Fashion Recommendations Dataset

Matthew Cameron Dickson
University of Pretoria
Department of Computer Science
Email: u15274170@tuks.co.za

1 Introduction

The globalization and digitization of the world has lead to an ever increasing generation of data from various sources especially with in organizations as they able to sell their products and services globally allowing them to spanned their business across multiple regions. This large complex data being generated is referred to as big data and is an important resource to stay competitive within the business markets in modern times.

The clothing retail sector can benefit from the use of big data in their organization in order to determine hidden trends that can give them a competitive advantage against there competitors either by reducing their costs or understanding their customer base. Stores such as H&M which have stores globally can benefit from technologies such Hadoop in processing their large transaction data that is generate daily across their multiple stores world wide [1].

This transaction data can be consider big data due to its large volume and velocity in which the data is generated. Hadoop is framework that allows for the distributed processing of big data across a cluster of computers [2] making it an ideal technology to use in processing large transactions data in order to get it into a form which can be used in analytics to extract knowledge.

Thus Hadoop and related technologies will be used in order to process and extract knowledge form the H&M personalized fashion recommendation dataset as a test case to see the effectiveness of using technologies such as Hadoop in processing retail transaction data and whether such a system reduces the processing time required for large data such as retail transaction data which is being generated continuously and in large volumes.

2 Hadoop

Hadoop consist of three main components these components are the following [3]:

-Hadoop Distributed File system (HDFS).
-MapReduce.
-Yet Another Resource Negotiator (YARN).

The HDFS is a distribution file system which provides fault tolerance functionality and was designed to run efficiently and reliably on commodity hardware. HDFS internal architecture revolves around a salve master relationship and thus provides high accessibility and throughput of large data. HDFS can be clustered which allows it to store data across multiple computers. Its splits files into fixed size data blocks which it then stores across the configured cluster [3]

MapReduce which is a programming model first introduce by google in order to process large amount of data in parallel. It fundamentally splits data and aggregates it again into a more reduce form [3]. It consist of two functions namely the map and reduce functions. The map function is responsible for filtering and sorting where it produces intermediate key, value pairs which then the input of the reduce function. The reduce function is responsible for aggregating the intermediate key value pairs in order to produce a reduce representation of the data to be finally use in an analysis task

YARN is responsible for allocating resources to all applications running in the Hadoop cluster. One of its main function is scheduling jobs that are required to be run on the cluster [4].

Hadoop also includes multiple add on components such as Apache hive, Apache pig, Sqoop and Apache spark that can be integrated into the Hadoop platform to enhance the functionality it provides.

3 Implementation

The technology stack that will be used in this system to process the transaction history which includes years from 2018 to 2020 of H&M stores will be a sandbox system provided from Nortonworks [5] which is a platform that is built with Hadoop and other integrated technologies like:

-Apache hive.
-Ambari.
-Apache pig.
-Sqoop.
-Oozie.
-ZooKeeper.
-Apache storm.
-Ranger.
-Apache Zeppelin.

Nortonworks sandbox environment was chosen because setting it up was easy and time efficient. The environment also came with many other helpful tools that work within a Hadoop cluster such as Ambari which made working with Hadoop and HDFS cluster easier. These extra helper tools would require alot of configuration if they were manually installed on a clean Hadoop cluster. The Hadoop cluster consist of only one node that was running a linux redhat 64 operating system. One node was only used due to resource limitation, however more nodes could be added relatively easy in the configuration if more processing power is needed. However, this was unnecessary due to this system only being a test case which was processing three years of H&M transactions.

