UNIVERSITY OF NAIROBI

OPEN SOURCE IMPLEMENTATION OF BUSINESS INTELLIGENCE SYSTEM FOR KENYAN UNIVERSITIES: A CASE OF THE TECHNICAL UNIVERSITY OF KENYA

BY

OBONYO ISHMAEL NYUNYA

A PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT FOR THE REQUIREMENTS OF THE AWARD OF DEGREE OF MASTER OF SCIENCE IN COMPUTATIONAL INTELLIGENCE, SCHOOL OF COMPUTING AND INFORMATICS, UNIVERSITY OF NAIROBI

JUNE 2015

DEDICATION

I hereby dedicate this research project to my dear parents; Joash Obonyo, Jane Obonyo and Wilkister Obonyo. This is due to their immense love, guidance and support I have received from them all the years of my life. May the Almighty God continue to grant them good and long life.

DECLARATION

Researcher's Declaration

This project report is my original work and has not been presented in any other institution for the purpose of an academic award.

SIGNATURE: ____________________ Date: ____________________

OBONYO, Ishmael Nyunya

Registration Number: P52/65641/2013

Supervisor's Approval

This project report has been submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computational Intelligence of the University of Nairobi with my approval as the University Supervisor.

SIGNATURE: ____________________ Date: ____________________

Dr. Elisha T. O. Opiyo

School of Computing and Informatics

University of Nairobi

ACKNOWLEDGEMENTS

My deep appreciation goes to my Supervisor Dr. Elisha Opiyo for his due support, guidance and encouragement during the entire period of this project. Special thanks go to Dr. Lawrence Muchemi and Dr. Agnes Wausi for their very insightful recommendations and support that made me improve my project and research skills in general. I extend sincere gratitude to my lovely fiancée Pauline for moral support and direction that made me complete the project successfully. I wish her well in her studies in Copenhagen University, Denmark. I send my appreciation to the staff of the Technical University of Kenya who sacrificed their valuable time to accept my interviews. I cannot mention all, but I say a big thank you to all who contributed in one way or the other towards my project. I salute you all! Thanks to the Almighty God for providing me with good life and hope up to the completion of this study.

ABSTRACT

Latest technological advancements in computer storage, networking and processor speeds have enabled organizations to develop innovative ways to intelligently collect data that was not possible before. However, this has led to the explosion of data and unprecedented challenges in making strategic and effective use of the available data. The current information systems such as executive information systems and decision support systems, among others, have not been able to provide crucial reports to decision makers. Decision makers require reports that are timely, accurate, actionable, and depicting the whole ‘business picture’. Business Intelligent (BI) systems come to the rescue of decision makers. BI systems provide timely, accurate, and actionable information to the right person enabling quick and correct decisions. Higher learning institutions like universities, being one of the organizations, require such systems for effective management. Past studies reveal little adoption of BI in Kenyan Universities. This research project was geared towards implementing a Business Intelligence System for Kenyan Universities; taking a case of the Technical University of Kenya. Kimball’s dimension modeling was used in designing a data warehouse. The system was implemented using Hadoop cluster integrated with R Statistical Software. Data warehouse was developed and analysis achieved using Hive Query Language and R through data visualization and dashboards. Comparison with the state-of-art open source BI tools was conducted. The project revealed great opportunities for open source tools in data analytics for universities. The study recommended a real-time BI system that would relay live status of events for effective decision-making and also incorporation of unstructured data from social media in improving the University analytics.

LIST OF ACRONYMS

BIOS
Basic Input/Output System 42

DaaS
Data as a Service 26

DSS
Desision Support System 1

EIS
Executive Information System 1

ES
Expert System 1

ETL
Extraction Transformation and Loading 13

GNU GPL
GNU General Public License 30

HBase
Hadoop Database 16

HDFS
Hadoop Distributed File System 14

HE
Hihger Education 4

HiveQL
Hive Query Language 16

HR
Human Resource 3

HUE
Hadoop User Experience 17

IBM
Industrial Business Machines 28

IDE
Integrated Development Environment 44

IS
Information System 1

IT
Information Technology 1

JDBC
Java Database Connectivity 43

KPI
Key Performance Indicator 31

MIS
Management Information System 1

OBI
Operational Business Intelligence 26

OBIEE
Oracle Business Intelligence Enterprise Edition 29

ODBC
Open Database Connectivity 43

OLAP
Online Analytical Processing 8
RDBMS
Relational Database Management System 16
RHIPE
R and Hadoop Integrated Programming Environment 18
ROLAP
Relational OLAP 29
SaaS
Software as a Service 26
SAS
Statistical Analysis Software 29
SQOOP
SQL to Hadoop 16
VM
Virtual Machine 42

