Wits University

School of Electrical and Information Engineering

ELEN7046 - Software Technologies and Techniques

Big Data Visualization using Commodity Hardware and Open Source Software

Individual Report

Author: Sidwell Mokhemisa

Student Number: 1229756

Abstract

XXXX.

July 2, 2016

Declaration of Originality

Contents

1 Introduction

XXX

1.1 Problem Statement

XXX

1.2 Solution Summary

XXX

1.3 Approach

2 Background

2.1 Data Sourcing

This project was .....

3 Lifecycle Methodology

In order for the team 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:

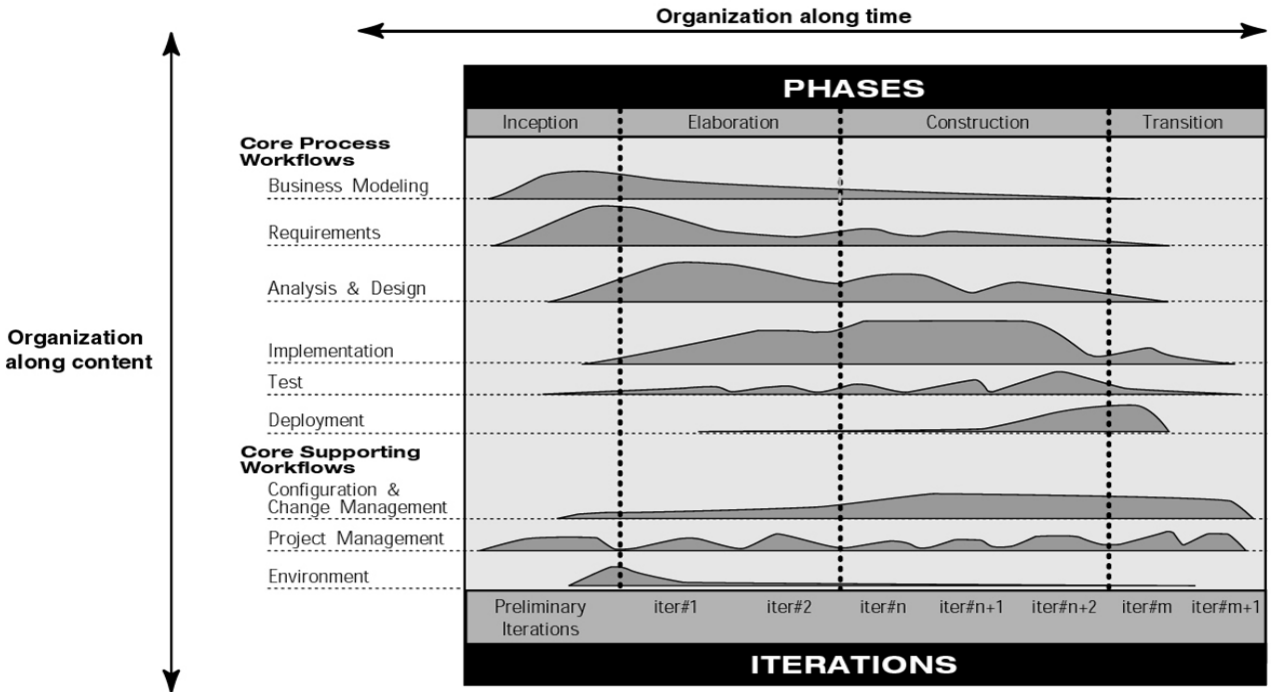

Figure 1: IBM Rational Unified Process (Source: RUP, Best Practices for Software Development Teams)

4 Requirements - Use Case Models

XXXXXX.

