CLOUD COMPUTING PROJECTS FA18

Gregor von Laszewski
Geoffrey C. Fox

laszewski@gmail.com

Cloud Computing Projects

Gregor von Laszewski

Cloud Computing Projects

1 Hadoop, Hive and Spark multi node Cluster set up on Amazon EC2 instances, fa18-516-29

firstname: “Shilpa” lastname: “Singh”

plese see format from our example

references missing

obviously your report is an early draft. You need to expalin what you do not in person to me, but in written form in this document

you need to have shell scripts that strat the vms, you can leverage cm4 if you like or design your own shell script while for example using the AWS CLI

you need to have a mechnism on telling us how you set up the clould. Science is about reproducability, If your hadoop and spark implementation can not be deployed by others, they can not be reproduced. Thus we can not replicate your benchmark. This is simple, just write a shell script on how to set up hadoop on your 4 machines.

If this can not be done. Write a manaul on how to do this including screenshots where needed.

github:

1.1 ABSTRACT

The goal of this project is to demonstrate the steps needed to set up a 4 node Hadoop Cluster with Spark and Hive on Amazon EC2 instances from the scratch and do comparison between the traditional mapreduce and Distributed Computing through Spark. It details the Hadoop and Spark configurations required for a 4GB and 8GB(memory) node Hadoop cluster for developmental and testing purpose.The objective is to do all the installations and configurations from the scratch.This process will be the same as doing the installation and set up on any 4 unix machines which have a static IP.

This should work on arbitrary numbers of VMS

how does your performance compare to EMR

headings must not be all caps

We will compare the processing times bewteen mapreduce computation and spark computation and prove how spark is much faster than mapreduce.

how does it stack up with EMR?

We show how to integrate spark as a compute engine in hive and bypass mapreduce in hive.

Lastly, we explores a sorting technique called secondary sort where we sort values in large data files by making use of map reduce framework without bringing all the data in memory of one node in a distributed environment.

1.2 KEYWORDS

Amazon EC2, Hadoop, Spark, Hive

1.3 INTRODUCTION

The project describes what are the minimum configurations required for a multi node Hadoop cluster set-up with Spark and Hive in AWS EC2 instances and how to establish a passwordless ssh connection between the instances.This is applicable for any unix/linux instances which can have a static IP and the same steps need to be followed for establishing a passwordless ssh connection and configuring the hadoop and Spark config files.The recommended amount of memory for an instance is 8GB and atleast 20GB of physical disk space.

1.4 SOFTWARE VERSIONS

- Hadoop 2.9.1
- Hive 2.3.3
- Spark 2.3.2

1.5 ARTIFACTS

links missing

- Project Proposal
- Project Code

1.6 REFERENCES

2 Image Classification using k-means on TensorFlow fa18-523-68

Selahattin Akkas
sakkas@iu.edu
Indiana University
hid: fa18-523-68
github:
code:

Keywords: Image classification, k-means, YFCC100M, TensorFlow

2.1 Abstract

TBD

2.2 Introduction

The project goal is clustering the Yahoo Flicker Creative Commons 100 Million (YFCC 100 Million) using k-means on TensorFlow. Since data set is very large and some media has no tags, it will be hard to measure the accuracy. Yahoo also shares a subset of the dataset which has 10 tagged classes. In this work 10 class dataset will be used.

2.3 Requirements

2.4 Design

2.5 Architecture

2.6 Dataset

Yahoo Flickr Creative Commons 100 Million (YFCC100m) dataset consists ~100 million photos(99.2 million) and videos(0.8 million). Medias in the dataset carry Creative Commons license [1]. Some medias have tags but in general the data is unlabeled. Therefore, total number of class is unknown and it will make the clustering harder. There is a subset of the dataset which have 10 classes. To see accuracy performance, this subset will be used in the project.

2.7 Implementation

2.7.1 TensorFlow

TensorFlow is a Machine Learning/Deep Learning framework developed by Google. It is continuation of DisBlief which is Google's internal use framework.

TensorFlow is widely used for Machine Learning/Deep Learning applications. It is easy to develop deep learning applications on TF. After training, applications can be easily deployed and used even on mobile phones [2].

2.8 Benchmark

2.9 Conclusion

2.10 Acknowledgement

2.11 Milestones and Time Plan

2.12 10/12 - 10/18:

- Clear the data if needed
- Extract the using VGG19 (It will be done one time and the extracted features will used many times)

2.13 10/19 - 10/25:

- Tensorflow installation
- Run the built-in k-means on single node and decide the data size. (I need to get the results in reasonable time)
- Run the built-in k-means on 2-3 nodes.

2.14 10/26 - 11/08:

- Implement k-means to TensorFlow and run it on single node
- Run own implementation on multiple nodes.

2.15 11/09 - 11/26:

- Fix the problems
- Write the final paper

3 Big Data and Commercial Data Analytics fa18-523-85

Bo Li
bl15@iu.edu
Indiana University Bloomington
hid: fa18-523-85
github:

3.1 Abstract

As internet developed, online shopping has become part of our daily life. Black Friday, a traditional deal day, has also transformed as a big day for online shopping. The internet retailers, such Amazon, also developed their specific strategy for the combat in online shopping, Amazon Prime Day. In the e-commerce area, the volume of the sales and the product data increased rapidly. It is necessary to develop a cost-effective way to deal with the big data. The technique of relational data management has developed a lot in the last decades. In today's business analysis scenario, the relational technologies seem cannot hold the large data since they are designed to deal with data which is much smaller in size. The developed internet techniques allow us to collect and store the trading data, which could be the most valuable materials for researching customer behaviors.

3.2 Keywords

Big data, human behavior, data mining, Python

3.3 Introduction

3.3.1 Commercial Data

In order to extract the knowledge behind the commercial data generated by hundreds of thousands of consumers for the use of leading managers to make the

decision, it is necessary to conduct a deep analysis to the commercial data, instead of generating simple reports. The deep analysis could hardly be done by SQL since the process relies on complex models. Without those models, it is impossible to get a profound understand of the commercial data [3]. People will not only need to find out what is happening now but also need to use data to make some predictions in order to make preparations for future events. For example, if the manager is able to predict the loss of the customer in the future, they can use a discount to attract the users again.

> ***"The quality of a product or service is an important determinant of consumer satisfaction, brand performance, and long-term brand success" [4].***

In the context of big commercial data, the traditional OLAP operations are not enough anymore to meet the requirements, we also need path analysis, time series analysis, graph analysis, what-if analysis, and some complex statistical models. Time series analysis, a useful method in the commercial data analysis since we have got lots of trading historical data [5]. The managers want to get some patterns in the data in order to find some chances to improve the revenue. By the trend analysis, they can even predict some changes in advance. In the financial area, analysts are able to develop some software to conduct the time series analysis of the trading data and find some profitable trading patterns. After further verification, they can use those profitable trading patterns to conduct real trade and make profits.

> ***"Managers and researchers usually obtain measures of perceived quality from customers through surveys or interviews, which are typically based on limited samples administered periodically" [4].***

Large-scale graph and network analysis also play a key role in commercial data analysis. The virtual social network is actually a description of the links between the entities. In the network, every independent entity will be converted to a node in the total graph, and the relationship between nodes will be converted to the link. By conducting a social network analysis, we can find some useful knowledge such as a small community in the whole group. This information could be used to advertise some new product if the community meets the requirement of the target group. We can also combine the individual behavior

analysis and the group behavior analysis.

> ***"The design decisions that determine what will be measured also stem from interpretation. For example, in the case of social media data, there is a 'data cleaning' process: making decisions about what attributes and variables will be counted, and which will be ignored. This process is inherently subjective" [6].***

3.3.2 Behavioral Data

> ***"The study of consumer analytics lies at the junction of Big Data and consumer behavior. Data provide behavioral insights about consumers; marketers translate those insights into market advantage. Analytics generally refers to tools that help find hidden patterns in data" [7].***

Behavioral big data (BBD) refers to very large and rich multidimensional data sets on human and social behaviors, actions, and interactions, which have become available to companies, governments, and researchers. A growing number of researchers in social science and management fields acquire and analyze BBD for the purpose of extracting knowledge and scientific discoveries [8]. Besides, the online retailers also want to figure out the profound meanings behind the consumer's actions. So the behavioral big data comes across with the research area and industry area, which results in different research methods. We use the methods in the research area to analysis the dataset with specific models designed to the dataset from Kaggle.

3.3.3 Necessity of Big Data

Why we use big data method to conduct the analytics? The answer lies in the character of the real dataset generated in the real trading system. The system records the deal with some specific fields, which could be some information on price, time, discount, product information, product status, etc. The user behavior could also be recorded such as the action of adding to list, the time between an order to payment, etc. If a user has the habit of adding many products to the list but only buy a few of them, you may observe the data and draw the conclusion

that the user is rational and could hardly be affected by the advertisement. But if you have the huge amount of data, maybe it is real-time data, the only way to picture those consumers is establishing models and define the features, then use the big data methods to research them.

> ***"The opportunities associated with data and analysis in different organizations have helped generate significant interest in commercial data analysis, which is often referred to as the techniques, technologies, systems, practices, methodologies, and applications that analyze critical business data to help an enterprise better understand its business and market and make timely business decisions" [9].***

The user behavior could also be recorded such as the action of adding to list, the time between an order to payment, etc. If a user has the habit of adding many products to the list but only buy a few of them, you may observe the data and draw the conclusion that the user is rational and could hardly be affected by the advertisement. But if you have the huge amount of data, maybe it is real-time data, the only way to picture those consumers is establishing models and define the features, then use the big data methods to research them.

3.4 DATASET

3.4.1 Commercial Data

The basic concept of commercial data is very easy to understand. The method of analyzing sales data provides a view to having a deep understanding of the sales data, which enables the managers to make a specific plan to conduct some strategies.

> ***"Technology's changing pace requires faster market analyses than traditional market analytics can handle. BDA might provide the real-time speed necessary to meet this challenge" [10].***

In the process of analysis of the commercial data, they often analyze the historical price, and the design of product line to mine the relationships behind

different departments, which could be a still base for a better sale strategy. There are some specific requirements of the commercial data, such as it must be objective and reliable, or it may mislead the decisions of the managers.

> ***"They make decisions based on rigorous analysis at more than double the rate of lower performers. The correlation between performance and analytics-driven management has important implications to organizations, whether they are seeking growth, efficiency or competitive differentiation" [11].***

3.4.2 Dataset Description

The dataset we use here is from a retail store, it contains the information of transactions. To get a profound understand of customer behavior in different things. The first step is to conduct a descriptive analysis of the dataset to get an overview. The second step is to make a deeper study with different variables such as gender, age, city. If the owner wants to find the potential links between different product categories or specific links between a group of people and products, they can do across analysis with different variables. The dataset is an ideal simple for classification and also clustering since all the needed information of the users is provided [12].

Data columns (total 12 columns): User_ID: 537577 non-null int64, Product_ID: 537577 non-null object, Gender: 537577 non-null object, Age: 537577 non-null object, Occupation: 537577 non-null int64, City_Category: 537577 non-null object, Stay_In_Current_City_Years: 537577 non-null object, Marital_Status: 537577 non-null int64, Product_Category_1: 537577 non-null int64, Product_Category_2: 370591 non-null float64, Product_Category_3: 164278 non-null float64, Purchase: 537577 non-null int64

3.5 TOOLS

3.5.1 Python 3.7

Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python's elegant syntax and dynamic typing, together with its

interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms [13].

3.5.2 Numpy

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed. This is a tutorial explains the basics of NumPy such as its architecture and environment. It also discusses the various array functions, types of indexing, etc. An introduction to Matplotlib is also provided. All this is explained with the help of examples for better understanding [14].

3.5.3 Pandas

Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis, manipulation tool available in any language. It is already well on its way toward this goal. Pandas is well suited for many different kinds of data. Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet. Ordered and unordered (not necessarily fixed-frequency) time series data. Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels. Any other form of observational / statistical data sets. The data actually need not be labeled at all to be placed into a pandas data structure [15].

3.5.4 Matplotlib.pyplot

Matplotlib.pyplot is a collection of command style functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc. In matplotlib.pyplot various states are preserved across function calls, so that it keeps track of things like the current figure and plotting area, and the plotting functions are directed to the

current axes (please note that “axes” here and in most places in the documentation refers to the axes part of a figure and not the strict mathematical term for more than one axis) [16].

