School of Electrical and Information Engineering

University of the Witwatersrand, Johannesburg

ELEN7046 - Software Technologies and Techniques

Individual Project Report

Big Data and Data Visualisation  
using Commodity Hardware  
and Open Source Software

Solution: Twit-Con-Pro

July 4th, 2016

Author: Gareth Stephenson (778919)

**Group Members:**

|  |  |
| --- | --- |
| Matsobane Khwinana | 779053 |
| Sidwell Mokhemisa | 1229756 |
| Dave Cloete | 1573016 |
| Kyle Trehaeven | 0602877N |

https://github.com/garethstephenson/ELEN7046 Group2 2016

# Abstract

The hardware and software used to construct a cost-effective cluster computing framework, which is capable of processing big data workloads, is presented. The system is designed to processes social media data related to American presidential and South African political party municipal elections for use by a data visualisation framework. An overview of the hardware infrastructure based on Raspberry Pi 3 Model B’s, as well as the software solution created using Apache Spark and Scala, is discussed. In conclusion, the system, which is the combination of economically priced hardware and open source software, along with targeted custom software applications, demonstrates itself to be a viable big data solution for small businesses, start-ups and academia.

Table of Contents

[Abstract i](#_Toc455324770)

[List of Figures iii](#_Toc455324771)

[List of Tables iii](#_Toc455324772)

[1. Introduction 1](#_Toc455324773)

[2. Background 1](#_Toc455324774)

[3. Requirements 1](#_Toc455324775)

[3.1 Functional 1](#_Toc455324779)

[3.2 Non-functional 1](#_Toc455324780)

[4. Approach 2](#_Toc455324781)

[4.1 Infrastructure Hardware 2](#_Toc455324783)

[4.1.1 Raspberry Pi 3 Model B 2](#_Toc455324784)

[4.1.2 TP-LINK 10/100Mbps Desktop Switches 2](#_Toc455324785)

[4.1.3 EZCOOL 450W Power Supply Unit 2](#_Toc455324786)

[4.2 Infrastructure Software 4](#_Toc455324787)

[4.2.1 Raspbian Linux 4](#_Toc455324788)

[4.2.2 Apache Spark 4](#_Toc455324789)

[5. Design Overview of the Data Transformation Applications 5](#_Toc455324790)

[5.1 Apache Spark Cluster Computing Components 5](#_Toc455324792)

[5.2 Category Count per Hour Application 6](#_Toc455324793)

[5.3 Category Count per Day Application 9](#_Toc455324794)

[6. Challenges 10](#_Toc455324795)

[7. Recommendations 10](#_Toc455324796)

[8. Conclusion 10](#_Toc455324797)

[References 11](#_Toc455324798)

[Appendix A Timesheet 12](#_Toc455324799)

[Appendix B Environment Configuration 13](#_Toc455324800)

[Appendix C Cost of Hardware 14](#_Toc455324801)

# List of Figures

[Figure 1: Raspberry Pi 3 infrastructure 2](#_Toc455324750)

[Figure 2: Modified ATX PSU with switches 3](#_Toc455324751)

[Figure 3: Raspberry Pi PSU 3](#_Toc455324752)

[Figure 4: TP-LINK PSU 3](#_Toc455324753)

[Figure 5: ATX PSU ratings 3](#_Toc455324754)

[Figure 6: Apache Spark's cluster computing components 5](#_Toc455324755)

[Figure 7: CategoryCountPerHour class 6](#_Toc455324756)

[Figure 8: Tweet data sample in JSON format 7](#_Toc455324757)

[Figure 9: Tweet class in Scala with accompanying HashTag class 7](#_Toc455324758)

[Figure 10: DAG of CategoryCountPerHour program flow 8](#_Toc455324759)

[Figure 11: CategoryCountPerHour JSON output sample 8](#_Toc455324760)

[Figure 12: Input and output classes for processing CategoryCount JSON data 8](#_Toc455324761)

[Figure 13: CategoryCountPerDay class 9](#_Toc455324762)

[Figure 14: CategoryCountPerDay JSON output sample 9](#_Toc455324763)

# List of Tables

[Table 1: RAM use per system 4](#_Toc455324764)

[Table 2: RAM footprints per system 5](#_Toc455324765)

[Table 3: Time to upload data sources 5](#_Toc455324766)

# Introduction

Data, in the age of modern computing, is being produced at an alarming, almost incomprehensible rate, and making sense of it all is the primary objective of the Big Data movement [1]. Seeking meaningful information from Big Data, whether by human analysis or machine learning algorithms, will reveal never-before seen insights into not only what is happening all around us, but why.