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

CHAPTER ONE

INTRODUCTION

1.1 Background to the Study

In order to effectively manage an organization, one requires access to the right information pertaining to that organization which subsequently enables monitoring of activities and assessment of the performance of the organization (Gangadharan & Swami, 2004). However, accessing this information remains a big challenge. Information systems (IS) collect and process vast amount of data in various forms that are not readily perceptible to decision makers. The flow of information within an organization determines the success of that organization (Davenport & Prusak, 1998). Many organizations struggle to collect data and retrieve information for making crucial decisions. Davenport & Prusak (1998) further point out that the large quantities of data collected by the information systems are only vital for operations within the organizations, but are hardly suitable for use in decision making geared towards business strategies and objectives. Decisions in organizations are made by human beings and not the management information systems, as a result, presentation of data plays a very important role in any decision making process, and for a strategy based on any decision to be successful; the information has to be comprehensive and perceptive (Herring,1992; Malik, 2005).

The importance of management information system as a support tool in organizations still remains inadequate. For instance, Olszak & Ziemba (2007) observe that for a long time management information systems (MIS) have been supporting organizations in their different tasks, however, today many information technology (IT) systems have undergone significant depreciation. The existing information systems (IS) such as management information systems (MIS), decision support systems (DSS), expert systems (ES), and executive information systems (EIS) have not always met decision makers’ expectations such as: making decisions under pressure; monitoring competition; possessing such information on their organizations that includes different points of view; and carrying out constant analyses of numerous data and considering different variants of organization performance (Olszak & Ziemba, 2007). Muhammad et al. (2014) denote that though some systems offer limited analytical features; these features just cover only their own application's database and cannot be used together with other systems in an organization, as a result, access to summarized as well as detailed information through a single user interface becomes a challenge.

Business Intelligence systems come to the rescue of decision makers. The purpose of the BI systems is to combine different data sources into information about processes in a company and provide this information in appropriate and timely way to the company management. Stackowiak et al. (2007) define Business Intelligence as the process of taking large amounts of data, analyzing it, and presenting a high-level set of reports that summarize that data into the basis of business actions, enabling management to make vital day-to-day business decisions. "*The implementation of business intelligence systems can contribute to improved information quality in several ways, including faster access to information, easier querying and analysis, a higher level of interactivity, improved data consistency due to data integration processes and other related data management activities (for example data cleansing, unification of definitions of the key business terms and master data management)*" (Popovic, Coelho, & Jaklic, 2009). In today's businesses, increasing standards, automation, and technologies have led to vast amounts of data becoming available in data warehouses, improved extract, transform and load (ETL) and OLAP reporting technologies (Ranjan, 2009) enabling Business Intelligence by sifting through large amounts of this data, extracting pertinent information, and turning that information into knowledge upon which actions can be taken. The benefits of having a Business Intelligence system cannot be overemphasized; universities being one of the organizations also require such systems. Today more and more organizations are turning towards BI for making better business decisions. This is true based on the Gartner's research report (Gartner Research, 2009) which states, "*For the fourth year in a row, BI applications have been ranked the top technology priority in the 2009 Gartner Executive programs survey of more than 1,500 Chief Information Officers (CIOs) around the world*."

Although the use of Business Intelligence tools is popular in industries, the adoption of open source tools is limited, with closed and commercial tools dominating the market (Golfareli, 2009). This project was geared towards the implementation of Business Intelligence system in the Kenyan universities using open source tools.

1.2 Statement of the Problem

Drawing from personal experiences with over three years of continuous service in a University in Kenya, this section described the underlining problems the Technical University of Kenya was facing in regards to access to valuable, correct, timely, and actionable information depicting the whole ‘business picture’ for effective decision making.

In the University, there were multiple data sources which are used in the generation of information for decision making. Data was mainly generated from the management information systems; these are basically structured data. Other kinds of data, which were unstructured in nature, were generated from log files, emails, social media, and official documents like memos. University decision makers needed information from multiple disconnected data across all departments or units of the University. Lack of consolidated data prevented seeing the ‘complete business picture’ of the University. Consolidating data from sources like accounting, student MIS, students’ portals, e-learning portals, Human Resource (HR)/Payroll, and websites, among others, to generate the required reports was complex. Different definitions existed across data stores creating difficult and unreliable data mappings. For example, student records would exist in student MIS, and financial systems and some cases wrong registration numbers would be entered during payment of fees. Too much time and money was spent on generating the required reports. Hence valuable team members were busy creating reports instead of making decisions on the report data. These reports were often rigid, not dynamic; preventing data analysis and drill-down. You would find that once a report is generated, people have to use it as is, so that in case of further inquiries another report had to be generated again. Data generated from different systems were usually unreliable. Data quality was poor and not validated and therefore not trusted fully. Concerns over data accuracy eroded confidence to make important decisions. During decision making, there was usually lack of timeliness of key information. Reports were often only available monthly or quarterly; however users needed them on-demand. Delays result in stale information, preventing optimal decision making. Key decisions were often delayed while waiting for data to be updated. Due to incomplete data for making important decisions; missing data prevented seeing a complete view of the university, analytics were incomplete and could not show an accurate representation of top issues and users were often unaware that certain data is missing and assumed it was not present.