3.5.5 Seaborn

Seaborn is a Python visualization library for statistical plotting. It comes equipped with preset styles and color palettes so you can create complex, aesthetically pleasing charts with a few lines of code. It is designed to work with NumPy and pandas data structures and to support statistical tasks completed in SciPy and statsmodels. Seaborn is built on top of Python’s core visualization library matplotlib, but it’s meant to serve as a complement, not a replacement. In most cases, you will still use matplotlib for simple plotting, and you’ll need a knowledge of matplotlib to tweak Seaborn’s default plots [17].

3.6 IMPLEMENTATION

3.6.1 Data Cleaning

Data cleaning is a necessary step for the whole process of data cleaning. The result of the data cleaning will have a significant impact on the efficiency of the model and the conclusions. In the practice, data cleaning often takes some time. There is also a domain about how to clean the dataset effectively. When multiple data sources need to be integrated, e.g., in data warehouses, federated database systems or global web-based information systems, the need for data cleaning increases significantly [18].

3.6.1.1 Pre-process

There are two things need to be done in this step. It is recommended to use the database to store the data since there are lots of advantages of this. If the size of the dataset is too huge to operate in the database, we can also store the data in text and operate it in the Python. The second thing is to have an overview of the data. Checking with the original data is a good way, but the dataset is often too big to get an overview. To extract a sample of the dataset is another choice of this, which will allow you to have the better understanding of the original

dataset.

3.6.1.2 Null test

The missing value is a common problem of the data analysis, and there are a lot of ways to deal with this problem. We often do this job in four steps. Locate the range of the missing values. In order to distinguish different variables with different importance, we calculate the missing percentage of each part of the dataset. According to the result of the calculation, we use different strategies to deal with those missing values. +??? shows the result of the null test, there are two data set have null value. +??? shows the real missing part of the dataset.

```
User_ID                       False
Product_ID                    False
Gender                        False
Age                           False
Occupation                    False
City_Category                 False
Stay_In_Current_City_Years    False
Marital_Status                False
Product_Category_1            False
Product_Category_2             True
Product_Category_3             True
Purchase                      False
dtype: bool
```

```
Product_Category_2 [nan  6. 14.  2.  8. 15. 16. 11.  5.  3.  4. 12.  9. 10. 17. 13.  7. 18.]
Product_Category_3 [nan 14. 17.  5.  4. 16. 15.  8.  9. 13.  6. 12.  3. 18. 11. 10.]
```

To the part with high importance and low missing rate, we give some values to the missing part by calculating. In some cases, we also use our experience to make up the missing part. To the part with high importance and high missing rate, we will try to fix the problem by finding other data sources. In some troublesome cases, we even delete the whole part of the dataset and claim the action in the result. To the part with low importance and low missing rate, we can care other things or estimate the missing value by some simple calculations. To the part of low importance and high missing rate, we choose to delete the whole part since it can not have a significant impact on the result of the analysis.

From Figure 1, we have already replaced the missing value with zero since it is in the domain of sales.

```
Product_Category_2 [ 0  6 14  2  8 15 16 11  5  3  4 12  9 10 17 13  7 18]
Product_Category_3 [ 0 14 17  5  4 16 15  8  9 13  6 12  3 18 11 10]
```

Figure 1: Null Fix

3.6.1.3 Delete unreasonable value

It is very common to find some strange value in the dataset. Such as a very big age more than 200. These extreme values will have a negative impact on the result since the features of the dataset will be partly represented by some extreme values.

3.6.2 Data Exploration and Processing

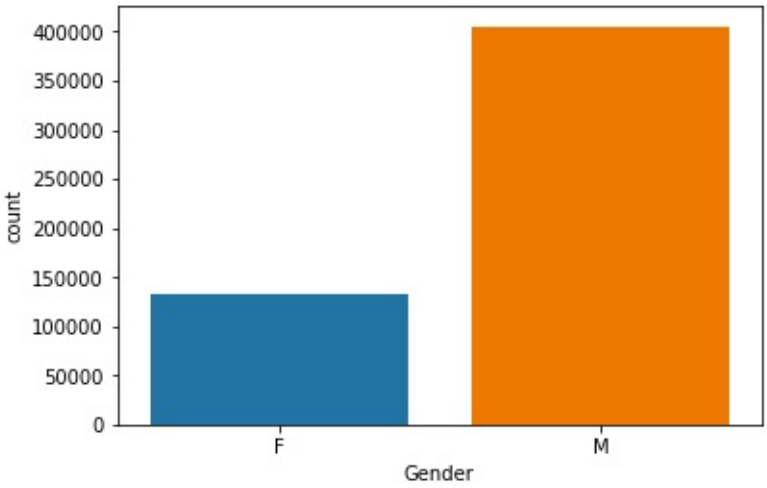
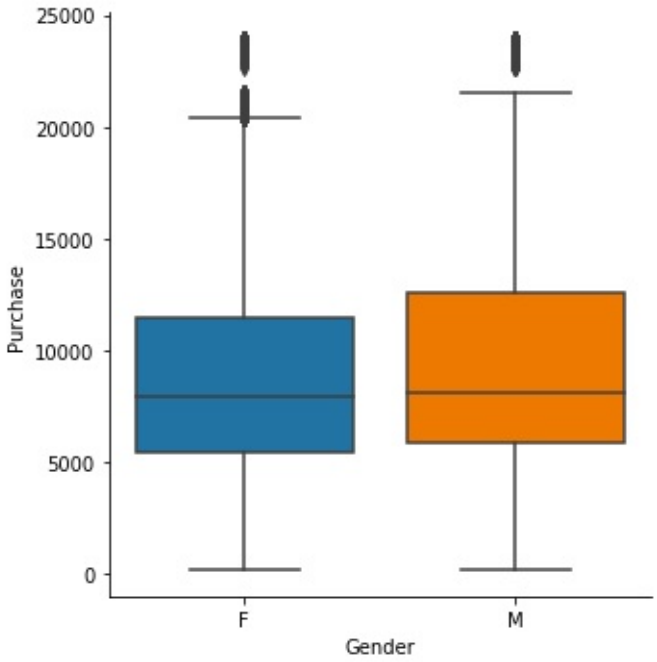
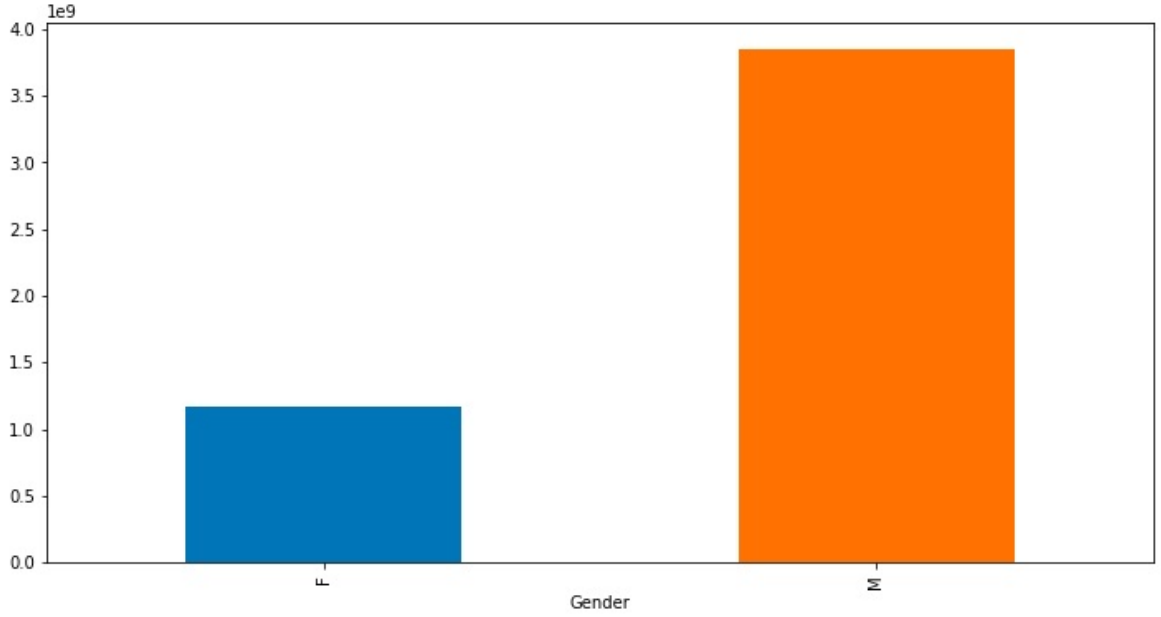


Figure 2: Gender

In the gender analysis in Figure 2, we can easily find that there are more males than females. The males have contributed more than the female. The result is also linked by the products classes since females have some specific domains to focus on. And the males also have their own preferences.



Generally speaking, females are easy to finish the act of purchase even the product is useless to them, just as the +??? shows. Purchasing was an action to meet the need of living, now it has evolved as an action of entertainment. Most of the females have the tendency of purchasing, but they have different degrees about that. Comparing with males, females are less rational in the process. They are easy to be affected by others, which is also a proof of their lack of rationality. On the other hand, females are sensitive to the details of products. Standing by this point, it is hard to make profits on females. So if the designer is able to care more about the details of the products, the products will be more attractive to customers.



The males also have contributed more than the females, about three times to the sales of females. In +???, we can easily find this. There are some reasons to explain this. Since females are easy to consume without lots of considerations they are often consuming on cheaper things such as thing for daily use.

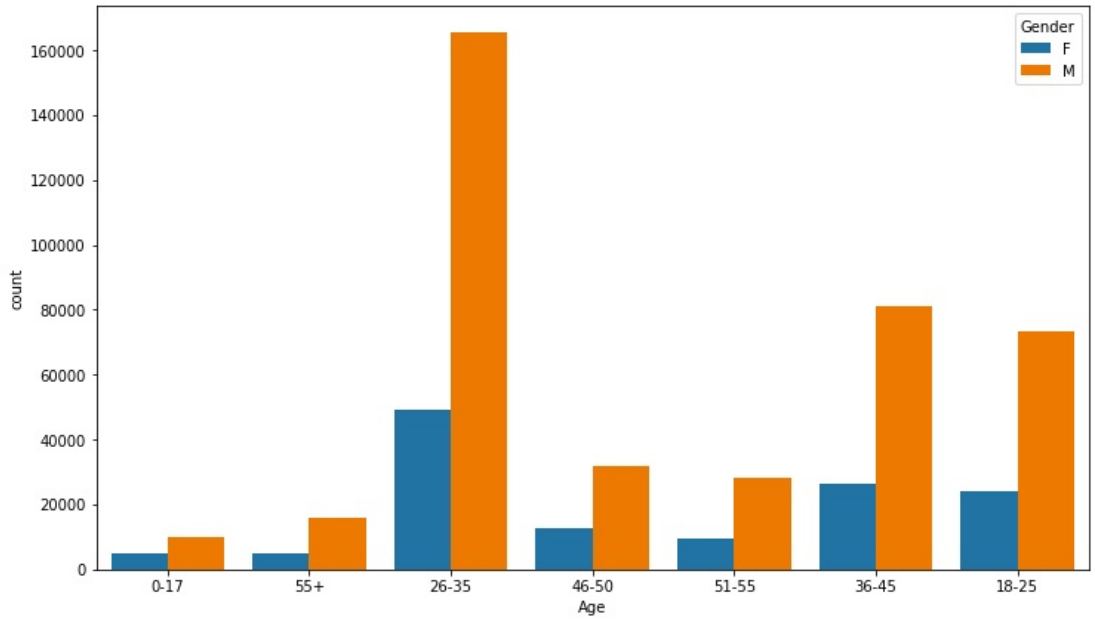


Figure 3: Age Gender

From Figure 3, it is obvious that the young people group, especially the young

males group, are the main contributors to the sales. Most of them are single or married a few years ago. So they do not have a big burden as other age groups do. They cost of education, the payment of the house, the savings for retirement, these factors are less considered by them. “Men Buy, Women Shop.” is a rule that Verde Group found in the research of men and women’s behavior of consuming.