4.1 View Elections Analytic Data

This section covers the details around the visualization Use Case. The diagram below depicts the actual Use Case followed by a table that further discusses the Use Case details:

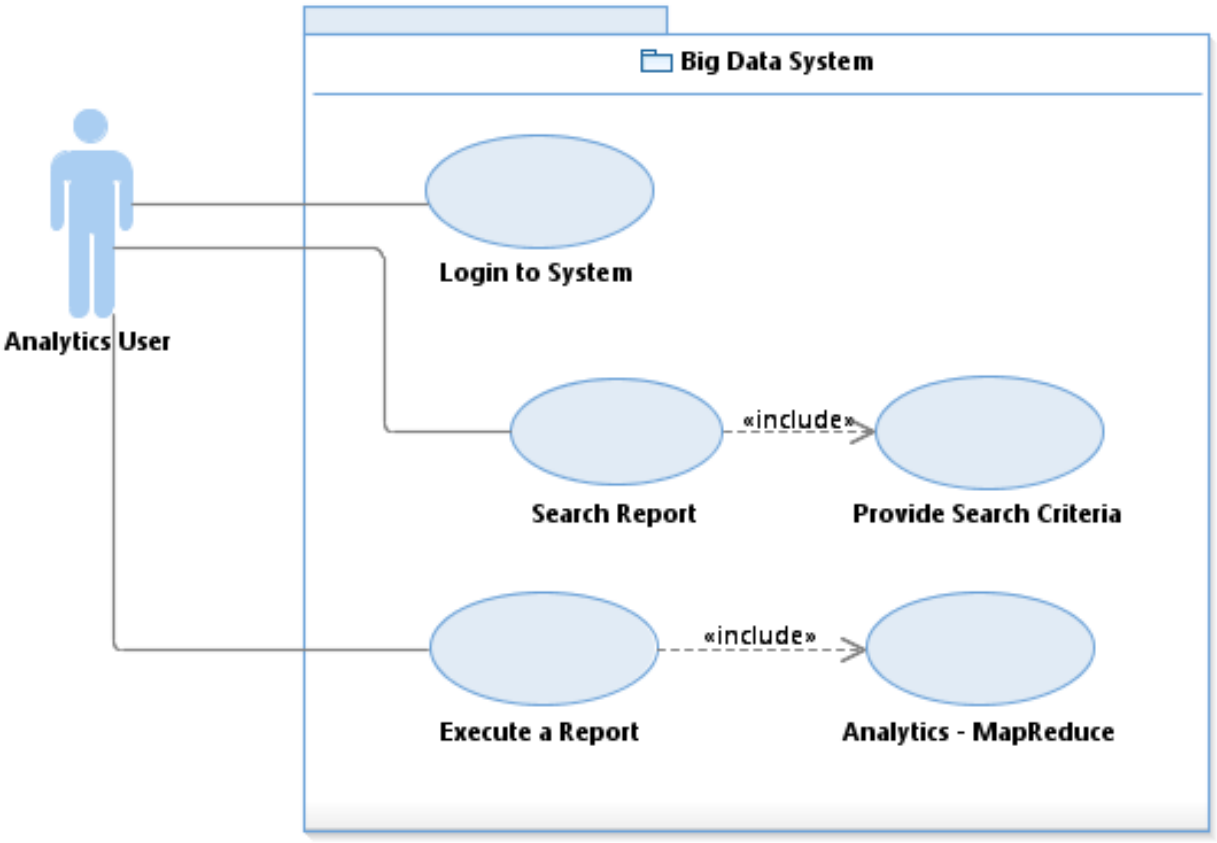

Figure 2: Use Case Diagram - View Twitter Elections Analytics

This table provides additional information to supplement the Use Case diagram.