This report focuses on the data transformation process, from building the infrastructure and installing and configuring the cluster-computing framework, to writing the software that creates the output to be used by the data visualisation component of the solution.

# Background

The objective of the project was to find a data source that matched the required field of interest selected by the group, to process that data into a form of information which would provide meaningful insight into that data, and to finally present that processed information via a visualisation framework so further examination could occur.

The field of interest selected by the group related to elections happening in both the United States of America and South Africa. Attention was given primarily to the individuals campaigning within the USA, namely Donald Trump and Hillary Clinton. For the South African elections, the areas of focus were the parties, namely the ANC, DA and the EFF. These identified subjects were used as the categories within the data transformation process.

Twitter was chosen as the data source for the group project, as it was felt this better portrayed the sentiment of the people, not just online and traditional media outlets. Also, the Twitter API proved easy to integrate with.

# Requirements

The requirements of the data transformation process, both functional and non-functional, relating to hardware and software are listed below.



## Functional

* Transform tweet data into the sum of categories per hour for any given day
* Transform the output from the above step into the sum of categories per day
* Transform the output from the above step into the total sum for each category
* Transform tweet data into the sum of positive and negative sentiment for each category for any given day
* Transform tweet data into a sum of commonly found words

## Non-functional

* Build a viable big data processing solution using low cost commodity hardware and open source software
* Purchase, configure and network a number of Raspberry Pi 3 Model B devices
* Create a single power supply to run the Raspberry Pi’s and network switches
* Install and configure cluster-computing framework software capable of on running on Raspberry Pi 3’s
* Learn Scala in order to create the software to fulfil the functional requirements of the project

# Approach

The data transformation process of the project was unique due to the dual outcomes expected of this particular endeavour. Not only did the data transformation process need to transform tweet data into meaningful information for the visualisation framework to consume, it needed to perform the data transformations on a platform that was affordable and powerful enough for it to be considered as a viable choice for small businesses, start-ups and academia.

The approach taken to build the hardware infrastructure was very exploratory in nature, as none of the members of the group had ever undertaken such a task before. A pragmatic approach to creating the data transformation software was employed, where each successive step leveraged the knowledge gained from the one before it.



## Infrastructure Hardware

### Raspberry Pi 3 Model B

Five Raspberry Pi 3 Model B units were selected to run the cluster computing framework infrastructure, cumulatively bringing 20 CPU cores and 5GB of RAM into the pool of available resources. Each Raspberry Pi 3 was equipped with a 32GB MicroSD card, class 10 specification with a minimum speed rating of UHS-1. The master node was equipped with a UHS-3 MicroSD card for better performance on read operations, as it was also acting as the file server.

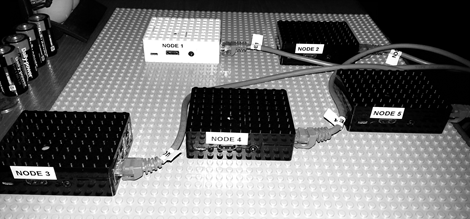


Figure 1: Raspberry Pi 3 infrastructure

### TP-LINK 10/100Mbps Desktop Switches

Two 5-port, 10/100Mbps TP-LINK desktop switches (TL-SF1005D) were selected for connecting the Raspberry Pi’s on a local area network (LAN). The speed of the desktop switches matched the speed of the Raspberry Pi’s on-board LAN module, so there was no need to invest in more expensive gigabit capable units.