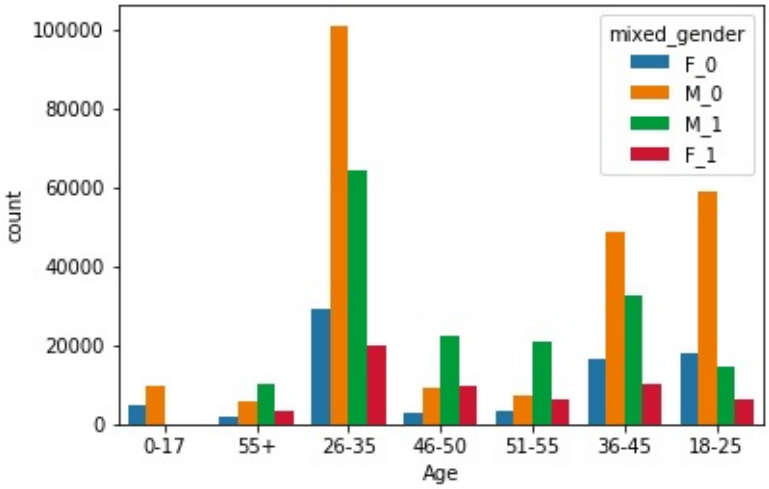


Figure 4: Age Mixedgender

Females are easy to interact with salesman since they are willing to hear some positive feedbacks. But men are more likely to consider other issues such as is there a position for his car, or is the wanted product in stock. It is well presented in Figure 4. In some domains, women invest more than men, such as time, emotions. For men, they are more likely to have a specific plan, once they achieve their goals, they are going to leave the store. Although women have a heavier burden on work, women are stilling playing a role to take care of others. The responsibility of this will help women to have a sharp awareness of shopping and a higher expectation. On the other hand, women have taken the task of shopping for the home for several generations, men could hardly find the fun in shopping.

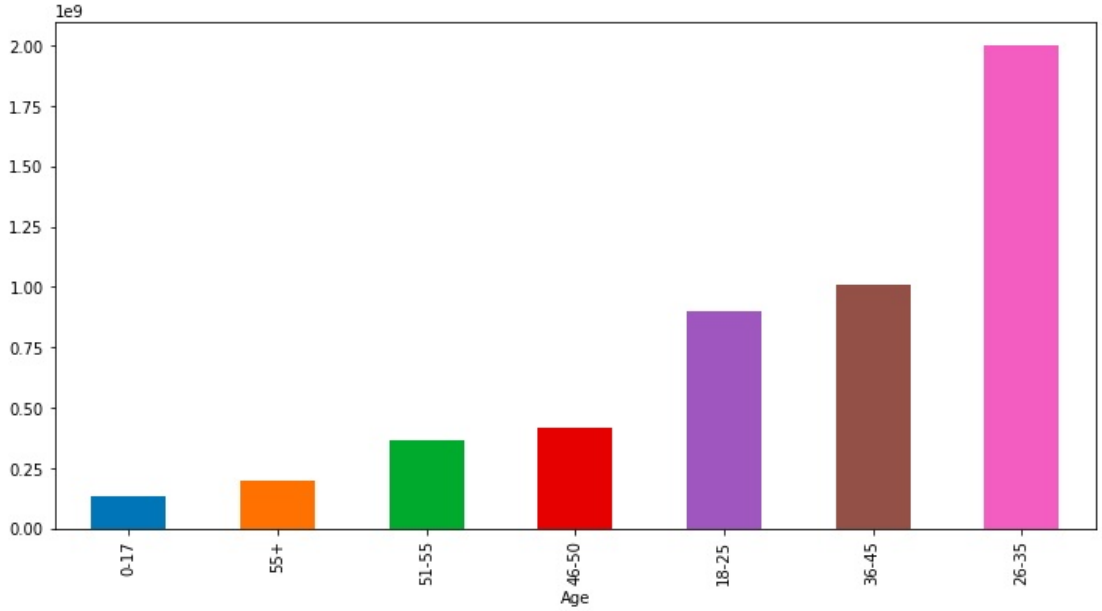


Figure 5: Age Purchage

The behavior of shopping has reflected a lot of differences between male and female. From Figure 5, we can see that Women are more likely to see shopping as part of their networking since they will get enough topics and materials about the products when talking about other women. In contrast, men are more likely to treat shopping as a task, once they have finished their task, they think they have done with this task and will not stay in the store.

3.6.3 Data Analysis and Data Visualization

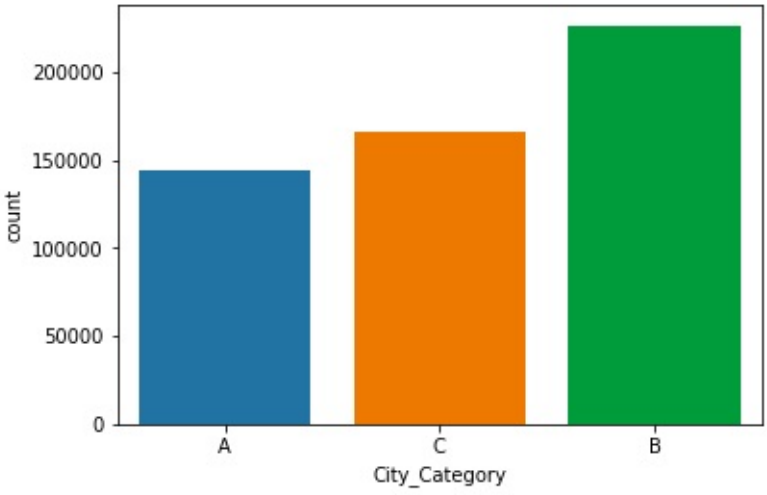


Figure 6: City

According to the Figure 6, we can see that the customers in this dataset are mostly distributed in three main cities. B city has more users than A and C. There are a lot of factors to effect this, such as the city is an industrial city or not, the city has a local big retailer or not. For the big city, people have a higher level of income, so they consume the most. For the subarea, the people usually have some reason to live far from downtown, part of the reason is the price of land and house, so they have a weaker purchase power than the group lives in the big city. Last but not least, the people who live in the countryside have the lowest purchase power.

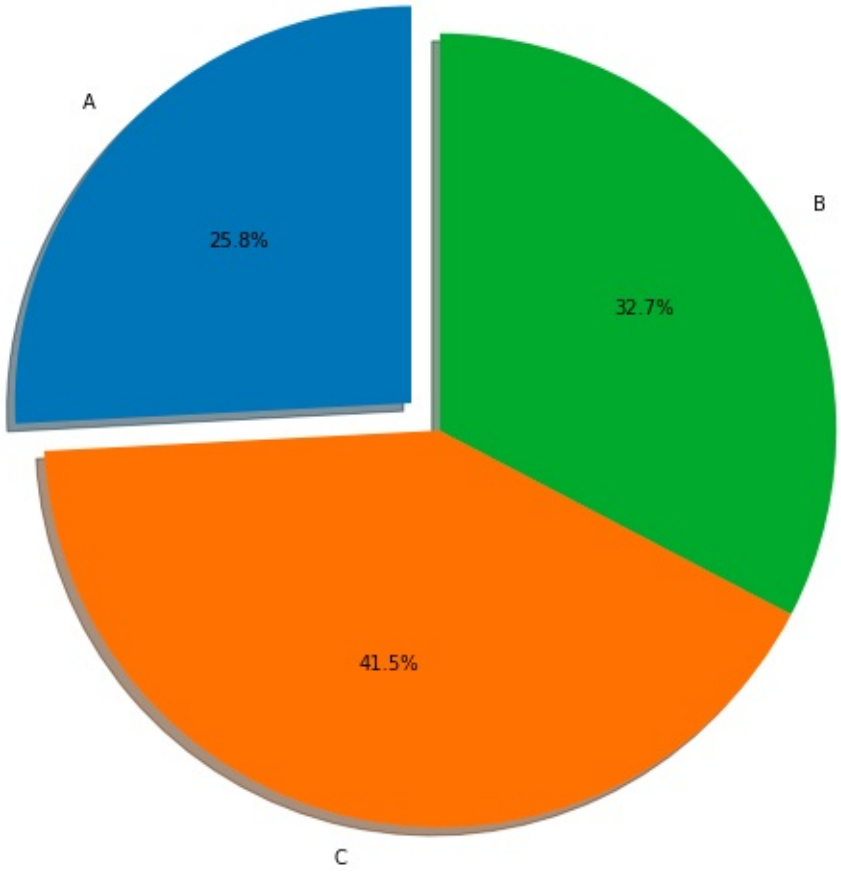


Figure 7: City Purchase

Surprisingly, in Figure 7, C city has the highest sales among the three cities. Although C city does not have the biggest customer group. This may attribute to the features of the customer group in C city. Due to the low purchase and the inconvenience they need to face when planning to go shopping, they have less time and money to invest in shopping. For the group that not live in the downtown, the time and money cost should be considered as a significant factor. In other cases, someone chooses to get the product by mail, which is also a factor that contributes to the lower purchase power of this group. Since they are not able to select and try the product by themselves, they are not willing to take the risk to pay for the mail charge. The group lives in the city is easy to get access to the product and they can get enough information form the salesman, it easy to make the deal since the service will push the deal.

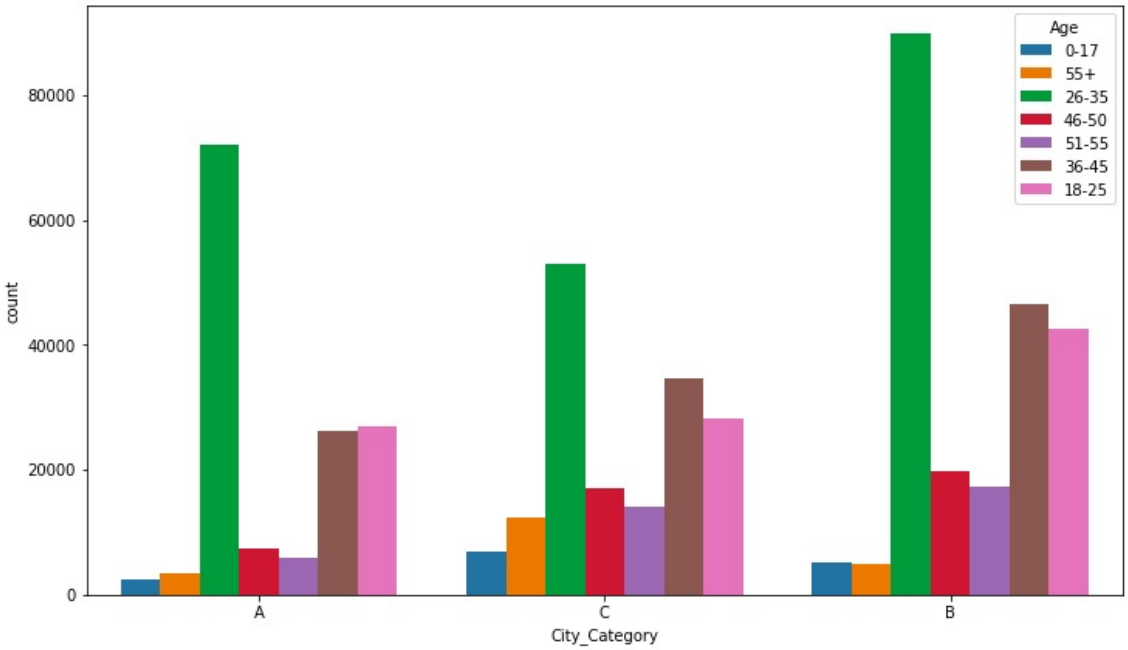


Figure 8: City Age

In each city, the young group after graduation is the main contributor to the sales. Figure 8 has showed that. As we mentioned before, they are more likely to be single, which means they have less pressure to save money for future use. There are several factors which can affect the sale converting ratio. The most important three are price, transportation, and discount. Part of the online shoppers will accept the suggestions from social media. And nearly half of them will listen to their family members or friends. Some of them are driven by the advertisements. Most of the customers are eager to see the product by themselves, especially for online shoppers. So the comments from the previous buyers will play a significant role in the process of deciding whether to buy the things or not.