Use Case ID:	UC01		
Use Case Name:	View Analytics Social Media Report overlaid on map background.		
Created By:	Sidwell	Updated By:	Sidwell
Date Created:	02/05/2016	Date Modified:	07/05/2016

Actor :	Analytics User
Description:	This use case describes how the user will use the system to run analytics based on social media data received from Twitter.
Pre-conditions:	Web browser opened and user logs onto the analytics site.
Post-conditions:	User views requested report overlaid on map background. Drill up/ down functionality provided by the application.
Normal Course:	1. Logon to the application. 2. Search report from list of available reports 3. Execute a report of choice
Frequency of Use:	
Alternative Courses:	None
Exceptions:	None
Includes:	1. Provide search criteria or Hashtag(s). 2. System runs report using Map Reduce and parallel processing in order to produce report results.
Special Requirements:	1. Ad-hoc access using most browsers (IE, Chrome, Safari).
Assumptions:	1. User login based on access to computer with browser and not necessarily integration to an LDAP compliant system. 2. Support for mobile apps once developed.
Notes and Issues:	

4.2 Acquire Twitter Data

This section covers the details around the data acquisition Use Case. The diagram below depicts the actual Use Case followed by a table that further discusses the Use Case details:

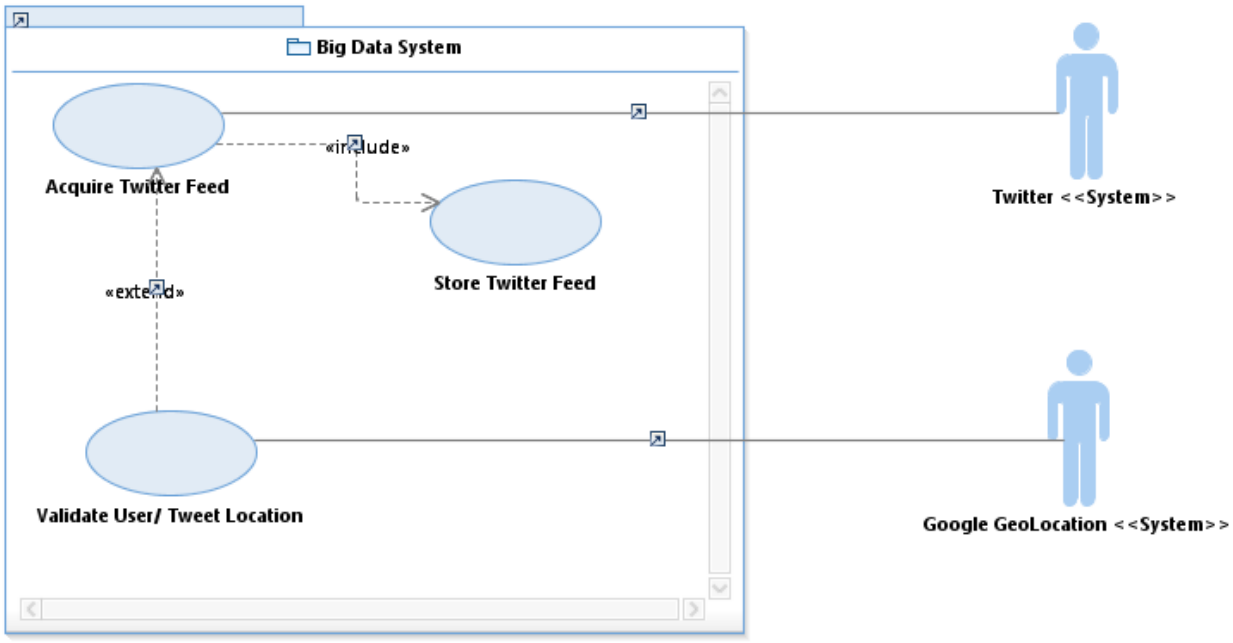

Figure 3: Use Case Diagram - Acquire Twitter Data

This table provides additional information to supplement the Use Case diagram.