### EZCOOL 450W Power Supply Unit

A PC ATX 450W power supply unit (PSU) (JSP-450P08N) was used as a power source to drive the Raspberry Pi’s as well as the switches. All the devices are rated to run between 5V and 5.1V, with the Raspberry Pi’s requiring a maximum of 2.5A and the switches requiring a maximum of 0.6A each. The ATX PSU is rated as being able to deliver up to 40A at 5V, so it has more than enough capacity to power the entire infrastructure. The PSU was opened up and had all the 3.3V and 12V leads removed, leaving only the 5V leads along with the ground leads. The connectors for PC devices and the PC motherboard were removed and discarded, and 5V micro-USB male B type cables were soldered on to the 5V leads. These would be used to power the Raspberry Pi’s. The 1.4mm DC male power plugs from the original PSUs of the switches were removed and soldered onto available 5V leads on the ATX PSU.

This configuration was agreed upon by the group as sufficient to serve as the base for the data transformation infrastructure hardware.



Figure 2: Modified ATX PSU with switches

|  |  |
| --- | --- |
| Figure 3: Raspberry Pi PSU | Figure 4: TP-LINK PSU |

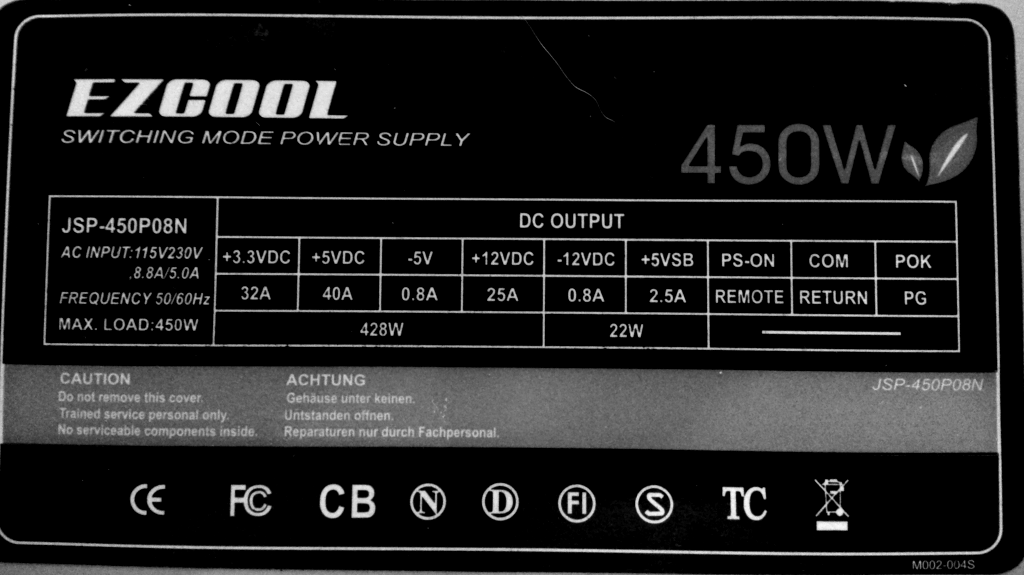


Figure 5: ATX PSU ratings

## Infrastructure Software

The Raspberry Pi environment is limited in system resources, specifically RAM. It only has 1GB of RAM which is partitioned for use between the system (accessible by the user) and the GPU.

The Raspberry Pi’s were configured with the least amount of RAM for the GPU (16MB), leaving the remaining space available for user applications. No windowing system (like the X Window System) was to be used, so any larger allocation of RAM for GPU use was unnecessary.

### Raspbian Linux

The Raspbian Linux operating system (OS), a derivate of Debian Linux, was installed on the Raspberry Pi’s as it has been tailored to best utilise the hardware’s resources and features. Further configuration was done on the OS to disable the daemons responsible for controlling the on-board Bluetooth and Wi-Fi modules, thus freeing up even more available RAM.

### Apache Spark

When the project was initiated, Hadoop was suggested as the cluster computing framework to use for the data transformation portion. It has a long, successful track record and there are numerous tutorials online on how to install it on Raspberry Pi’s. However, Hadoop’s original design is based on a distributed file system that is very disk intensive, a feature that does not work well with the Raspberry Pi’s. An alternative solution was offered, namely Apache Spark. It was promoted as being faster than Hadoop by 10 to 100 times, between disk and memory based operations respectively. It also advertised that is could work with a variety of file systems, local disk based, networked and distributed (partitioned) file schemes, including Hadoop’s.