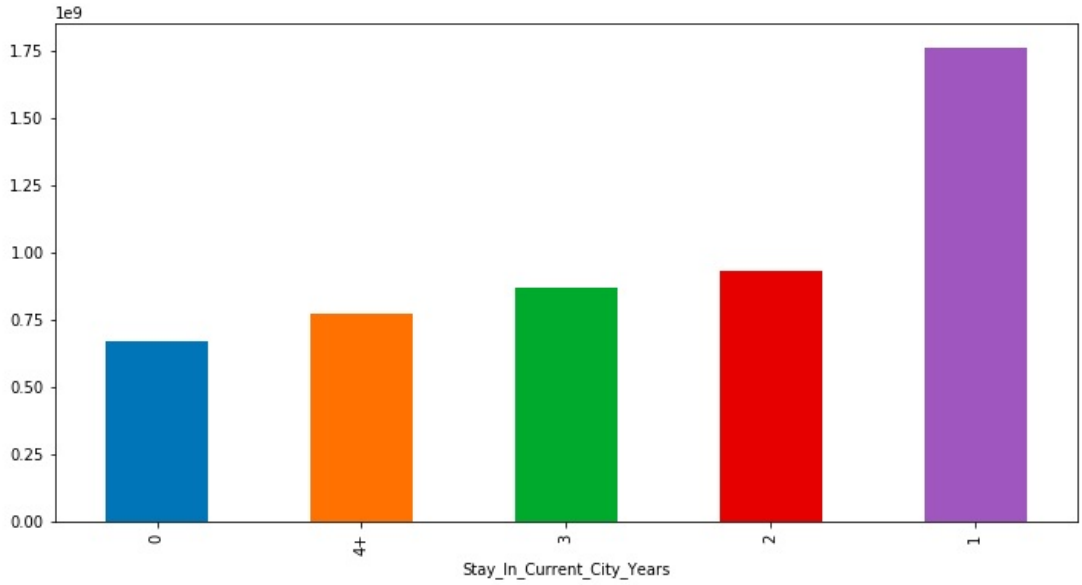


Figure 9: fa18-523-85-13-stay-purchase-bar

According to Figure 9, the people who stay in the city for one year have contributed the most part of the sales. Since they have settled down to the city, rather than those who have not to stay in the same city for one year, they have the need to buy things. On the other hand, they have adapted to the environment in the city, so they are more likely to pursue some higher level things.

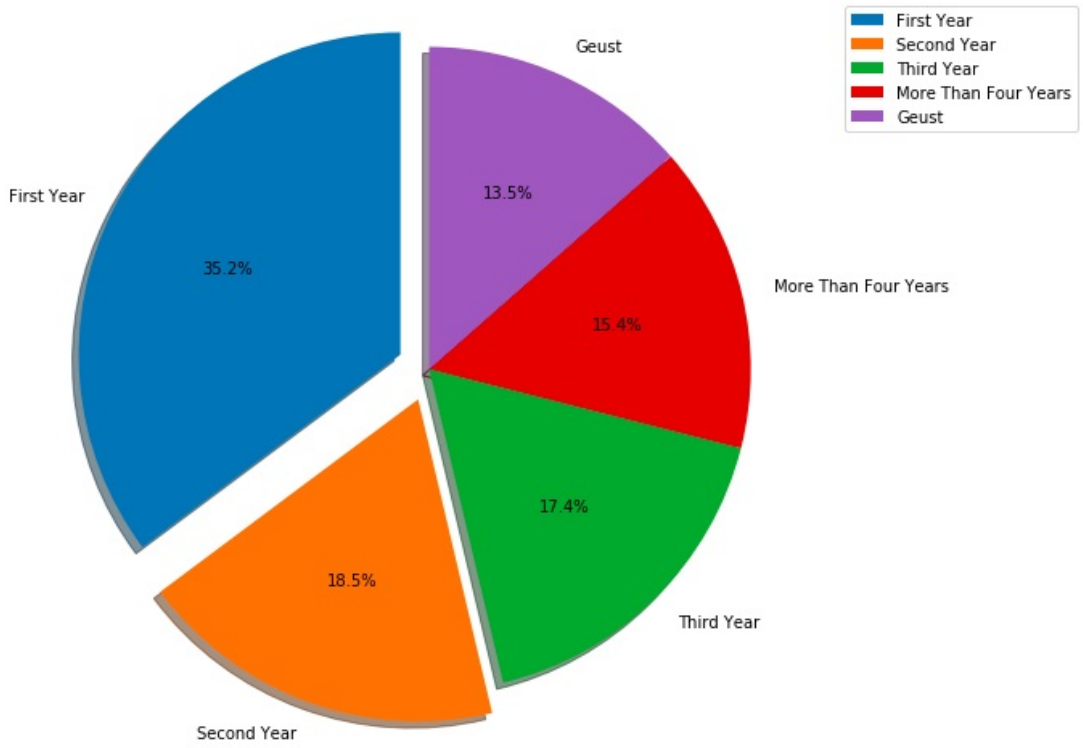


Figure 10: Stay

Figure 10 shows that the first year users in the specific city are the majority of the group. The other kinds of the group are similar in size. The females are enjoying the process of shopping, so they are tending to spend more time to compare different products from different retailers. The discount will lead to the final purchase since they are sensitive to the prices of different products. More than half of the customers will give up the purchase due to the cost of delivery, which should be a key point to improve the sales.

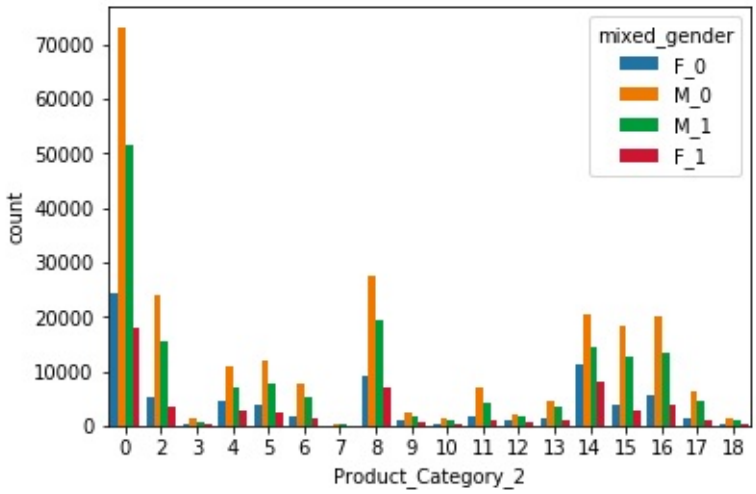


Figure 11: Product Mixed

To have a better understanding of the gender and product, we use the marital status to divide the customer and conduct across analysis with a product, just like Figure 11 shows. In each age group, it is the single male group that consumes a lot than other groups. The single female group contributes more than the married female group. A possible solution is to include the cost of delivery into the price of the product. On one hand, when customers see the price of the product at the first time, the anchor effect will leave them an impression of the price, so the additional fee of delivery will push them to give up the purchase. Showing the whole price at the first time will be a better choice. On the other hand, to serve the customers, the company can establish a stable relationship with those companies, which means a lower delivery price for every customer.

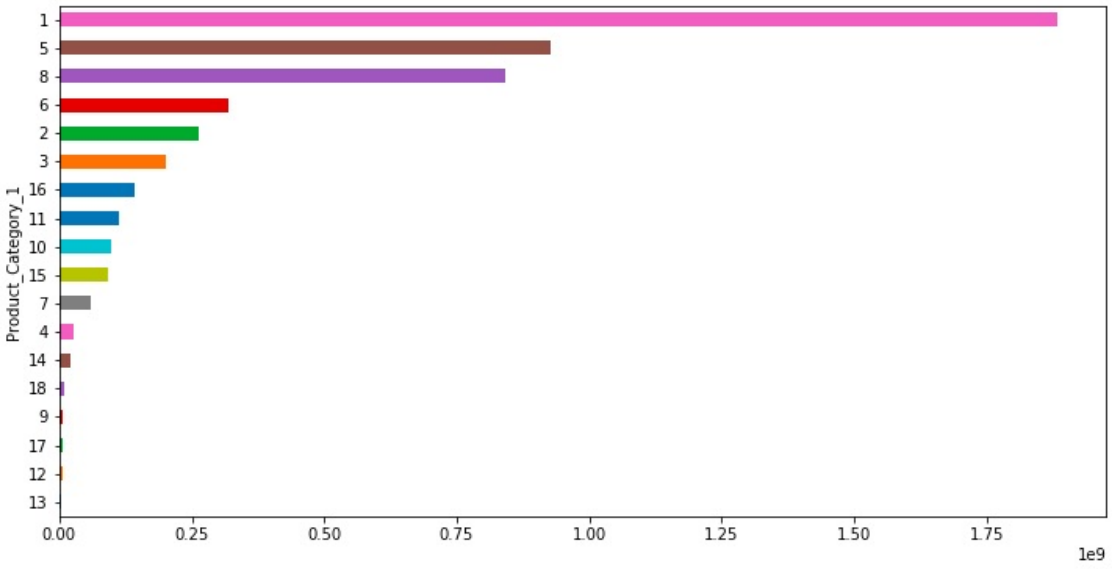


Figure 12: Product1 Bar

From Figure 12 we can see: for the first product category, although the top three items have occupied more than half of the sales, it is still more diverse than the second product category. For sales, customer behavior is a key role to care about. The process of decision making is worth to be researched. When people have the plan to consume, there are often two ways to make the decision, high indecision and low in a decision [19]. Entering is the time and money they want to invest in the purchase when the strategy has already been decided. There are a lot of factors to affect the entering way, such as the price of the product, the value of the product.

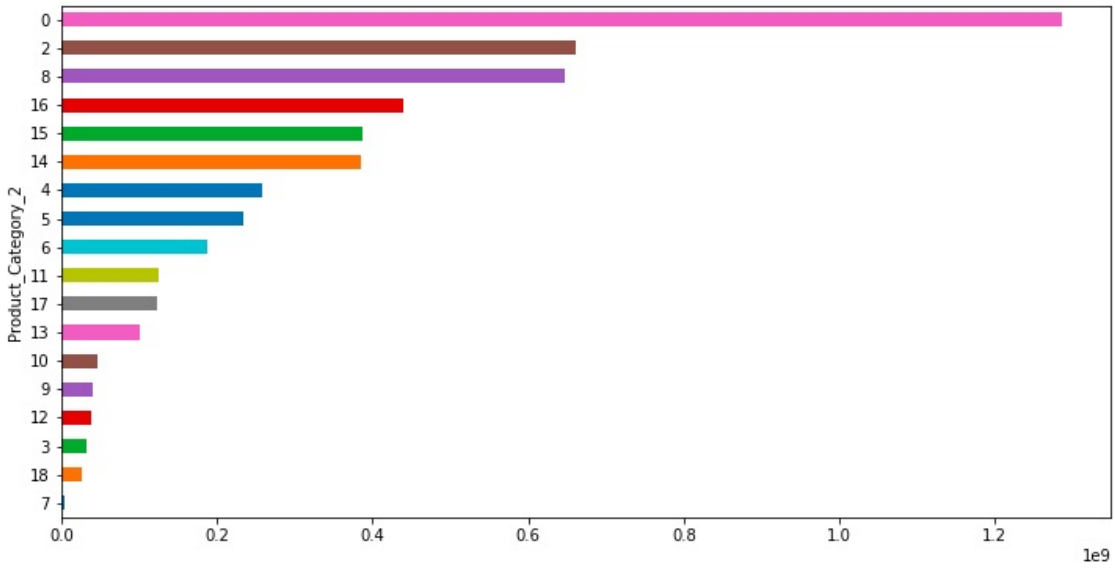


Figure 13: Product2 Bar

In the second product category like +??? shows, the top one item has occupied most of the sales, which means a single product consists the majority of the revenue. Due to the centralized profit, the economy of the city may have a greater risk than the other two cities. The product category is also one of them. Once the consumer wants to buy a gift for his father, he will spend a lot of time and money on this purchase. As a result, he has a strong tendency to make the deal, and then finish the purchase with an acceptable price. That is an example of high entering. In contrast, for the low entering consumers, they are not going to consider a lot of factors before making the decision, so the band and the preference will affect a lot [20].

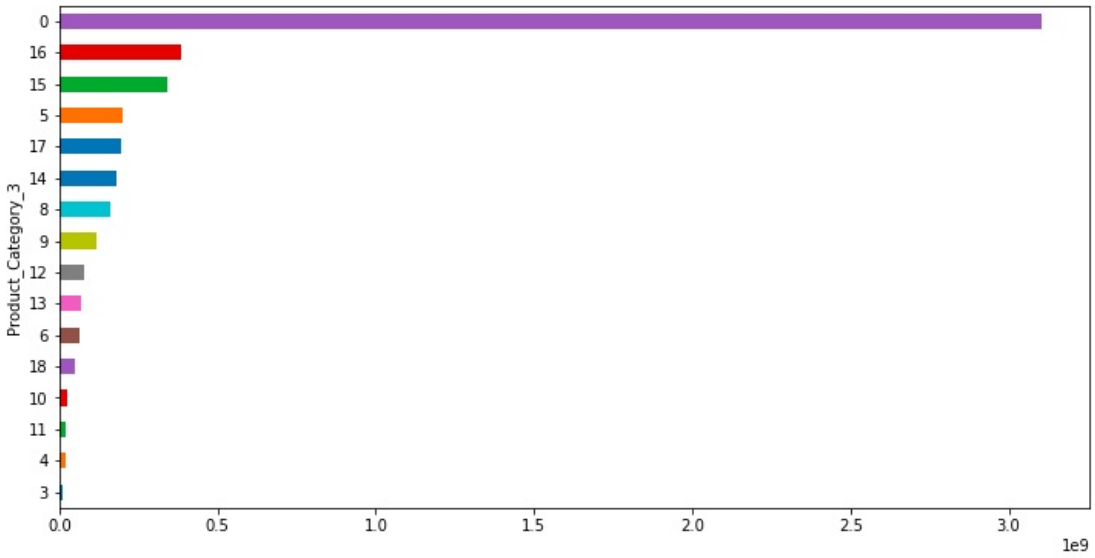


Figure 14: Product3 Bar

The third product category is similar to the first one, like Figure 14 shows, which has more diversity than the second one. Multiple products make up the whole revenue means a diversity in risks. The managers can use a hint to affect the process of making decisions. By endless advertising, they can stimulate the interest of consumers. To make this works, the managers need to find the potential needs of the consumers and convert them to the actual needs. There are two kinds of needs, psychological need and functional need when the manager is able to combine them, he will be successful since the product will meet the expectation of the consumers [21]. For the searching of the information, there are two kinds of storage environments for information. The internal environment contains the previous information, such as the product category in their mind. The external environment contains comments from other uses, differences of a price.

