Frog-B-Data

Project Plan

Version 4.0

2 May, 2016

Revision Signatures

By signing the following document, the team member is acknowledging that he has read the entire document thoroughly and has verified that the information within this document is, to the best of his knowledge, accurate, relevant and free of typographical errors.

Name	**Signature**	**Date**
Sushant Ahuja		
Cassio Lisandro Caposso Cristovao		
Sameep Mohta		

Revision History

The following table shows the revisions made to this document.

Version	Changes	Date
1.0	Initial Draft	7 October, 2015
1.1	Revised errors	25 October, 2015
1.2	Updated Schedule	15 December, 2015
2.0	Updated Iterations and Schedule	21 January, 2016
3.0	Updated Iterations and Schedule, Updated Development Tools and Scope	20 April 2016
4.0	Final minor changes	2 May, 2016

Table of Contents

1 Introduction

1.1 Purpose

The purpose of this document is to provide an overview of the planning for project Frog-B-Data. This document covers the project overview, specifies resources required for this project, sets milestones and identifies the roles and responsibilities of all the team members. This document also gives specific instructions on the monitoring and reporting mechanisms and discusses risk management techniques for this project.

1.2 Overview

This document includes the following four sections.

Section 2 - Project Overview: Gives a detailed idea of the scope and objectives of our project and also provides an insight on the background of the system being developed.

Section 3 - Resource Specification: Lists the software, hardware, and individuals involved in the development of the system

Section 4 - Project Management: Specifies the milestones and deliverables for this project, roles and responsibilities of the team members, weekly team meetings, along with risk analysis and planning.

Section 5 - Glossary of Terms: Lists all the technical terms that are mentioned in this document with their definitions.

2 Project Overview

2.1 Scope and Objectives

In the near future Big Data is going to touch every business and every person on this planet. MIT Technology Review reported that currently less than 0.5 % of all data collected is being analyzed and used; therefore its potential is huge. We are Frog-B-Data and our senior capstone project is a Big Data research project in which we setup and compare three environments: stand-alone Java, Apache Hadoop and Apache Spark. Apache Hadoop and Apache Spark are setup as clusters with three nodes. We handle the application dependencies using Apache Maven and develop them on Eclipse IDE, using Mahout and ML libraries. Apache Hadoop has been the go-to framework for Big Data applications, but is slowly being replaced by Apache Spark which is gaining more popularity. We perform four comparison tests: Word Count, Matrix Multiplication, Recommendation using Co-occurrence Matrix or Collaborative filtering algorithms, and K-means clustering. Non-structured data files of sizes ranging from a few Megabytes to 10 Gigabytes are being used for comparative studies. Based on the results, we will build our own recommender system in the preferred framework.

2.2 Project Background

Data now streams from everywhere in our daily lives: phones, credit cards, computers, tablets, sensor-equipped buildings, cars, buses, trains and the list goes on and on. We have heard so many people say “There is a Big Data Revolution”. What does that mean? It is not the quantity of data that is revolutionary. The Big Data revolution is that now we can do something with the data. The revolution lies in the improved statistical and computational methods which can be used to make our lives easier, healthier and more comfortable.

Familiar uses of Big Data to a common man include “recommendation engines” used by Netflix and Amazon, credit card companies, and tech giants like Facebook. In the public realm, there are all kinds of applications: allocating police resources by predicting where and when crimes are most likely to occur; finding associations between air quality and health; or using genomic analysis to speed the breeding of crops like rice for drought resistance. However, this is a very small fraction of what can be done and what is being done. The potential for doing good is nowhere greater than in public health and medicine where people are dying everyday just because data is not being properly shared.

Nowadays, it’s not just about mining data and analyzing results, it is about using data smartly. The purpose of smart data is to filter out the noise from the Big Data and hold the valuable data to solve business problems. There are no formulae to convert Big Data into smart data, but if we understand the clues in the questions around the data and analyze data qualitatively, we can use it smartly.

3 Resource Specifications

3.1 Software

i) **Development Tools**
   (1) Hadoop 2.7.1
   (2) Spark 1.5.1
   (3) Java 8.6
   (4) Scala 2.11.7
   (5) Eclipse Mars 4.5
   (6) Python 2.7.9

ii) **Supporting Tools**
   (1) Core FTP 2.2
   (2) GitHub
   (3) Slack
   (4) Sublime Text
   (5) FileZilla

iii) **General Utilities**
   (1) Adobe Photoshop CC 2015 (v2015.0.1)
   (2) Microsoft Office 2013
   (3) Google Drive

iv) **Operating Systems**
   (1) Ubuntu 15.04

3.2 Hardware

i) **6 PCs running Ubuntu 15.04**
   (1) Intel Core i5-4570S CPU @ 2.90GHz *4 (64-bit)
   (2) 8 Gigabytes RAM
   (3) 500 Gigabytes HDD

ii) **TCU Brazos Web Server**

3.3 Contacts

i) **Client**
   (1) Dr. Antonio Sanchez – a.sanchez-aguilar@tcu.edu

ii) **TCU Faculty Advisor**
   (1) Dr. Donnell Payne – d.payne@tcu.edu

iii) **Team Frog-B-Data**
   (1) Sushant Ahuja – sushant.ahuja@tcu.edu
   (2) Cassio Lisandro Caposso Cristovao – cassio.capossocristovao@tcu.edu
   (3) Sameep Mohta – sameep.mohta@tcu.edu