Use Case ID:	UC02		
Use Case Name:	Acquire social media feed from twitter to enable big data analytics.		
Created By:	Sidwell	Updated By:	Sidwell
Date Created:	02/05/2016	Date Modified:	07/05/2016

Actor :	Twitter and Google GeoLocation
Description:	This use case describes how data is collected from twitter based on subscribed topics stored in a database for later use in analytics processing.
Pre-conditions:	1. Application logs into twitter with provided credentials and starts streaming all the data that complies with subscribed topic(s). 2. Topics to subscribe to are configured on the system beforehand. 3. For each tweet streamed, the application through its orchestration service attempts to verify location from which tweet was sent, or from profile of user sending twitter using Google GeoLocation Service. 4. Where location could not be verified, the tweet is stored in the database without location information.
Post-conditions:	1. Developed application authenticates and streams data. 2. Streamed data is stored in the database with location information where location could be determined.
Normal Course:	1. Configure election related topics to subscribe to (both US and SA). 2. Allow application to log onto both twitter and Google GeoLocation. 3. Stream tweets through orchestration service while attempting to verify location by validating certain data via Google GeoLocation Service. 4. Store all tweets regardless of location information availability.
Frequency of Use:	
Alternative Courses:	None
Exceptions:	None
Includes:	1. Storing of tweeter feeds in a database.
Special Requirements:	1. Username token provided by twitter. 2. Username token provided by Google GeoLocation Service. 3. Internet access to connect to both services.
Assumptions:	1. Availability of infrastructure resources to harvest more than a million twitter records and store them.
Notes and Issues:	

5 Assumptions and Constraints

Early

6 Design Decisions

The table below details all the key design decisions made in the delivery of the solution:

- **History Data:** History data/ batch....

7 Solution Design

7.1 High Level Design: Component Architecture

The high level component model below depicts the key features of the solution delivered for the project.