3.7 Conclusion

In the analysis, we have found that the big data analysis can mine a lot of information behind the dataset. By using big data technologies, we are able to establish a strategy for a business. Such asset a specific strategy for the young group of the consumers [22]. Also, we are able to make the whole target more clear, especially the blueprint of data, technology, operating. According to the

blueprint, we can break down the steps for conducting, which could lead to the practice of big data strategy.

> ***"For effective marketing, it is essential to identify a specific group of customers who share similar preferences and respond to a specific marketing signal. Customer segmentation applications can help identify different communities (segments) of customers who may share similar interests" [23].***

The blueprint of application could help the managers to recognize the value point of big data in two aspects, business model innovation, and business value chain. Based on the big data, we can combine the Internet to conduct the strategy [24]. Standing at a higher level, we could design the strategy by considering all the factors. To conduct the strategy on the business value chain, we need to analyze the applications in different scenarios. Use big data in the product design step to catch the needs of consumers, use big data in the product production process to improve the quality of the product, use big data in the sales process to send the information of the product to the targeted group.

> ***"Finally, retail managers learn very early in their management training programs that the bottom line for most decisions in retailing is sales" [25].***

The data blueprint could help us to establish the framework and specify the data collection standards. The framework of big data is supported by the classification of the data, which contains basic data of the business, behavior data and the picture data of customers [26]. To make the data source clear, we need to divide them into different kinds, internal data from the company, data from the product chain, data from the upper or lower companies and social data. If we can use the technologies in an appropriate way, big data will save us a lot of time [27]. Considering the needs of the company and the tendency, we need to compare different main technologies, from different perspectives. Hardware platform, data storage and management, calculation, data analysis, data visualization, and data safety. Comparing with their advantages and disadvantages, then make the choice of technology. To help the company establish an effective and durable management system, it is necessary to collect the data in the process stream.

3.8 ACKNOWLEDGEMENT

We would like to thank Professor Gregor von Laszweski for all his support in this project. We would also like to thank the TAs for their support to help me finish this project.

4 Topic: Big data in SAP Ariba fa18-523-86

Jeff Liu
liujeff@iu.edu
Indiana University
hid: fa18-523-86
github:

Keywords: Big Data

4.1 Abstract

We're in the midst of a big data revolution. Around the world 2.5 quintillion bytes of data is produced daily. That's a mind-boggling amount. And how about this: approximately 90% of the world's data was generated in the last two years alone. For years, companies have been collecting and analyzing massive amounts of information – everything from structured data on production, marketing, sales, HR, finance, facilities and operations to transaction-level data on suppliers, customers and partners. The explosion of data created has led organizations into a period of necessary innovation. The imperative to transform this data into intelligent insights has never been greater and no industry or line of business is immune from the big data revolution. It therefore comes as no surprise that Procurement is turning to big data to drive digital change [28]. Ariba was founded to solve the problem of inefficient enterprise procurement through the network, in 2012, SAP US acquired Ariba for $4.3 billion and was renamed SAP Ariba, right now, 2.9 million companies worldwide are connected using the SAP Ariba network platform, using Ariba's SaaS solution to manage expenses and business activities. SAP Ariba has built a complete corporate trading network.

4.2 Introduction

SAP Ariba is how companies connect to get business done. On the Ariba Network, buyers and suppliers from more than 3.3 million companies and 190 countries discover new opportunities, collaborate on transactions and grow their

relationships. Buyers can manage the entire purchasing process, while controlling spending, finding new sources of savings and building a healthy supply chain. And suppliers can connect with profitable customers and efficiently scale existing relationships – simplifying sales cycles and improving cash control along the way. The result is a dynamic, digital marketplace, where more than $1.6 trillion in commerce gets done every year[29]. Companies are increasingly relying on Ariba to enable them to connect and collaborate in today's digital economy which means demand for certified Ariba consultants is on the rise. SAP Ariba, a market leading cloud, network procurement offering is the fastest growing business for SAP. SAP Ariba is enabling friction-less commerce between companies, through the exchange of direct and indirect goods and services making SAP's vision of Run Simple, Run Real-Time and Run Networked a reality for customers, SAP Ariba's cloud-based solutions make it easier to collaborate and compete.

4.3 ARCHITECTURE

Perhaps more than any other business unit, procurement stands at the forefront of digital transformation. What began as snatching efficiency gains by digitizing and automating paper-, time-, and labor-intensive processes has evolved today into real-time analytics, enterprise-wide visibility and global supply chain optimizations. By harnessing the transformative power of digitalization, procurement's influence goes well beyond mere cost cutting and, today, profoundly impacts competitiveness and corporate strategy.

From sourcing and orders to invoice and payment, SAP Ariba's procurement solutions span the entire e-procurement process, and deliver serious bottom-line benefits:

- 60% lower operating costs
- 1% to 8% reduction in supply chain costs
- 50% to 75% faster transaction cycles
- 90% fully automated transaction processing
- 60% improvements in order accuracy
- 5% to 20% increases in revenue with new customers
- 30% or more higher share of wallet with current customers
- 15% improvement in retention rates among customers.

Automating key processes such as procurement, orders, invoicing and payment is a must if your business is to remain competitive. But to thrive, more must be done with the vast amount of data — internal and external — available to procurement. Mining, cultivating and analyzing transactional and operational data —together with data from customers, prospects, partners, and suppliers — empowers procurement executives to gain visibility into direct and indirect spending across global accounts. This visibility empowers procurement pros to identify opportunities for consolidation, cost reduction, and process improvement, and, in line with procurements' evolution to the role as trusted advisor, participate in more strategic and business-meaningful goals, such as discovering new business models, driving product and service enhancements, and identifying untapped revenue streams[fa18-523-86-www-Ariba-Procurement].

4.4 Implementation

With B2B ecommerce adoption hitting an inflection point, companies operating global supply chains are moving away from legacy systems and expensive EDI (electronic data interchange) toward ubiquitous and affordable online platforms and cloud-based solutions that reduce costs, streamline procurement and payment processes, and provide visibility across global supply chains. Leveraging advances in cloud, mobility, big-data analytics and IoT, today's digital supply chains are more data driven and operate faster, better, stronger, and leaner than ever before. According to research presented in an INFOGRAPHIC created by The Economist Intelligence Unit, the gains from digital supply chains are impressive:

- Today, 22% of manufacturing executives have "complete visibility" into their supply chains, up from 9% in 2013.
- 44% of senior executives say their supply-chain function uses sophisticated tools that leverage big data to help with day-to-day decision making.

Cloud-based supply chain solutions and business networks present great options for strategically tackling this problem by providing a robust, unified platform that enables close collaboration. According to research by SAP Ariba, collaborative supply networks deliver quantifiable business value, including:

- 60% improvement in order accuracy
- 50% - 70% faster transaction cycles
- More than 90% “touchless” transaction processes[30].

Ariba can now integrate with solutions such as SAP ERP and S/4 HANA, and can use SAP HANA and SAP Leonardo technology for data analysis to provide customers with more insights. As SAP deepens the integration of Ariba with other software, users can more easily connect, share and query their own, partner ERP, CRM, SCM and e-commerce software. Procurement systems capture a vast amount of data, including sourcing information, weather reports, manufacturing and delivery data, supplier data, purchasing data, catalogue data. There are undoubtfully very valuable insights within that data, but the challenge for procurement is to understand and use it to make better, informed decisions. Big data, predictive analytics and prescriptive guidance can further scale with cognitive computing power to provide better information that will enhance situational awareness and speed-to-decision. The resulting business value will improve the way procurement roles and processes are executed, making both more effective, valuable and sustainable[28]. The Ariba Procurement Content Solution has a single user interface and is integrated with the Ariba network, the world’s largest online trading community, to quickly implement product catalogs and their maintenance. With the right contract pricing, users simply search for the content that the supplier provides about the goods and services they need. Users will also enjoy it when they use it, because this solution is designed to make users feel like they are shopping on their favorite websites. Once your e-procurement begins to include user-friendly content, you will be able to deliver on its promises of improved normative and bottom-line balances without any difficulty.

4.5 Conclusion

Today, most procurement organizations are faced with three core challenges when working with big data: * Digitising processes: The first challenge for procurement is to access unstructured data that resides in scanned documents, email inboxes and spreadsheets. The key to unlocking the data’s potential is to ensure it is stored in a digital format that can be analyzed. * Driving insights: It is then vital to be able to analyses and utilize the digitized data. Human limitations, such as skills and time, as well as technical challenges can hinder

find patterns, process the information intelligently and accurately, and derive insights. Customers are challenged to figure out how to tap into cognitive capabilities and build them into their processes. * Enabling talent and skills: Once this is in place organizations need to foster a change in culture and mindset that will help them embrace these disruptive technologies. Big data, predictive analytics and cognitive computing require new skills and new ways of working, such as self-service analytics.

Advances in technology and new concepts like intelligent computing fueled by SAP Leonardo provide ample opportunities to predict and respond more effectively to customer and market demands. This insight can guide buyers or requesters to create request for information or a contract template based on potentially unlimited amounts of information analyzed across multiple databases. Connecting people and information guided by “intelligent” procurement systems will fundamentally change how companies buy and sell[fa18-523-86-www-Ariba-BigData]. The digitalization of business is real, and it is here to stay. The combination of new technologies and skilled talent working with intelligent systems may provide the competitive edge that businesses need to stay ahead of competitors in the market. But it will also fundamentally elevate the importance of procurement to providing guidance on innovations, mitigating risks and securing a sustainable supply chain. As with any revolution there will be winners and losers and those that make the most of big data will come out on top.

5 Secchi Disk Visibility Recognition, fa18-423-08

format not followed

use the sample and do not make it beautifu;, the sample will do formating correct

integrate refermnces

could we use master worker and not master slave

stream of data is not shown in design

arhitectur how to connect boat to cluster not shown

Yuli Zhao Indiana University Bloomington, Indiana yulizhao@iu.edu

github:

This student focusses at this time on developing programs. the results need to be clearly documented. He has not shown yet the effort he spend.

A simple program to use filters to filter for the measuing tape is missing.

5.1 Abstract

Computer vision is designed to imitate human vision. In order to imitate human vision, enabling computer to see is not enough, computer needs to be able to analyze what it sees. What computer sees is essentially frames of images which means nothing to it, but computer is able to analyze the images and extract useful information out following specific instructions from the programmer. Due to the nature of computer, it is able to perform analyzes on a large scale with specific instructions. This paper will discuss how computer vision is used in determining water turbidity through recognizing the secchi disk and extracting measurements as useful information, and how computer vision is applied to a big data set.

5.2 Introductions

Secchi disk is an eight inch circular disk has alternating black and white color on the surface. It is used to determine the turbidity or the transparency of the water. The way it works is that secchi disk is lowered into the water slowly using a tape that has measurements, and the researcher would record the measurement when the secchi disk is no longer visible. But one single measurement does not offer many information to be useful, useful information can only be extracted by researchers when there are many more measurements. But it would take researcher enormous amount of time to keep in track with each and every measurement. And computer vision is needed here to replace all that workload. Instead of having researcher record every measurement, a camera can be placed by the measurement that has visual on both the tape measurement and the secchi disk. The process of lowering secchi disk can be repeated in different locations by an automated boat, and each process will return result in the format of a video. Then computer will be able to extract the necessary information from the video and mitigate the workload for the researcher.

