Wits University

School of Electrical and Information Engineering

ELEN7046 - Software Technologies and Techniques

Big Data Visualization using Commodity Hardware and Open Source Software

Authors:	*Student Number:*
Gareth Stephenson	778919
Matsobane Khwinana	779053
Sidwell Mokhemisa	1229756
Dave Cloete	1573016
Kyle Trehaeven	0602877N

Abstract

The project aims to provide a low cost solution to big data processing problems, enabling a commercially viable option to individuals, small businesses and academia using commodity hardware. This report explores the use of Open Source technologies Apache Spark and Scala as well as the low cost and scalable Raspberry Pi's. After building a prototype system to source, process and visualize data from twitter, it was concluded that

commodity hardware is a viable small-scale solution to working with Big Data.

June 26, 2016

Declaration of Originality

Contents

1 Introduction **2**
- 1.1 Problem Statement . . . 2
- 1.2 Solution Summary . . . 2
- 1.3 Approach . . . 2

2 Literature Review **2**

3 Lifecycle Methodology **2**

4 Assumptions and Constraints **3**
- 4.1 Tweet Locations . . . 3
- 4.2 Pros and Cons determinations . . . 3

5 Design Decisions **4**

6 Success Criteria **4**

7 Solution Design **5**
- 7.1 High Level Design: Component Architecture . . . 5
- 7.2 Detailed Designs . . . 6
  - 7.2.1 Data Acquisition (Batch) . . . 6
  - 7.2.2 Data Acquisition (Streaming) . . . 6
  - 7.2.3 Data Processing . . . 6
  - 7.2.4 Data Visualization . . . 6
- 7.3 Operational Model: Infrastructure Design . . . 6
- 7.4 Possible Extensions . . . 7

8 Conclusion **7**

9 Appendices **8**

1 Introduction

This report presents the work done by Group 2 in response to the project brief for ELEN7046: Software Technologies and Techniques.

The report will broadly focus on the following topics:

- The Methodology followed to execute the project;
- The Architecture of the solution developed for the project; and
- The different technologies used to deliver the solution.

1.1 Problem Statement

There is an abundance of big data available to individuals and companies with very limited capacity to refine this data into meaningful information, particularly for small scale endeavours. Big Data processing is often locked behind high cost barriers to entry, and individuals, start ups and academics may find it difficult to be involved in Big Data processing.

1.2 Solution Summary

Commodity hardware is available to provide a means by which the barrier to entry for Big Data projects can be overcome. Simple, low cost components can be leveraged to address each of the parts of a big data processing solution, whether it be data sourcing, transormation, or visualization; and can be scaled according to needs or as required.

1.3 Approach

2 Literature Review

3 Lifecycle Methodology

In order for group two to successfully deliver this project, a development methodology based on IBM Rational Unified Process (RUP) was followed albeit tailored

to cater for the specific needs of this project.

The diagram below depicts the IBM RUP model:

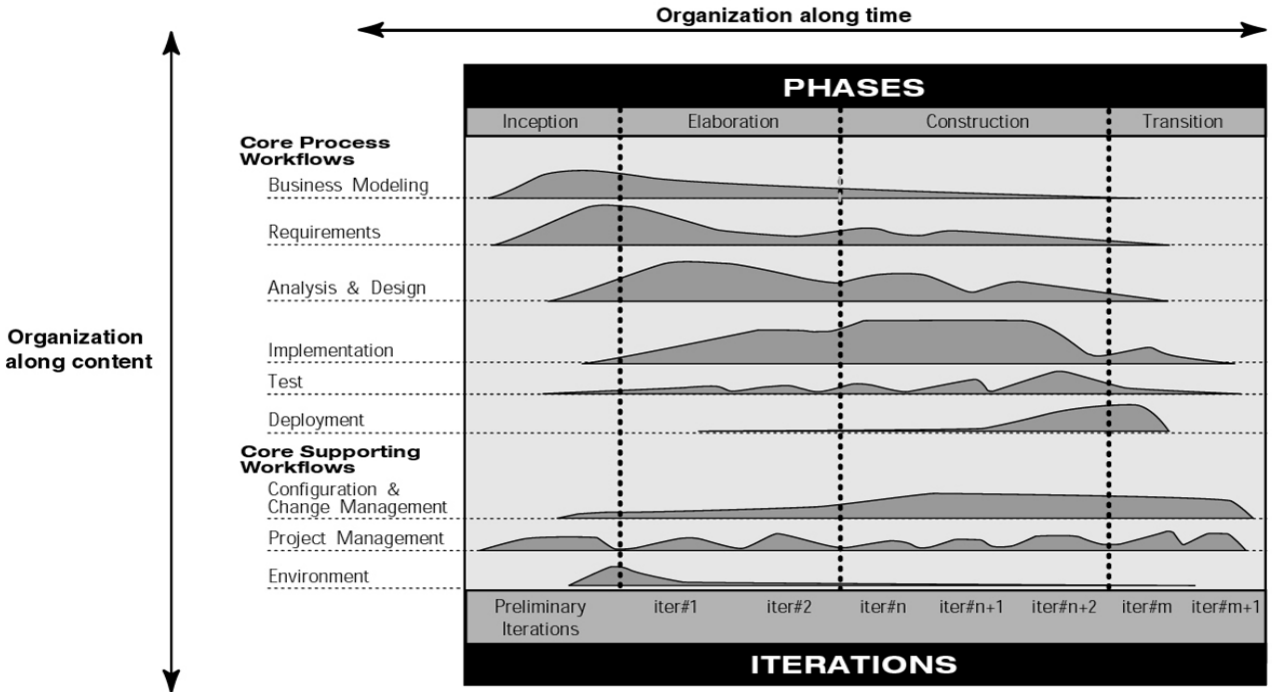

Figure 1: IBM Rational Unified Process (Source: RUP, Best Practices for Software Development Teams)

4 Assumptions and Constraints

4.1 Tweet Locations

The group encountered tweet location issues.

4.2 Pros and Cons determinations

Rudimentary algorithm for determining Twitter statements (tweets) that are against or for a particular candidate was adopted...

5 Design Decisions

6 Success Criteria

- Build a system that uses commodity hardware to solve for big data problems;
- Provide a solution that covers the data sourcing, transformation and presentation of social media data from Twitter relating to the United States and South African elections. The end result must be visualization that provides insight to the sentiment of election candidates on Twitter.

7 Solution Design

7.1 High Level Design: Component Architecture

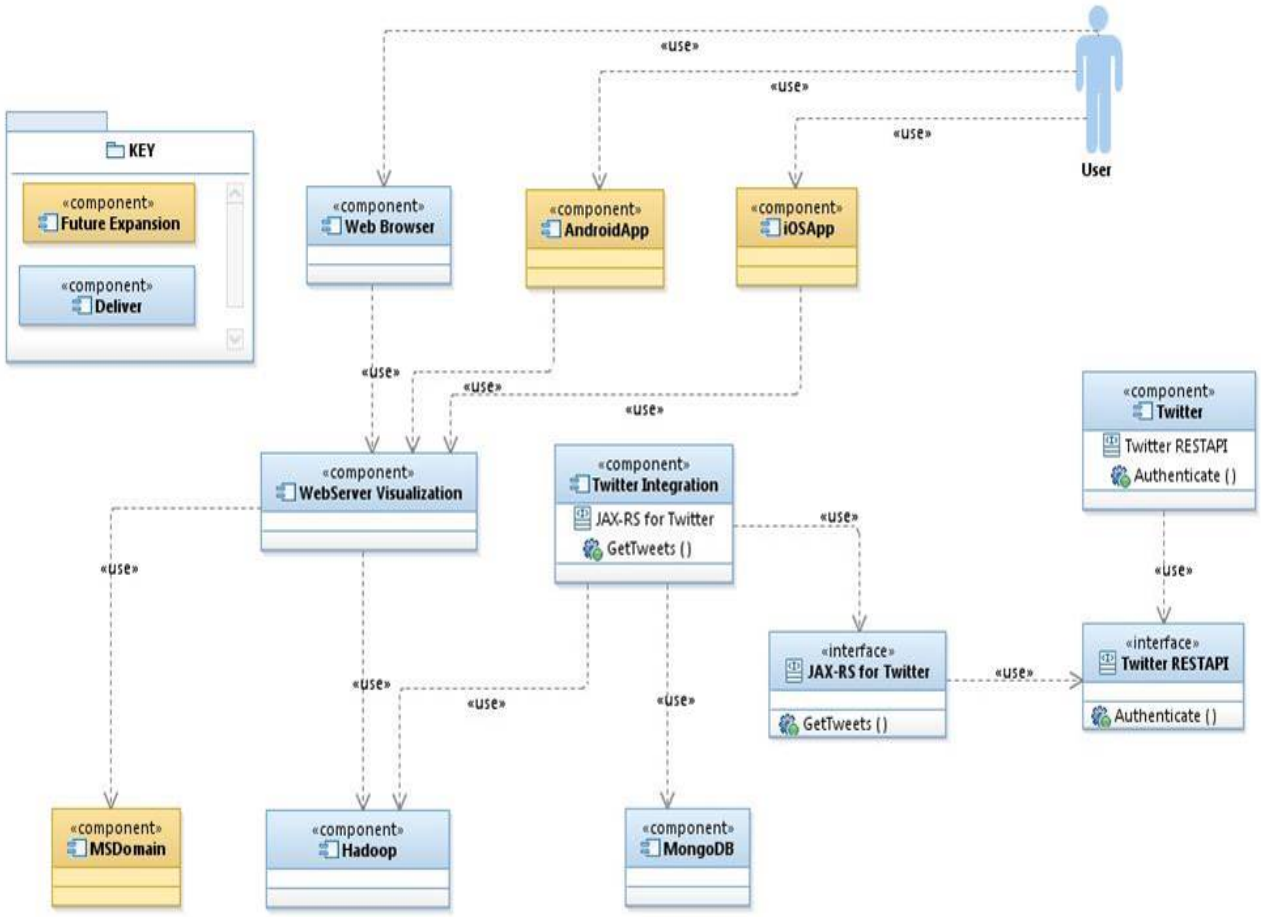

Figure 2: High Level Component Model.