Both Hadoop and Apache Spark were installed on the Raspberry Pi infrastructure to see which solution would suit the needs of the project best.

Linux’s Network File System (NFS) service was installed and enabled on the Raspberry Pi master node so the slaves in the Apache Spark configuration could operate on a single source of data, with the least amount of memory overhead.

When using Hadoop’s Distributed File System (HDFS) with Hadoop, the RAM used by the slave nodes increased from 99MB to 125MB, and from 181MB to 206MB on the master node. In addition to extra RAM use, the time to upload the source data (1GB of plain text files) to HDFS took around 8 minutes, as the data was partitioned across all the data nodes within the cluster. The same source data took only 2 minutes to copy to the Apache Spark master node when using Linux’s scp utility.

Table 1: RAM use per system

|  |  |  |
| --- | --- | --- |
| System | RAM used (MB) | RAM used (%) |
| Raspbian OS | 35 | 3.6 |
| Apache Spark (master) | 170 | 17.47 |
| Apache Spark (slave) | 115 | 11.82 |
| Hadoop (master) – Empty HDFS | 181 | 18.6 |
| Hadoop (slave) – Empty HDFS | 99 | 10.17 |
| Hadoop (master) – Populated HDFS | 206 | 21.17 |
| Hadoop (slave) – Populated HDFS | 125 | 12.85 |

When comparing the amount of RAM available for data transformation operations in table 2 and time to load data into the network in table 3, Apache Spark was the primary candidate due to its RAM utilisation and speed of file access.

Table 2: RAM footprints per system

|  |  |
| --- | --- |
| System | RAM footprint (%) |
| Apache Spark (master) | 21.06 |
| Hadoop (master) – Populated HDFS | 24.77 |
| Apache Spark (slave) | 15.42 |
| Hadoop (slave) – Populated HDFS | 16.44 |

Table 3: Time to upload data sources

|  |  |
| --- | --- |
| Data destination | Time to upload |
| Linux NFS | 01:53 |
| Hadoop HDFS | 08:02 |

# Design Overview of the Data Transformation Applications

The data transformation applications were written in Scala, the language in which Apache Spark was created. Apache Spark comes with a rich Scala API, allowing for easy use of the Resilient Distributed Dataset (RDD), the fundamental building block used in parallel programming operations in Apache Spark.

A brief overview of Apache Spark’s cluster computing components will be discussed so a better understanding of how parallel programming within the context of the framework is managed and executed.