5.3 Data Collection

The data taken needs to be uniformed in the first place to avoid further data cleaning. For each video taken, the tape measurement needs to be placed on the right side of the video at a fixed position. In order to increase the identifiable feature of the secchi disk, four geometric shape of pentagons should be placed on the secchi disk, one on each sector. Pentagons on the white sectors are filled with color black, pentagons on the black sectors are filled with color white.

Original

5.4 Data Flow

One of the purposes of this project is to create a continuous data flow from the raw video input, to the output of whether if the secchi disk is visible and the measurements of the tape. First, the automated boat needs to equip a hard drive disk to store recorded videos. At the end of each week, the hard drive needs to be

picked up and the video storage needs to be uploaded into a cloud storage then cleared from the hard drive. To maximize efficiency, multiple machines are needed to implement a master/worker architecture. One machine will be the master, the rest of the machines are workers. The master machine retrieves raw video data from the cloud, and assigns the data to one of the worker machines. Each worker machine is programmed to break down the video into frames and perform analysis on each frame, then return two lists of results with secchi disk visibility and the measurements of the tape. When a worker machine is done with one analysis, it will tell the master machine that it is available, and the master machine will assign another raw video to the worker machine. This cycle will repeat to maximize the efficiency of analyzing the big data.

Original

5.5 Worker Machine

The first step worker machine needs to do is to take apart the raw video into separate frames. The analysis need to be performed on each individual frames. Within each frame, the worker machine needs to run two functions. One function is to correctly identify the visibility of the secchi disk. The other function is to correctly identify the position of the tape measurement, and then run an Optical Character Recognition (OCR) on the position that reads the number of measurement. The worker machine will return a list of pair data matching each frame. At last, the worker machine will let the master machine know that it has finished all its work and available to work on the next one.

Secchi Disk Visibility	Depth Measurement
True	1’5"
True	2’5"
True	3’5"
False	5’10"
False	8’10"

5.6 Approach: Use CNN to Determine the Visibility of

Secchi Disk

The position of the secchi disk constantly changes in the process due to the water turbulence. Instead of trying to locate the exact position of the secchi disk, a convolutional neural network can be used to train a model that predicts the visibility of the secchi disk in each frame. Using a convolutional neural network approach requires many labeled data to begin with in order to get an accurate model. An example of data is a frame of a video, the label would be either “yes” or “no”, representing the secchi disk is visible or not in that certain frame. In order to train an accurate convolutional neural network model, a large number of data needed to be labeled, approximately 10,000 frames or more. To ensure the accuracy of the model, the training data needs to cover most scenarios. It would take a researcher enormous amount of time to label every frame. Instead, the researcher could design a survey where it is going display a frame and a “yes” or “no” option. The researcher can recruit 100 volunteers to do 100 of these secchi disk identification questions each. Each answer along with the frame will be stored in a CSV file where a frame is stored as a one-dimensional array.

5.7 Training of Convolutional Neural Network Model

Before training the model, the label of the training data needs to be processed with one-hot encoding. To build a convolutional neural network model, Keras is a researcher friendly tool to create and test the model, the parameters should be chosen at the researcher’s discretion. A sample code is provided in the project code section.

5.8 Tape Measurement Optical Character Recognition(OCR)

The fixed position of the tape on every frame, allows the researcher to crop out the section of the frame, then run pytesseract OCR on the section. The instruction to install pytesseract can be found at www.github.com/tesseract-ocr/tesseract.

6 CMD5 Plugin to Create a Docker Swarm Cluster on 3 Raspberry PIs hid-sp18-709, hid-sp18-710

please update to markdown

see our example https://github.com/cloudmesh-community/proceedings-fa18/tree/master/project-report this report: https://github.com/cloudmesh-community/hid-sp18-709/edit/master/project-report/report.md

the original latex had structural issues so pleas ervisit the section heding levels of all sections.

Andres Castro Benavides, Uma M Kugan
Institution: Indiana University
Streetaddress: 107 S. Indiana Avenue
City: Bloomington
State: Indiana
Postcode: 43017-6221
email for Andres Castro: acastrob@iu.edu
emailfor Uma Kugan: umakugan@iu.edu

6.1 Abstract

Information technologies are evolving from mainly one-host environments to more distributed environment. Docker Swarm makes it possible to avoid having a single point of failure and instead, have multiple nodes that can be properly balanced and contain replicas of the information. Currently, Dockers must be individually downloaded, installed and configured on each physical computer in order for the desired computers to work in swarm mode. This paper details the development of a plug-in that would allow CloudMesh to deploy a Docker Swarm cluster. The creation of this plug-in would be the first step towards the development of a tool which would allow larger debian based networks to work as container oriented virtual environments with optimized usage of resources.

Keywords

6.2 INTRODUCTION

6.2.1 Docker: Swarm mode, Current Use, Installation and Configuration

Docker is the technology used for containerization for software development. It is an open source tool which makes it easy to deploy applications. Applications are packaged in containers and then it is shipped to all the platforms that is supposed to work with. Applications are divided into manageable sizes and all the dependent functions are added and individually packaged. Both Linux and Windows are supported by Docker.

Docker Swarm is a clustering and scheduling tool for Docker containers. A swarm is nothing but multiple Docker hosts which run in swarm mode and act as managers to manage delegation and workers will run swarm services. A given Docker host can be a manager or a worker or it can perform both roles. If any of the worker node becomes unavailable, manager schedules that node's tasks on other nodes. A node is an instance of the docker engine participating in the swarm [31].

A swarm is made up of multiple nodes. We need to execute "docker swarm init" to enable swarm mode and to make current machine a swarm manager, run docker swarm join on other machines to add them to the swarm as workers and run docker node ls on the manager to view the nodes in this swarm.

Docker Swarms are used to orchestrate processes, optimizing the use or resources across clusters. In other words, the use of Docker Swarms allow individual computers to work as a cluster, sharing their RAM, processors, physical memory, among other features or abilities. The docker, when used in swarm mode, evaluates the assets across the network and manages tasks in real time. Each computer can contribute its assets to complete tasks in the most efficient way. It is dynamic and adapts based on the available resources and current demands.

In order to set up a Docker Swarm, there needs to be direct access to each machine that will be used as a node which is how we will call an instance of Docker that will be part of the swarm. In order to set up the nodes, the docker must be independently installed and configured on each machine. Then, each machine must be added to the swarm, allowing it to communicate or interact with the other nodes.

This processes not only requires human resources like technicians working on installation and configuration, but also demands these actions be repeated manually on each individual node or manager. While this can be done virtually, it still requires individual attention in the setup of each machine. In order to optimize the setup of Docker Swarms, CloudMesh could be utilized to centralize installation and configuration of every node and manager.

6.2.1.1 Inside Docker

The four main internal components of docker are Docker Client and Server, Docker Images, Docker Registries, and Docker Containers.

6.2.1.2 Docker Client and Server

The docker server gets the request from the docker client and then process it accordingly. Docker server and docker client can either run on the same machine or a local docker client can be connected with a remote server running on another machine [32].

[Docker Architecture] [33]](https://github.com/cloudmesh-community/hid-sp18-709/blob/master/project-markdown/images/High-level-overview-of-Docker-architecture.png)

6.2.1.3 Docker Images

Base image are the Operating system images such as Ubuntu 14.04 LTS, or Fedora 20 which creates a container to run Operating system. The docker file contains a list of instructions to build an image. When using docker, we start with a base image, boot up, create changes and those changes are saved in layers forming another image [34].

6.2.1.4 Docker Registries

Docker images are placed in docker registries. It is same as source code repositories where images can be pushed or pulled from a single source.

6.2.1.5 Docker Containers

Docker image creates a docker container. Containers have everything for the application to run on its own.

6.2.2 Benefits of using Docker

6.2.2.1 Open Source Technology

The Docker containers are based on open standards which means that anyone can contribute to the Docker tool and at the same time customize it for their needs, if the features they are looking for is not already available.

6.2.2.2 Portability

Docker makes distributed applications to be dynamic and portable which can be run anywhere which makes it extremely popular among developers.

6.2.2.3 Sharing

Docker is integrated with a software sharing and distribution mechanism that allows for sharing and using container content which helps the tasks of both the developer and the operations team.

6.2.2.4 Elimination of Environmental Inconsistencies

Any changes made in one environment will be shared across other environments or all the applications can exist in the single environment.

6.2.2.5 Resource Isolation

Resource isolation adds to the security of running containers on a given host. Docker uses Namespaces technology to isolate work spaces called containers. Namespace is created when container is run and access is limited to that namespace only. Every container in Docker will have its own work space which makes it easier debug if there are issues with any particular container.

6.2.2.6 Easy Integration

Docker can be easily integrated into a variety of infrastructure tools like Amazon Web Services, Ansible, IBM Bluemix, Jenkins, Google Cloud Platform, Oracle Container Cloud Service, Microsoft Azure to name a few.

6.2.2.7 Better Security

Docker provides a interface for developers and IT teams to define and manage their security configurations for applications as it navigates from one stage to another.

6.2.2.8 Docker - Use Cases

The Docker platform is the only container platform to build, secure and manage the variety of applications from development to production both on premises and in the cloud. It also creates room for innovation, increases time to market, highly agile. Docker supports diverse set of applications and infrastructure for both developers and IT. It transforms IT without having to re-tool, re-code or re-vamp existing applications, policies or staff [35].

6.2.2.9 DevOps

The main goal of DevOps is to eliminate the gap between the developers and IT operations team. Docker with DevOps get the developers and operations team to work together so that they both understand the challenges faced by each other, apply DevOps practices [35].

6.2.2.10 CI/CD

Continuous Integration and Continuous Deployment are the most common use cases of Docker. Continuous Integration testing and Continuous Deployment allows developers to build codes, test them in any environment. Docker integration with Jenkins and GitHub making it easier for developers to build codes, test them in GitHub and trigger a build in Jenkins and adding the image in Docker registries [35].

6.2.2.11 Docker Containers As A Service

Docker help any organization to modernize their application architecture. It can deploy scalable services securely on a wide variety of platforms, improving flexibility and maximizing capacity. Best use case for Docker installation is the US Government where they enhanced their applications and made their components and services of their system and easily transportable/shareable with other agencies within the government [35].

6.2.3 Docker - Services:

6.2.3.1 Docker Engine

Docker Engine is the foundation for the application platform which is used for creating and running Docker containers. It is supported on Linux, Windows, Cloud and Mac OS. It is lightweight, open source and integrated with a work flow to build and containerize applications. User interface is very simple and it makes the environment easily portable from single container on single host to multiple applications on a many number of hosts [35].

6.2.3.2 Docker Enterprise

Docker Enterprise provides an integrated platform for both developers and IT operations team where container management and deployment services are together for end-to-end agile application portability. It is easy to manage, monitor and secure images both within the registry and those deployed across various clusters [35].

6.2.3.3 Docker Hub

Docker Hub functions as a hosted registry service that helps you store, manage, share and integrate images across various developer work flows. Integration testing is done each time when the image is shared [35].

6.2.3.4 Docker Compose

Docker Compose is a tool that developers deploy to define and run all multi-container Docker applications. Single host can be used to isolate multiple environments, even if they are of the same name. Data volume is copied automatically from old container whenever a new container is created. Compose uses the previous configuration to create the new container which reduces the time for replicating the same changes to the environment [35].

#@ CloudMesh

CloudMesh is an innovative tool that allows communication and interaction between cloud based solutions. Not all clouds are docker based and there are different types of virtual and cloud environments. Through CloudMesh, data can be shared and utilized by cloud solutions that are not otherwise programed to communicate with each other. Cloud mesh does not just manage a series of clouds, but centralizes and deploys them as one main system that manages the data resources.

Quote von Laszewski:

> ***Cloudmesh is a project to easily manage virtual machines and bare metal provisioned operating systems in a multicloud environment. We are also providing the ability to deploy platforms.***

6.3 Creating CloudMesh plug-ins

6.4 What it currently does and has the potential to do:

By creating CloudMesh plug-ins, it is possible to extend its potential from different kinds of cloud based environments interconnection to deployment of a container management system, in this case, Docker.

Utilizing CloudMesh to Centralize Docker Swarm Installation Cloud Mesh does not have a plug in that allows you to deploy container solutions on physical networks. Create a plug in that would allow Cloud Mesh to deploy container solutions, in this case, the Swarm mode of Docker, to a physical Debian based network, in this case, a series of raspberry pies. Could be used as a model to deploy other types of container oriented solutions. It is taking a simple network. Debian based network and allowing it to centralize resources and assigning tasks and optimizing different functions by installing a container management system, called Docker Swarm.