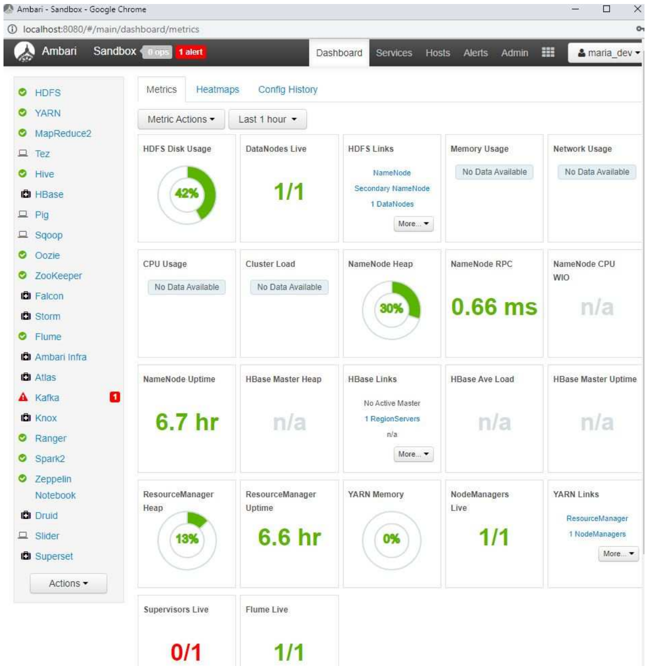
The Nortonworks Hadoop sandbox system was used in order to bin the transactions into years and month using map reduce functions. This will be done in order to see what months out of the years most transactions occurred and which years had the most transactions for all H&M stores and out of the transaction data which year was the most profitable for the H&M group.

This can be beneficial as it could provide knowledge for H&M management and allow them to come up with future strategic plans such as which months they should include more inventory in their stores to meet customers demand. It also provides knowledge which years where under performing in terms of customer sells and revenue and could give management the required reasoning to audit those years to further understand the root causes of the bad performance for those particular month/years.

4 Visualization

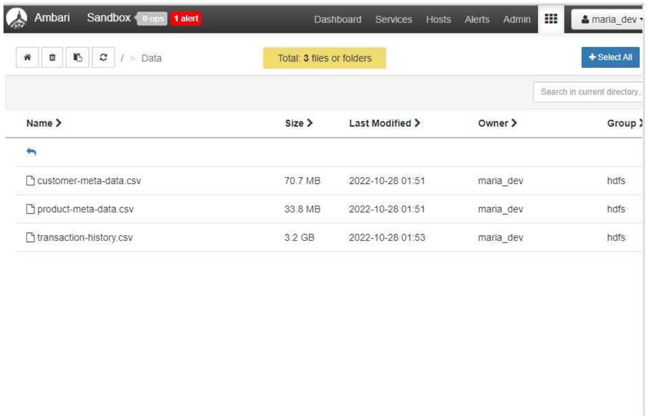
The results of using map reduce operations on the transaction history were put into a python script for visualization and the observations during the analytics were the follows:

- June and July were the most profitable months during the three years of 2018, 2019 and 2020
- February was the worst performing months for the years of 2018, 2019 and 2020
- 2020 had the least amount of value for H&M stores. One possible reason for this is that during Covid less people were buying expensive clothes and opting to buy cheaper clothes.
- 2019 had the largest amount of transactions and value out of the years.



HDFS Disk Usage 42%
DataNodes Live 1/1
HDFS Links: NameNode, Secondary NameNode, 1 DataNodes
Memory Usage: No Data Available
Network Usage: No Data Available
CPU Usage: No Data Available
Cluster Load: No Data Available
NameNode Heap 30%
NameNode RPC 0.66 ms
NameNode CPU WIO n/a
NameNode Uptime 6.7 hr
HBase Master Heap n/a
HBase Links: No Active Master, 1 RegionServers, n/a
HBase Ave Load n/a
HBase Master Uptime n/a
ResourceManager Heap 13%
ResourceManager Uptime 6.6 hr
YARN Memory 0%
NodeManagers Live 1/1
YARN Links: ResourceManager, 1 NodeManagers
Supervisors Live 0/1
Flume Live 1/1

Fig. 1. Ambari UI



Total: 3 files or folders

Name	Size	Last Modified	Owner	Group
customer-meta-data.csv	70.7 MB	2022-10-28 01:51	maria_dev	hdfs
product-meta-data.csv	33.8 MB	2022-10-28 01:51	maria_dev	hdfs
transaction-history.csv	3.2 GB	2022-10-28 01:53	maria_dev	hdfs

Fig. 2. HDFS files

- It can also be observed that 2018 had the least amount of transaction for H&M however still out performed 2020. This was probably due to people buying more expensive clothes during this year compared to 2020 due to external factors such as covid.