## Apache Spark Cluster Computing Components

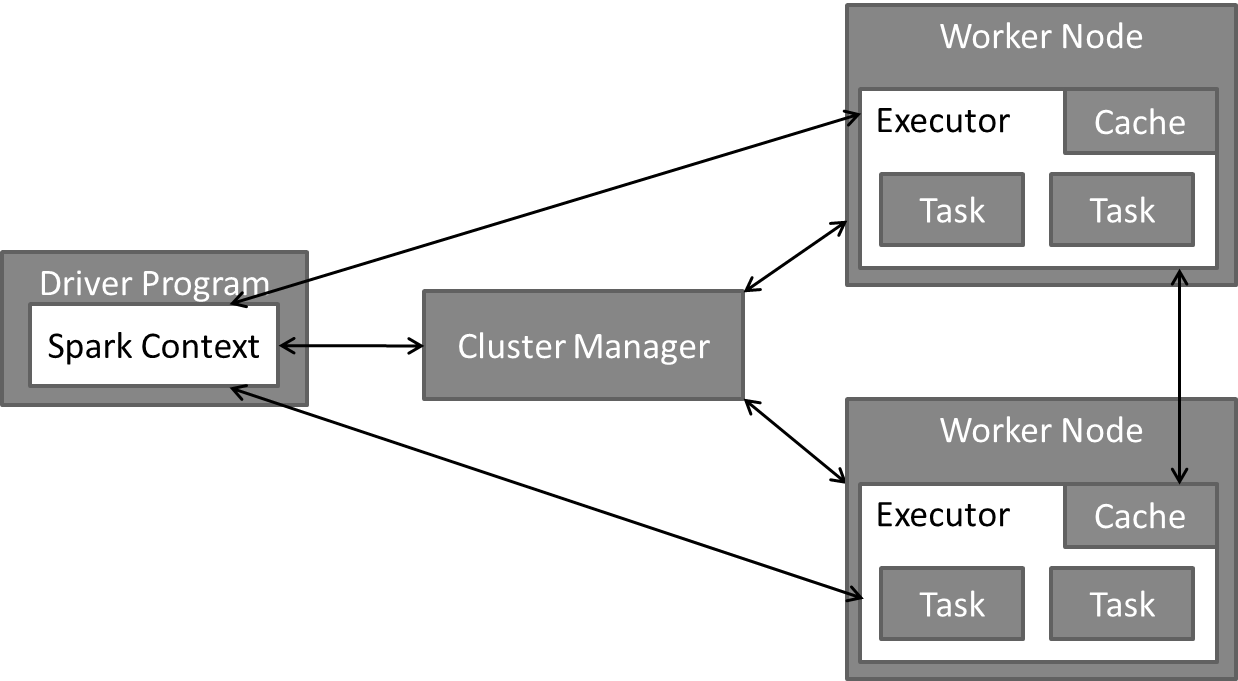


Figure 6: Apache Spark's cluster computing components

The applications created to perform map-reduce operations within Apache Spark need to run using an instance of a *SparkContext*. These are called *Driver* Programs in Apache Spark.

The *SparkContext* then connects to a *Cluster Manager* which allocates resources across nodes for the application. In the context of this project, the *Cluster Manager* is Apache Spark’s standalone cluster manager.

The *SparkContext* then acquires *Executor*s on *Worker Node*s in the cluster, which will run computations and store data for the applications hosting the *SparkContext*.

The applications are then sent through to the *Executors*, followed by the *Task*s that the *Executor*s need to execute.

*Executor*s can communicate with each other using peer-to-peer networking (like bit torrent), in order to transmit shared data (broadcast variables).

Once the *Executor*s have completed their *Task*s, they communicate back to the *Driver* Program.

The data transformation operations consist of two main applications: *CategoryCountPerHour* and CategoryCountPerDay. The logical progression of data flow and transformation of each application is explained in detail below.

## Category Count per Hour Application

The data transformation operation steps consist of:

* Loading tweet data from plain text files into *Tweet* objects
* Performing various map, filter and reduce operations on those *Tweet* objects related to the categories and timings mentioned in sections 2 and 3.1
* Outputting a JavaScript Object Notation (JSON) text file for use within the visualisation framework

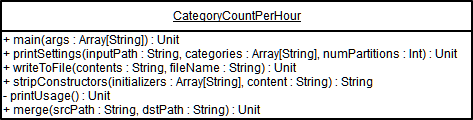


Figure 7: CategoryCountPerHour class

Apache Spark executes applications via a standard *main* method which accepts command-line parameters. When submitting the application to Apache Spark, the parameters are supplied as a space separated list.

To execute the *CategoryCountPerHour* application, the following command needs to be issued:

spark-1.6.1-bin-hadoop2.6/bin/spark-submit --class org.TwitConPro.CategoryCountPerHour TwitConPro-assembly-1.0.jar /data/20160610 Trump,Clinton 16

The command is broken down as follows:

* spark-submit is the execution engine, which requires a class to run. In this case, *org.TwitConPro.CategoryCountPerHour* is supplied. This in turn calls the *main* method described above
* The second parameter, TwitConPro-assembly-1.0.jar, is the Scala jar file that contains the *org.TwitConPro.CategoryCountPerHour* class
* The third parameter is the path to the tweet data files to be transformed
* The fourth parameter is a comma separated list of keywords to use as categories during the transformation operation
* The fifth parameter is an optional parameter used to instruct Apache Spark on how many cores/partitions to use to process the data