In order to simulate the deployment of a Docker Swarm cluster, this Cloudmesh project develops a Cloudmesh plug in, that deploys a Docker Swarm cluster on three Raspberry Pi, allowing them to be part of this multi cloud environment.

The cloud mesh allows you to use Methods to deploy the Docker Swarms as container management tools, to the raspberry pi's.

6.5 Raspberry Pi as Platform

The Raspberry Pi is a credit-card-sized computer with ARM processor that can run a Linux desktop operating system. Raspberry PI can plug into TV and a keyboard. It is a little computer which can be used for many of the things that desktop PC does, like spreadsheets, word processing, browsing the internet, playing games and also to play high-definition video. Raspberry is not intended to replace personal computer as its OS support, memory etc are limited when compared to Laptop [36].

6.5.1 Differences between Laptop and a Pi

Raspberry Pi uses an ARM based processor like ARM Cortex A7 or A53 depending upon the model while the traditional PC/Laptop uses a conventional x86 /x64 Processor from either Intel or AMD. Embedded systems had low cost and low power requirements and since ARM processor used in the Raspberry Pi is used in embedded systems, A raspberry pi consumes very much less power than a laptop. The processor is also much slower than most Intel/AMD processors used in PCs, so complex programs can not be executed. Pi does not have any wireless networking capability like WiFi, Bluetooth etc when

compared to laptop. Pi comes with 1 GB ram for version 3 while most laptops have 2GB/4GB RAM that can be easily expanded to 16GB. Laptops can have secondary storage for about 1 TB. It also supports Flash based storage which tends to be more expensive per bit than traditional Magnetic Hard drives. Therefore the Raspberry Pi will have a smaller storage capacity than a traditional PC.A binary built for either system will not execute on the other. Images or binaries that was not created by you or from true source may pose a potential threat. Docker swarm cluster can be built easily on Raspberry Pi with just two basic commands: swarm init and swarm join [37].

6.6 Docker and Big Data Platform

It is always been a challenge to maintain or even to have a control deployment environment. It is very difficult to identify any issues without proper deployment environment. Most of the times, issue can be fixed as simple as disabling a service or just uninstalling a software or slightly tweaking the environment. This can be easily achieved only when we have complete control of the environment [38]. It is very difficult to manage a distributed environment whether in cloud or not. There are lot of manual effort whenever there is an installation across multiple nodes. Docker allows anyone to quickly create, launch and test Docker containers very easily. Container offer lightweight isolation and virtualization, yielding reduced overhead, faster deployments and restarts, and simplified migration. There are lot of frameworks like, Google's Kubernetes, CoreOS, Multi-Container orchestration, etcẇhich comes in handy with Dockers and Docker is very lightweight when compared to a Virtual Machine. Even though Docker comes very handy in addressing many of these issues, main selling point is building consistent environments which are very easy to replicate. Especially in big data environment, instead of installing every single component from the Hadoop ecosystem, required for their development or testing environment, we can just create it once and use it any number of times and everywhere. Docker allows usage of different versions on the same tool for different jobs without any conflict. Docker containers are a great way of deploying services at scale and giving isolation to services that run on the same host and improving utilization and we can even use Dockers for scheduling batch analytical jobs.

6.7 Docker Critique

Docker was not designed to support the long-running containers that are needed to support production systems. While Docker gets a lot of visibility from the development and DevOps communities, its operational maturity still leaves a big void. There are no logs from containers and hence logging is difficult in a distributed Docker environment. Dockers need separate orchestration, provisioning and automation [39]. Managing a huge amount of containers is challenging, especially when it comes to clustering containers. Running a container need root access and due to security and governance policy, many companies may not grant root privileges for everyone. In some companies, only software from official/trusted sources can be installed on their machines. Since Docker is not included in Red Hat Enterprise Linux 6, it needs to be installed from docker.com, which is an untrusted source [40].

6.8 Methods: Proposed Solution

The solution was created for a specific type of hardware and software, but is modular enough to be extended to different environments with similar features, such as basic architecture -which include but is not limited to ARM single boarded computers- and an operating system based on Debian, such as Debian, Raspbian, Ubuntu, etc.

6.8.1 Hardware

For the current proposed solution, the different pieces of hardware were chosen based on criteria such as Compatibility and Price.

The following is a list of the hardware that was used and below that list there is a description of each piece of hardware that was used.

- 3 Raspberry Pi
- 3 Micro SD Cards with a capacity of 64 GB
- 3 USB to Micro USB Cables for power supply to the Raspberry Pi
- 1 External monitor for the configuration step only.

6.8.2 Raspberry Pi

For this experiment, the 3 machines that were used were Raspberry Pi 3 Model B. Raspberry Pi are single boarded computers, that come in a small presentation. They have been developed with education and extension in mind, making them very popular in the academic and entrepreneur communities. The specifications of the model that has been used for this experiment are the following:

- CPU: 1.2 GHZ quad-core ARM Cortex A53 ARMv8 Instruction Set.
- GPU: Broadcom VideoCore IV @ 400 MHz
- Memory: 1 GB LPDDR2-900 SDRAM
- USB ports: 4
- Network: 10/100 MBPS Ethernet, 802.11n Wireless LAN, Bluetooth 4.0

[41]

The Raspberry Pi are interacting with each other using a private wireless network, and they have been assigned static Internet Protocol Addresses. In this case 192.168.1.85, 192.168.1.86 and 192.168.1.87.

6.8.3 Micro SD Cards

Because of its architecture, Raspberry devices require the use of Micro SD Cards to contain the Operative system and other files. They emulate the Hard drive resource used on other kinds of computers. The reason that it is required to have at least 16 GB of memory, is because there will be several pieces of software installed in the devices, each one of them with different requirements:

Docker Memory Requirements [31]:

- 8GB of RAM for manager nodes or nodes running DTR.
- 4GB of RAM for worker nodes.
- 3GB of free disk space.

So at least 12 of the GB would be required for Docker and 4 GB used for the proper functioning of Raspbian. [42]

Taking these requirements in consideration, there should be a minimum of 16GB of free space in the MicroSD in order to perform this experiment.

The Micro SD cards used were San Disc Memory Cards with a 64GB capacity.

6.8.4 Micro USB Cables

3 USB to Micro USB Cables for power supply to the Raspberry Pi Since these small computers don't use the regular power supply chords, they are equipped with MicroUSB ports to power the device. All of these devices are plugged to a main power outlet that allows to charge multiple devices at the same time. There are other options to power the devices include, such as attaching them to external batteries.

6.8.5 External monitor

Since the Raspberry Pi are headless machines, they require to be accessed directly for the initial set up and after that it is possible to continue the configuration and installation process using any kind of remote access, like SSH or RealVNC. For this initial connection, any kind of screen that is HDMI compatible is useful. In this case the initial setup of the Raspberry Pi was performed on a Toshiba 55 inch HDTV with HDMI port. After that they were accessed from a Laptop computer with Linux Ubuntu 17.10, using Remmina via ssh using XORG.

6.8.6 Initial input devices

In order to set up the devices. The Raspberry Pi will require a set of initial input devices attached to each computer. For this exercise, a USB enabled standard keyboard and a USB enabled standard mouse were used.

6.8.7 Software

6.8.7.1 Raspbian

Currently, the default way to deploy the operating system to the Raspberry Pi is by using an Operating System installation Manager called Noobs -which stands

for “New Out Of Box Software”-. This manager can be downloaded directly from the Raspberry Pi website and it includes several Operating system options, among them:

- Raspbian
- Pidora
- LibreELEC
- OSMC
- RISC OS
- Arch Linux

Since Raspbian is the default Operating system and most commonly used, this experiment decided to use it. This is also helpful because there is material available in different websites with instructions on how to install Docker in Debian based Machines. Raspbian is Debian based. Another important reason is that Docker has as a requirement that the Linux kernel version on which it will be installed is 3.10 or higher. The Kernel version of the version of Raspbian that was used is 4.9.

The version of Raspbian that was used has the following specifications:

- Kernel version: 4.9
- Release date: 2017-11-16

6.8.8 Docker

There are several versions of Docker available. Each version with their own advantages and disadvantages. Because of the architecture used by Raspberry Pi -ARM instead of AMD-, the Docker version used is ***Docker for Debian ARM***. With the following Specifications:

- Version 17.09.0-ce
- Release 2017-09-26

This version of Docker is Community Edition, which means that it is available for free and can be installed on bare metal or cloud infrastructure. This flexibility is good for the experiment, because it will be installed on Raspberry Pi, which are considered physical devices or bare metal Machines [31].

6.8.8.1 Prerequisites

There are several reasons to have the pre requisites that the user will find in this document. They will be explained in a separate section. Before using the proposed solution, the user's environment needs to meet the following requirements:

6.8.9 Raspbian Installed

Raspbian must be installed and configured on all Micro SD Cards. For this, the user may download Noobs from https://www.raspberrypi.org/ and copy it to a formated Micro SD Card. Once the Raspberry Pi has the MicroSD loaded with noobs in place and has the input devices and display attached to it, the user may follow the OS installation guide found on: >http://raspbian.org/

It is advisable to be hooked up to the network where the user is planning on implement this solution before running Noobs for the first time. This will allow the user to download newer packages or Raspbian and avoid interruptions in the process.

This requirement exists because there is a function that is being explored to capture Raspberry Pi images to be deployed later on and avoid the present pre requisite, but it is not ready yet.

6.8.10 Update OS repositories

In order to ensure that the user is accessing the latest version available of the software, it is important to update the Raspbian repositories. In this case, the user can access the Terminal and enter the following commands:

```
sudo apt-get update
```

to update the list of available repositories and then

```
sudo apt-get upgrade
```

to upgrade the available packages. The first time that the user runs one of these commands, the root password will have to be entered. This process might take a few minutes [43].

6.8.11 Remote access setup

Enable SSH on the Raspberry Pi. After Raspbian installation, enable SSH on all your Raspberry Pi machines.

To do this, the user has to add a line in the file `sshd\_config` found in the directory `/etc/ssh/` The line has to go at the end of in the `Authentication section`. It has to contain the following string:

> ***`PermitRootLogin yes`. [44]***

6.8.12 Changing hostnames

In order to keep the three Raspberry Pi organized it is highly advisable to assign an exclusive and distinctive hostname to each Raspberry Pi. The three Raspberry Pi have the following static IP addresses:

```
1.  pi85 - 192.168.1.85
2.  pi86 - 192.168.1.86
3.  pi87 - 192.168.1.87
```

By default, all Raspberry Pi devices will have the same Host Name.

To change this feature on each machine, the user will have to modify the line that contains `127.0.1.1` and as hostname it includes the string `raspberrypi` in `/etc/hosts` file, in most of the cases it is the last line in the file. Then, the user may type the desired hostname instead of the word raspberrypi and save the file and close it. This part can be done by using the text editor that comes by default with Raspbian, an editor called `nano`. It is not advisable for the users to modify the rest of the entries, at least as part of this project.

Once the file is modified, the user will have to initialize the hostname with the `hostname.sh` script this can be done using the following line in the Terminal:

> ***`sudo /etc/init.d/hostname.sh`***

To check if the modification has worked as expected, the user may check the hostname of the machine from the Terminal by running the command: `hostname -I`

6.8.13 Steps Followed

6.8.13.1 Testing shell commands prior to integrations with Cloudmesh

Since Raspberry pi is not currently listed under the supported operative systems for Docker or Cloudmesh, The process of deploying Docker and configuring the swarm Mode was successfully tested on the Raspberry Pi first using the commands that are intended for Debian. Once the Swarm was configured, the three Raspberry Pi devices were left on for over 24 hours and it was not observed any kind of abnormal behavior, like looping services in the OS or overheating.

6.8.13.2 Purchasing the hardware

The different hardware components were purchased via Amazon.com and took anywhere between 2 to 5 days to arrive. The different components can also be purchased through multiple on line sources or local electronics stores.

6.8.13.3 Installing the components via ssh into every node.

The following steps were followed on each device: Usig the TV as an external monitor, the USB input devices: keyboard and mouse, and the Raspberry Pi with Raspbian installed. An ssh key was generated and the device was accessed using Remmina via a XORG connection from a computer equipped with Linux Ubuntu 17.10 Artful Aardvark.