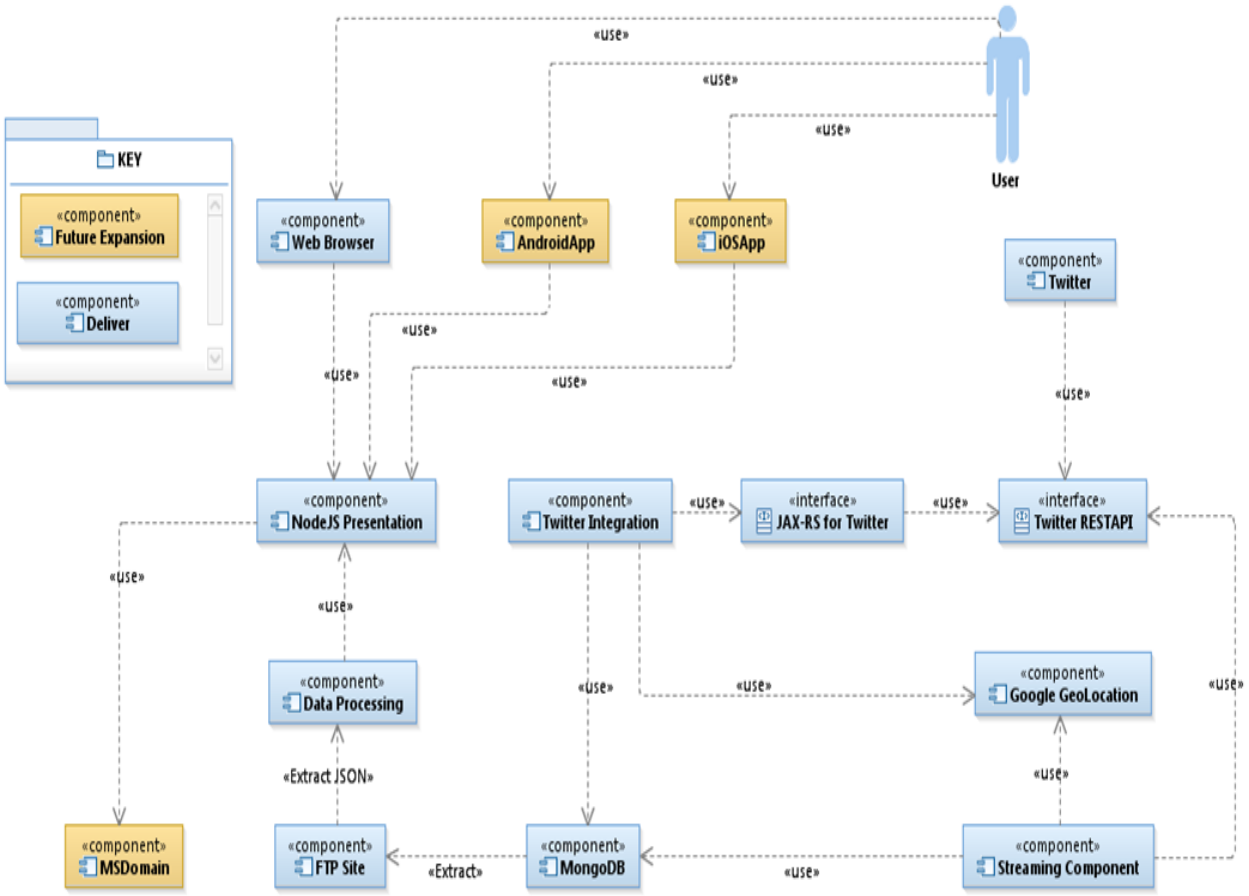

Figure 4: High Level Component Model.