4 Project Management

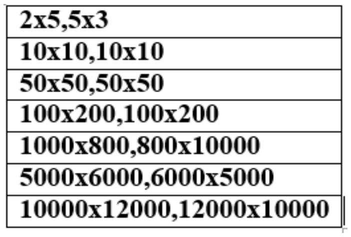
4.1 Milestones and Deliverables

Skeleton Website	4 October, 2015
Software Engineering Presentation	15 December, 2015
Project Plan v1.0	15 December, 2015
Requirements Document v1.0	15 December, 2015
Design Document v1.0	15 December, 2015
Iteration 1	15 December, 2015
Iteration 2	2 February, 2016
Faculty Presentation	2 February, 2016
Iteration 3	20 March, 2016
Iteration 4	6 April, 2016
SRS	8 April, 2016
NTASC Presentation	16 April, 2016
Developer Manual	26 April, 2016
User’s Manual and Research Results	26 April, 2016
Final Presentation	28 April, 2016
Complete Documentation	2 May, 2016

4.2 Iteration Descriptions

Iteration 1 15 December, 2015

- Setting up 6 Linux machines
  - Hadoop and Spark on 2 machines as manager nodes
  - 2 worker nodes for Hadoop and Spark each
- Initial Software Tests
  - **Word Frequency** count on all the three environments with text files of different sizes (100MB, 500MB, 1GB, 5GB and 10GB).
  - **Large Matrix** multiplication with the following matrix sizes:



2x5,5x3
10x10,10x10
50x50,50x50
100x200,100x200
1000x800,800x10000
5000x6000,6000x5000
10000x12000,12000x10000

Iteration 2 2 February, 2016

- Setting up a cluster with at least 2 worker nodes for Hadoop and Spark each
- Running Mahout on Hadoop systems
- Running MLlib on Spark systems
- Starting to understand basic recommender algorithms for Hadoop and Spark

Iteration 3 20 March, 2016

- Build recommendation system for Hadoop using Apache Mahout
- Using K-means clustering on Hadoop cluster

Iteration 4 6 April, 2016

- Build recommendation system for Spark using MLlib
- Using K-means clustering on Spark cluster
- Work on making the Hadoop recommender more scalable and reliable

4.3 Team Member Roles and Responsibilities

- Sushant Ahuja – Project Lead, Algorithm Design Lead
- Cassio Cristovao – Technical Lead, Website Architect
- Sameep Mohta – Documentation Lead, Testing Lead

4.4 Monitoring and Reporting Mechanisms

4.4.1 Meetings

Meetings with the client and faculty advisor take place every Wednesday at 11 AM. In addition to this team Frog-B-Data meets every Friday and Sunday at 2 PM.

4.4.2 Communication

Team Frog-B-Data communicates through Slack which is a team communication application and makes communication between team members very easy and convenient. We also use GitHub to upload the code that we write so that it remains in one place. We use Google Drive to upload all the documents for their convenient access at all times. All the team members also remain in contact through email and phone, if needed.

4.4.3 Requirements Control

Frog-B-Data shall be developed according to the Requirements Document. The system will be tested to ensure that the requirements are being met at the end of each iteration. If the client wants a change in the system, the change would be implemented after complete analysis of the proposed change.

4.4.4 Weekly Activity Reports (WARs)

Weekly Activity Reports will be posted on the project website by each group member on Sunday of every week. Below is the link where one can find all the reports.
http://brazos.cs.tcu.edu/1516FrogBData/wars.html

4.4.5 Walk-throughs

Website Skeleton walk-through	*8 October, 2015*
Project Plan v1.0	*Week of October 12, 2015*
Requirements walk-through	*Week of October 19, 2015*
Design walk-through	*Week of November 9, 2015*
Iteration 1 walk-through	*Week of December 7, 2015*

4.5 Risk Management

Contingency	Probability	Severity	Mitigation Strategy
Project Not finished	Low	Catastrophic	Work efficiently In a timely manner
System Failure	Moderate	Critical	Have backup machines
Data Loss	Moderate	Moderate	Backup of Data properly
Group Member unavailable at critical time	Moderate	Moderate	Make sure every team member knows what they are doing
Root Access not available at critical times	Moderate	Moderate	Be prepared and plan in advance

5 Glossary of Terms:

Apache Hadoop: Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets.

Apache Mahout: An Apache software used to produce free implementations of distributed scalable machine learning algorithms that help in clustering and classification of data.

Apache Spark: Apache Spark is an open source cluster computing framework which allows user programs to load data into a cluster's memory and query it repeatedly.

Big Data: Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions

K-means clustering: A way of vector quantization used for cluster analysis in data mining.

Map Reduce: A programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster.

MLlib: Apache Spark's scalable machine learning library.

Root Access: Access to install various software and related items on Linux machines.

Scala: A programming language for general software applications.

Smart Data: Use of Big Data to make it more valuable and help in decision making to solve problems.