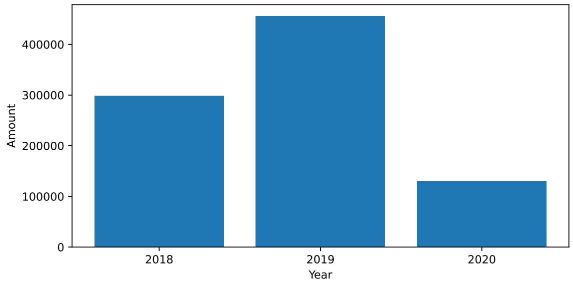


Fig. 3. Value made per year

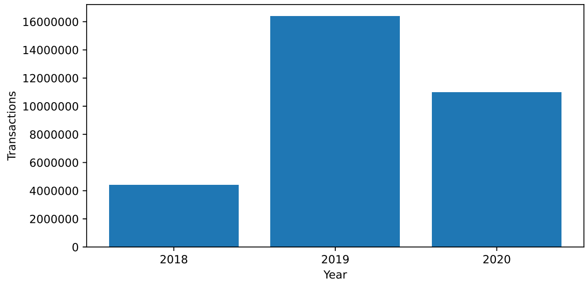


Fig. 4. Transactions made per year

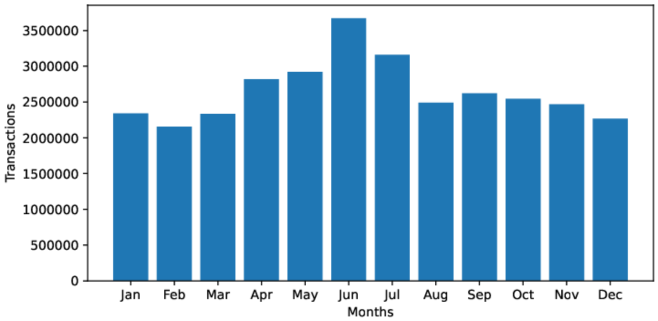


Fig. 5. Transactions made per Month

5 Conclusion

Transaction data is a good candidate for distributed processing as transaction data is generated at high velocity and volume especially in well known global retails organizations such as H&M. H&M can use such a system that uses Hadoop for distributed processing to reduce and aggregate their transaction data for insight about their customers and their market performance in a more timely manner. This processed data can then be used for analytics in order to better improve their services and relationship with customers. Using the transaction history of H&M a test case was derive to see if map reduce could be used to process the data into a more suitable form for visualization as a result it was shown that the highest transactions for H&M were in the June and July months and that out of all the years of transaction history 2018 had the smallest amount of transaction but 2020 had the least amount of value for H&M this could be due to the it being in the year when Covid was gripping the world and people were not buying expensive clothes.

References

[1] H&M South Africa website. https://www.hm.com/za/.

[2] Hadoop website. https://hadoop.apache.org/.

[3] Beakta, R., 2015. "Big data and hadoop: A review paper". *international journal of computer science information te,* **2**, 01.

[4] Subbulakshmi, T., and Manjaly, J. S., 2017. "A comparison study and performance evaluation of schedulers in hadoop yarn". In 2017 2nd International Conference on Communication and Electronics Systems (ICCES), pp. 78–83.

[5] NortonWorks. https://www.cloudera.com/products/hdp.html.