7.2 Detailed Designs

7.2.1 Data Acquisition (Batch)

7.2.2 Data Acquisition (Streaming)

7.2.3 Data Processing

7.2.4 Data Visualization

7.3 Operational Model: Infrastructure Design

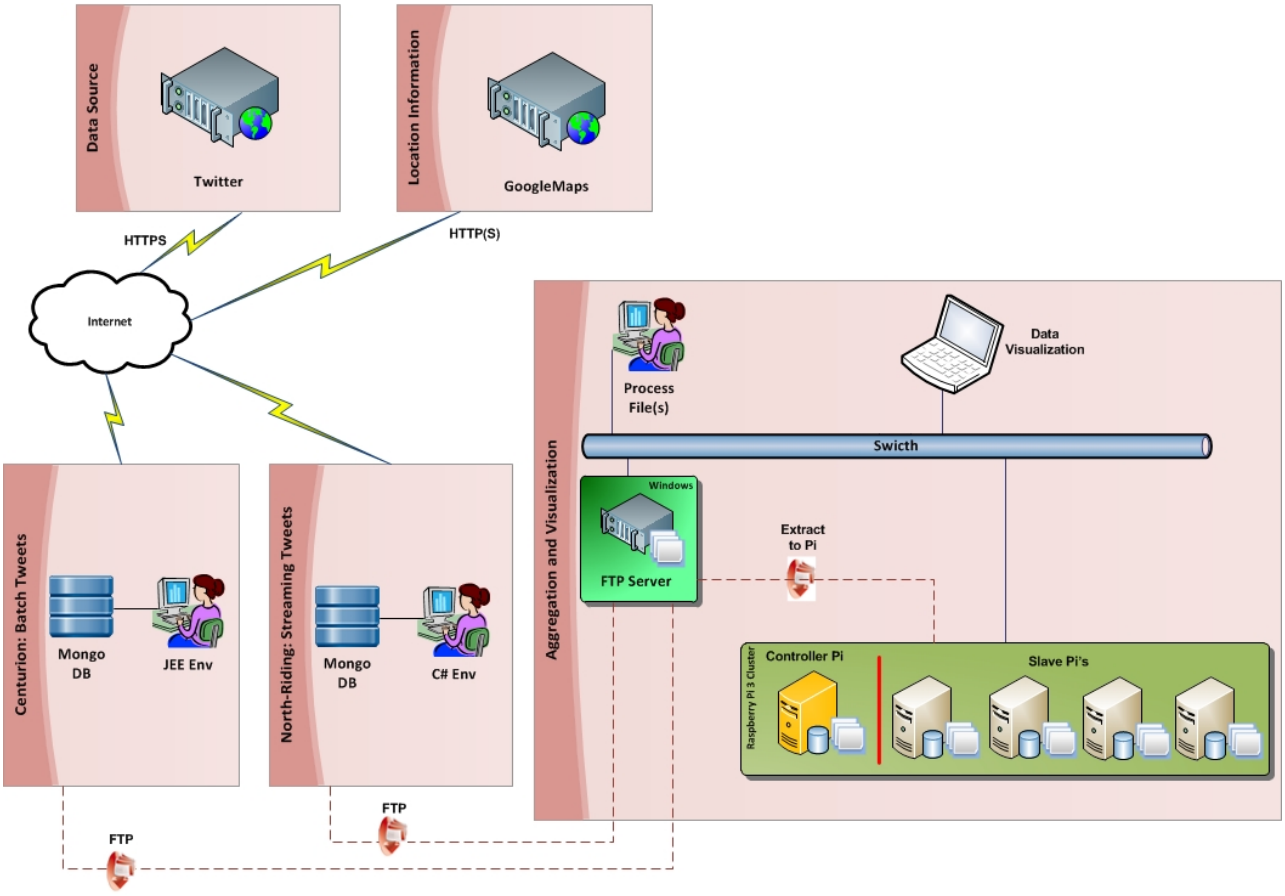

Figure 3: Operational Model: Phyical

7.4 Possible Extensions

- **Personnel Shortfall:** Inexperience in the management team is a potential risk, due to the possible oversight and inaccuracies . Leach does not have sufficient IS Management experience*(pg 502 paragraph 4)*, the project may suffer if leach continues at his current position

8 Conclusion

All this hardware and software is available to anybody interested in Big Data processing.

The hardware is cheap and the software is free.

The learning curve in the beginning can be quite steep but is ultimately very rewarding in terms of what can be achieved with so little financial investment.

References

[Van Vliet, 2008] Van Vliet (2008). Software Engineering Principles and Practice

[Frederick P. Brooks, 1975] Frederick P. Brooks (1975). The Mythical Man-Month

9 Appendices