A sample of the tweet data is presented in figure 8:

|  |
| --- |
| {  "createdBy": "Particle News",  "createdAt": ISODate("2016-06-10T03:02:05Z"),  "coords": ["latitude", 37.3541079, "longitude", 37.3541079],  "favouriteCount": 0,  "hashtags": [],  "twitterID": NumberLong("741103109980618753"),  "inReplyToName": "",  "inReplyToStatusID": NumberLong(-1),  "inReplyToUserID": NumberLong(-1),  "isRetweet": false,  "language": "English",  "place": "",  "sensitive": false,  "quotedStatusID": NumberLong(-1),  "retweeted": false,  "retweetedCount": 0,  "tweetText": "A former US ambassador to the Middle East pointed out the inherent flaw in the Trump... https://t.co/75IUDa9rIn https://t.co/ixChZTEPy8",  "tweetURL": "https://twitter.com/jess247news/status/741103109980618753"  } |

Figure 8: Tweet data sample in JSON format

This data is mapped into *Tweet* object instances presented in figure 9:

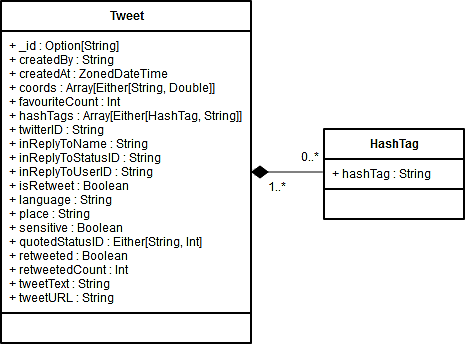


Figure 9: Tweet class in Scala with accompanying HashTag class

The mapped *Tweet* objects are then sorted by date, filtered by hour per day and category, then reduced by category into the sum of unique categories per hour.

The application flow is visually represented via the directed acyclic graph (DAG) in figure 10:

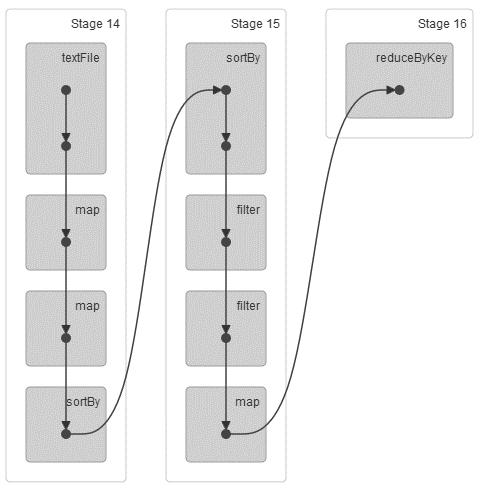


Figure 10: DAG of CategoryCountPerHour program flow

A sample of the output from the *CategoryCountPerHour* application is presented in figure 11:

|  |
| --- |
| [{  "Date": "2016-06-10T03:00:00Z",  "Data": [  { "Category": "Trump", "Count": 9947 },  { "Category": "Clinton", "Count": 6125 }  ]  }, {  "Date": "2016-06-10T04:00:00Z",  "Data": [  { "Category": "Trump", "Count": 11512 },  { "Category": "Clinton", "Count": 7130 }  ]  }] |

Figure 11: CategoryCountPerHour JSON output sample

Just as the text data of tweets were mapped to a Scala *Tweet* class, the converse also applies when producing the JSON text data above. The following Scala classes were used to produce the JSON output:

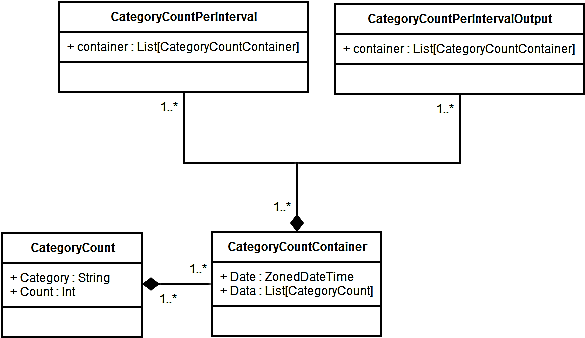


Figure 12: Input and output classes for processing CategoryCount JSON data

The top-level JSON array in figure 11 is represented by the *CategoryCountPerIntervalOutput* class in figure 12. The JSON objects inside that array are represented by the *List[CategoryCountContainer]* class, which is an array of objects based on the *List[CategoryCount]* class in the *Data* attribute.