The components were installed in the following order:

Updated the Raspbian packages Installed Python 3.6.2 and Python 2.7.13 via PIP and also Installed Cloudmesh: following the instructions found in: https://github.com/cloudmesh/ Installed Docker CE ARM via Terminal using the following command:

```
curl -sSL https://get.docker.com \| sh=\" as suggested in https://www.raspberrypi.org/
```

Started the swarm and assigned a master node

```
sudo docker swarm init --advertise-addr 192.168.x.x
```

Created the remaining two nodes, with

```
curl -sSL https://get.docker.com \| sh=\" as suggested in https://www.raspberrypi.org/
```

And then, using the docker swarm join command the token generated when the master node was created, they were added to the swarm.

6.9 Installing and configuring Docker Swarm

6.9.1 Manager

Since Docker requires at least one computer to be a Manager and Cloudmesh also requires at least one main configured piece of equipment, a Raspberry Pi was chosen to be the main device, in this case, the Raspberry Pi with the IP address 192.168.1.85. The following command was run on the Terminal or that device to set it as the manager:

```
sudo docker swarm init --advertise-addr 192.168.1.85
```

6.10 Workers

The other two Raspberry Pi devices. In this case, the Raspberry Pi with the IP address 192.168.1.86 and the one with 192.168.1.86, were defined as simple worker nodes.

To define the workers, the following command was used:

```
sudo usermod -a -G docker USER`
```

and to work as part of the swarm the command used was:

```
docker swarm join --token *** 198.168.1.85:2377`
```

As a last step, it was confirmed that all the nodes were added by using the following command:

```
sudo docker node ls
```

6.11 Additional Research

6.11.1 Other functions considered

Initially, for this case, it was considered an option to developed a function called CaptureImage and a second function called Deploy Raspbian. As their names suggest, the first one intended to capture an image or backup of a Raspberry Pi. This first function would receive the IP address or hostname of the desired machine and the desired location to store the captured image, alongside the corresponding credentials and wrap a dd shell command similar to the following:

```
dd if=/dev/mmcblk0 bs=1M ` gzip -QUOTE \| dd of=imageDir\|
```

Among the challenges faced, this line was returning an invalid syntax, most likely because of the use of the variables. Since there was not a lot of time, the team decided to postpone this function.

The second function was called DeployRaspbian and would receive the route and name where the image would be deployed, i.e.`/dev/bkp` and image name and route, i.e.`~/Desktop/raspbian.gz`. The shell command that would be wrapped would be:

```
gzip -dc diskNm PIPE sudo dd of=imageName bs=1m conv=noerror,sync
```

More information on this topic can be found in the section called Backup ***www.raspberrypi.org***.

Among the challenges, there is no clarity on whether the image can be deployed over a lan connection and this point there is not enough time to run tests. Also, since this is a copy of a previously used Raspbian, there is a chance that there might be conflicts related to the IP addresses that might be stored in different files of the OS.

6.11.2 Final code

The final version of the code can be found on:

https://github.com/cloudmesh-community/hid-sp18-709/tree/master/project-code

6.12 Other options considered

Other options of coding were considered during the development of this solution. Since all of the deployment can successfully be done via terminal in

Raspbian, two main options were considered:

Option 1. A bash script for every part of the deployment and wrap it in python. This option would have been less dynamic and wouldn't make the best use of the available resources, but at the same time it could have ben easier to addapt to linux Operating systems other than Raspbian.

6.13 Conclusions

It is possible to create the plug in. Using the SH sub process included in python 2.5-3.5. The team was able to try the steps one at a time at the level of py scripts, but encountered an error previously mentioned in this document when trying to implement it as part of cms. Also, as the professor suggested, this same system can be implemented as a different abstraction for deployments such as an abc class similar to the following:

```
class deployment

  def prepare
      prepares installation including downloads and other installs needed

  def deploy
      deploys the package or software

  def configure
      does some configurations

  def test(test)
      does a test wheer a name is passed of a test (you could have multiple)
      the name all woudl be running all tests
```

Since most of this work was working with bash commands tunneled through python scripts and implemented in CMS, Once this is fully functional, it is very possible that the same methodology can be followed to add more layers of complexity, i.e. Kubernetes.

It would be important to consider that the fact that the passwords would have to be either hard coded or transferred in plain text has to be seen as a vulnerability, that has to be addressed either by adding an encryption/decryption module or finding another way to safely access the root of the target device.

6.14 Work Breakdown

- ***The authors would like to thank Dr. Gregor von Laszewski for his support and suggestions on this project.***
- Uma Kugan 50% of the document and codeing.
- Andres Castro Benavides 50% of the document and testing and reviewing the code
- Gregor von Laszewski significant help with markdown, suggestions to the report, structure, correction of some issue,
  contribution of all the cm5 module on which this project relies, transition to markdown

[1] S. Kalkowski, C. Schulze, A. Dengel, and D. Borth, "Real-time analysis and visualization of the yfcc100m dataset," in ***Proceedings of the 2015 workshop on community-organized multimodal mining: Opportunities for novel solutions***, 2015, pp. 25–30 [Online]. Available: http://doi.acm.org/10.1145/2814815.2814820

[2] TensorFlow, "TensorFlow lite." Oct-2018 [Online]. Available: https://www.tensorflow.org/lite

[3] S. Debortoli, O. Müller, and J. vom Brocke, "Comparing business intelligence and big data skills," ***Business & Information Systems Engineering***, vol. 6, no. 5, pp. 289–300, 2014.

[4] S. Tirunillai and G. J. Tellis, "Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation," ***Journal of Marketing Research***, vol. 51, no. 4, pp. 463–479, 2014.

[5] G. George, M. R. Haas, and A. Pentland, "Big data and management." Academy of Management Briarcliff Manor, NY, 2014.

[6] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," ***Information, communication & society***, vol. 15, no. 5, pp. 662–679, 2012.

[7] S. Erevelles, N. Fukawa, and L. Swayne, "Big data consumer analytics and the transformation of marketing," ***Journal of Business Research***, vol. 69, no. 2, pp. 897–904, 2016.

[8] G. Shmueli, “Research dilemmas with behavioral big data,” ***Big data***, vol. 5, no. 2, pp. 98–119, 2017.

[9] H. Chen, R. H. Chiang, and V. C. Storey, “Business intelligence and analytics: From big data to big impact,” ***MIS quarterly***, pp. 1165–1188, 2012.

[10] Z. Xu, G. L. Frankwick, and E. Ramirez, “Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective,” ***Journal of Business Research***, vol. 69, no. 5, pp. 1562–1566, 2016.

[11] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, “Big data, analytics and the path from insights to value,” ***MIT sloan management review***, vol. 52, no. 2, p. 21, 2011.

[12] M. Dagdoug, “Black friday.” online, 2018 [Online]. Available: https://www.kaggle.com/mehdidag/black-friday/home

[13] Python, “Python.” online, 2018 [Online]. Available: https://docs.python.org/3/tutorial/

[14] Numpy, “Numpy.” online, 2016 [Online]. Available: https://www.tutorialspoint.com/numpy

[15] Pandas, “Pandas.” online, 2015 [Online]. Available: https://pandas.pydata.org/pandas-docs/stable/index.html

[16] Pyplot, “Pyplot.” online, 2017 [Online]. Available: https://matplotlib.org/users/pyplot_tutorial.html

[17] Seaborn, “Seaborn.” online, 2018 [Online]. Available: https://community.modeanalytics.com/python/libraries/seaborn/

[18] E. Rahm and H. H. Do, “Data cleaning: Problems and current approaches,” ***IEEE Data Eng. Bull.***, vol. 23, no. 4, pp. 3–13, 2000.

[19] Y. Liu, “Big data and predictive business analytics,” ***The Journal of Business Forecasting***, vol. 33, no. 4, p. 40, 2014.

[20] J. F. Hair, M. Celsi, D. J. Ortinau, and R. P. Bush, ***Essentials of marketing research***. McGraw-Hill/Higher Education New York, NY, 2008.

[21] P. Malik, "Governing big data: Principles and practices," ***IBM Journal of Research and Development***, vol. 57, no. 3/4, pp. 1–1, 2013.

[22] D. J. Power, "Using 'big data'for analytics and decision support," ***Journal of Decision Systems***, vol. 23, no. 2, pp. 222–228, 2014.

[23] S. Fan, R. Y. Lau, and J. L. Zhao, "Demystifying big data analytics for business intelligence through the lens of marketing mix," ***Big Data Research***, vol. 2, no. 1, pp. 28–32, 2015.

[24] C. Donnelly, G. Simmons, G. Armstrong, and A. Fearne, "Digital loyalty card 'big data'and small business marketing: Formal versus informal or complementary?" ***International Small Business Journal***, vol. 33, no. 4, pp. 422–442, 2015.

[25] L. W. Turley and R. E. Milliman, "Atmospheric effects on shopping behavior: A review of the experimental evidence," ***Journal of business research***, vol. 49, no. 2, pp. 193–211, 2000.

[26] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," ***Information Sciences***, vol. 275, pp. 314–347, 2014.

[27] T. H. Davenport and J. Dyché, "Big data in big companies," ***International Institute for Analytics***, vol. 3, 2013.

[28] Marcell Vollmer, "Big data – what's the big deal for procurement?" Web page, Sep-2017 [Online]. Available: https://www.cbronline.com/big-data/big-data-whats-big-deal-procurement/

[29] Ariba Inc., "SAP ariba live: The journey continues." Web page, Mar-2018 [Online]. Available: https://www.ariba.com/about/news-and-press/sap-ariba-live-the-journey-continues

[30] Baskar Radhakrishnan, "SAP ariba — procurement and supply chains for the digital world." Web page, Mar-2017 [Online]. Available:

https://us.nttdata.com/en/blog/2017/march/sap-ariba-procurement-and-supply-chains-for-the-digital-world

[31] Docker, “Docker ce release notes,” ***Docker Documentation***. San Francisco, CA, Dec-2017 [Online]. Available: https://docs.docker.com/release-notes/docker-ce/

[32] J. Turnbull, ***The docker book: Containerization is the new virtualization***. New York, USA: James Turnbull, 2014.

[33] F. Paraiso, S. Challita, Y. Al-Dhuraibi, and P. Merle, “Model-driven management of docker containers,” in ***9th ieee international conference on cloud computing (cloud)***, 2016, pp. 718–725 [Online]. Available: https://www.researchgate.net/figure/High-level-overview-of-Docker-architecture_fig1_308050257

[34] B. B. Rad, H. J. Bhatti, and M. Ahmadi, “An introduction to docker and analysis of its performance,” ***International Journal of Computer Science and Network Security (IJCSNS)***, vol. 17, no. 3, p. 228, 2017 [Online]. Available: http://paper.ijcsns.org/07_book/201703/20170327.pdf

[35] Hackernoon, “Docker-the popular containerization technology for an effective software development.” 2017 [Online]. Available: https://hackernoon.com/docker-the-popular-containerization-technology-for-an-effective-software-development-4e2cddc5a329

[36] R. Pi and R. Pi-Teach, “FAQS,” ***Raspberry Pi***. Cambridge, UK [Online]. Available: https://www.raspberrypi.org/help/faqs/

[37] A. Ellis, “5 things about docker on raspberry pi.” Sep-2016 [Online]. Available: https://blog.alexellis.io/5-things-docker-rpi/

[38] V. Murugesan, “Why we chose docker to build our data processing platform.” 2015 [Online]. Available: http://bigdata-madesimple.com/why-we-chose-docker-to-build-our-data-processing-platform/

[39] S. Charrington, “Running hadoop on docker, in production and at scale.” 2015 [Online]. Available: https://thenewstack.io/running-hadoop-docker-production-scale/

[40] P. Hauer, “Discussing docker. Pros and cons,” ***Philipp Hauer’s Blog***. Oct-2015 [Online]. Available: https://blog.philipphauer.de/discussing-docker-pros-and-cons/

[41] B. Benchoff, “Introducing the raspberry pi 3,” ***Hackaday***. Feb-2016 [Online]. Available: https://hackaday.com/2016/02/28/introducing-the-raspberry-pi-3/

[42] R. Pi and R. Pi-Teach, “SD cards,” ***SD cards - Raspberry Pi Documentation***. Cambridge, UK [Online]. Available: https://www.raspberrypi.org/documentation/installation/sd-cards.md

[43] Debian, “DebianPackageManagement,” ***DebianPackageManagement - Debian Wiki***. 2017 [Online]. Available: https://wiki.debian.org/DebianPackageManagement

[44] L. Bailey, L. Novich, T. Hildred, and D. Jorm, “Red hat customer portal,” ***5.2.2. Enable root login over SSH***. 2012 [Online]. Available: https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/6/html/v2v_guide/preparation_before_the_p2v_migr enable_root_login_over_ssh