7.2 Solution Sequence Diagrams

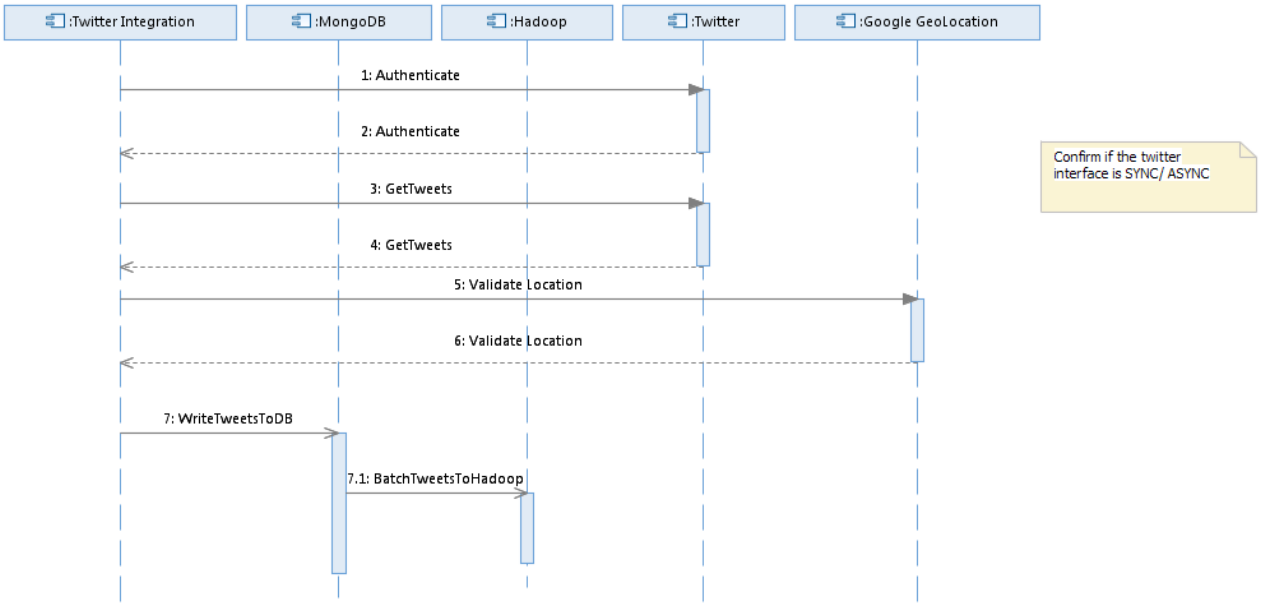

Figure 5: Sequence Diagram: Data Integration - History

7.3 Operational Model: Infrastructure Design

The Raspberry Pi

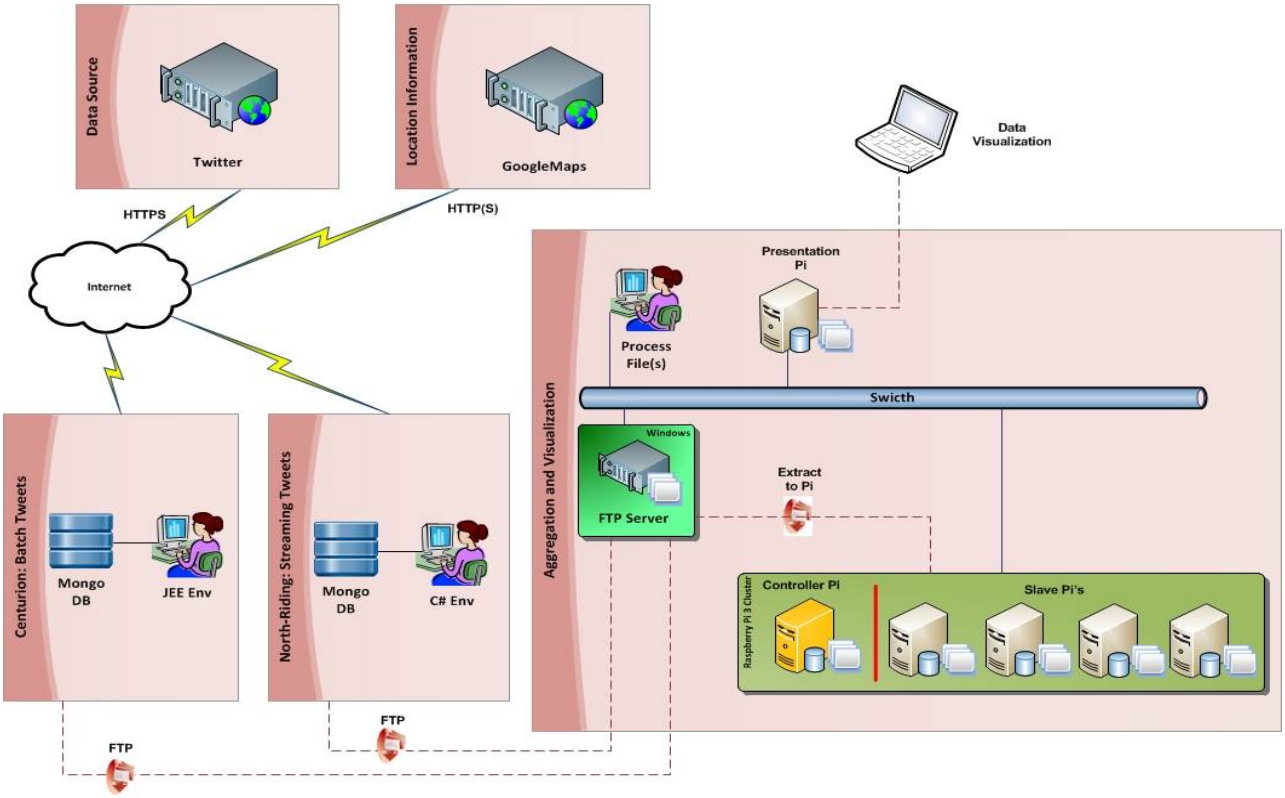

Figure 6: Operational Model: Phyical

The project

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

[1] . S. Madam. From Databases to Big Data. IEEE Computer Society, 2012.

[2] . V. Kumar, R. Yuvaraj, C. Anusha. Effective Distribution of Large Scale Datasets Clustering Based on MapReduce. 2016.

9 Appendices