## Category Count per Day Application

The data transformation operation steps consist of:

* Loading the JSON output of the *CategoryCounterPerHour* application into a *CategoryCountPerInterval* object (see figure 12)
* Flattening the hourly based data into a given day by summing up the totals per category mentioned in sections 2 and 3.1
* Outputting a JavaScript Object Notation (JSON) text file for use within the visualisation framework

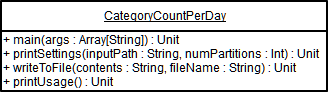


Figure 13: CategoryCountPerDay class

The output presented in figure 11 is used as the input for the *CategoryCountPerDay* application, and is supplied as a parameter to the *main* method.

To execute the *CategoryCountPerDay* application, the following command needs to be issued:

spark-1.6.1-bin-hadoop2.6/bin/spark-submit --class org.TwitConPro.CategoryCountPerDay TwitConPro-assembly-1.0.jar /data/hourlydata.json 16

The command is broken down as follows:

* spark-submit is the execution engine, which requires a class to run. In this case, *org.TwitConPro.CategoryCountPerDay* is supplied. This in turn calls the *main* method described above
* The second parameter, TwitConPro-assembly-1.0.jar, is the Scala jar file that contains the *org.TwitConPro.CategoryCountPerDay* class
* The third parameter is the path to the output file to be transformed
* The fourth parameter is an optional parameter used to instruct Apache Spark on how many cores/partitions to use to process the data

A sample of the output from the *CategoryCountPerDay* application is presented in figure 14:

|  |
| --- |
| [{  "Date": "2016-06-10T00:00:00Z",  "Data": [  { "Category": "Trump", "Count": 165018 },  { "Category": "Clinton", "Count": 108596 }  ]  }, {  "Date": "2016-06-11T00:00:00Z",  "Data": [  { "Category": "Trump", "Count": 113174 },  { "Category": "Clinton", "Count": 55724 }  ]  }] |

Figure 14: CategoryCountPerDay JSON output sample

Notice that the *Date* attribute shows the date with a zeroed time offset, as it is a representation of an entire day’s count of categories.

# Challenges

The challenges experienced during the project were both technical and nontechnical.

On the technical side, learning how to create the physical cluster computing framework was a new challenge, having had no prior experience in that field. Many mistakes were made with regards to how the environment should be setup and tuned, especially when using hardware like the Raspberry Pi’s, which are resource constrained. Many times the OS on the MicroSD cards became corrupted due to the Raspberry Pi’s not dealing well with unexpected power loss.

Learning Scala was a unique challenge, as the programming experience in group was mostly in Object-Oriented languages. Scala provides a hybrid OO/functional programming paradigm which took some time to adjust to, particularly around working with immutable state.

On the nontechnical side, the work allocation seemed to be unevenly balanced. The majority of the work was in the hardware setup, cluster computing software configuration and data transformation programming, as well as the data visualisation component setup and programming. This had a major impact on the quantity of work produced for the final product.

# Recommendations

A data stream instead of flat JSON files is recommended for a solution such as this. It would fit the intent of the application much better as data could be processed near real-time instead of being batched. Processing historic data in large batches is very expensive to compute, as large datasets require extra management overhead to partition the data into smaller pieces in order to be processed in the resource constrained environment of the Raspberry Pi’s.

# Conclusion

It is possible to create a cost-effective big data processing system using commodity hardware and open source software. The adage “many hands make light work” is certainly true when applied to the world of cluster computing, leveraging the power of parallel computing. The Raspberry Pi cluster computing system presented in this group project is capable of handling the workload issued to it, and would handle various other scenarios quite well.

# References

|  |  |
| --- | --- |
| [1] | S. Sardana and S. Sardana, “Big Data: It's Not A Buzzword, It's a Movement,” Forbes, 20 November 2013. [Online]. Available: http://www.forbes.com/sites/sanjeevsardana/2013/11/20/bigdata/#3b507d136be7. [Accessed 1 July 2016]. |

# Appendix A Timesheet

# Appendix B Environment Configuration

# Appendix C Cost of Hardware