The University needed to optimize resources, foresee new opportunities and seize them; report accurate information and in best way possible to government, sponsors and the public; also to answer critical questions like which locations and demographics they get students from; and amongst others. To successfully address these, there needed an effective quick responses requiring access to timely and accurate student, research and operational data; and this called for Business Intelligence. This study was geared towards addressing these issues by implementing a Business Intelligence System.

The global Higher Education (HE) market has become fiercely competitive. The Universities, for example, need to optimize resources, foresee new opportunities and seize them; they need to report accurate information and in best way possible to government, sponsors and the public; they also need to answer critical questions like which locations and demographics they get students from; and amongst others.

To successfully address these, there should be effective quick responses requiring access to timely and accurate student, research and operational data; and this call for Business Intelligence. This study was geared towards addressing these issues by implementing a Business Intelligence system for Universities in Kenya using open source tools, taking a case of the Technical University of Kenya.

1.3 Objectives of the Study

1.3.1 General Objective

The main aim of conducting this project was to implement a Business Intelligence system for Universities in Kenya using open source tools taking a case of the Technical University of Kenya.

1.3.2 Specific Objectives

In pursuit to implement a business intelligence system for Universities in Kenya using open source tools, this project was aimed at realizing the following specific objectives:

- To investigate appropriate open source tools in the implementation of Business Intelligence systems.
- To design and develop a university data warehouse.
- To implement dashboards for different BI users.
- To perform analytics pertaining to universities in Kenya.

1.4 Research Questions

This project study was expected to answer the following research questions:

- Which open source tools are appropriate for implementing Business Intelligence systems?
- How can a university data warehouse be designed and developed?
- How can dashboards for different BI users be implemented?
- What kinds of analytics can be performed pertaining to Universities in Kenya?

1.5 Significance of the Study

In the execution of this project, the following will be beneficial both to the developers and scholars in the BI community:

- The study will reveal the current tools and approaches towards implementing Business Intelligence systems in the Kenyan universities. This will enable developers who are enthusiastic to come up with BI systems by tapping into the available opportunities using the open source tools.

- Universities in Kenya will have a reference point when it comes to implementation of BI projects.

- The study will add to the knowledge base, the different techniques and methods of developing optimal open source Business Intelligence Systems in the Kenyan universities. The study will also open new insights of research areas that can improve BI implementation in organizations.

1.6 Scope

This project was intended to deliver a BI information system based on the typical business processes of the Universities in Kenya. The system incorporated processes relating to students for example course application, selection (short listing), admission, deferment, course transfers, registration, fee invoicing and payment, class attendance, library usage, examination, graduation and alumni management. Processes relating to human resource management, customer relations management, and social corporate responsibly, among others were not incorporated.

1.7 Limitations

Due to the cost and time constraints, this study encountered the following limitations:

- Not all data sources were incorporated, only data relating to students were considered. In real scenario, all data sources would be considered both from internal and external sources. Sourcing data from social media requires extra tools or programming that would be time consuming.

- Sample data were generated for test where real data were unavailable; in real life scenarios some data may not be readily available or accessible for analysis.

- Hadoop was installed in a single node pseudo-distributed architecture, however, in order to realize the full potential of Hadoop framework on large data sets, a cluster of commodity servers is usually required. This made development and testing a bit slow; again large data samples could not be tested effectively.

1.8 Assumptions

In the execution of this project, some assumptions were however made. First, the reports and analysis generated using sample data were assumed to be the same as for the case of data in real life. Then, data acquired from the Technical University of Kenya would apply for any other university in Kenya. All Universities in Kenya apply somehow the same curriculum and therefore assumed to have the same business processes. Lastly, the BI system will work for any quantity of data provided the hardware requirements are met since the software is expected to scale out without degradation when more data is incorporated.

1.9 Definition of Key Terms

Some terms had been used to clearly bring out the topic of this research and were elaborated as follows:

- **Business intelligence:** a set of tools and techniques for the consolidation of data from multiple sources (both internal and external) relating to processes of an organization and transforming this data into meaningful and useful information for business analysis purposes.

- **Business intelligence system:** an information system that employs business intelligence tools to produce and deliver information.

- **Business intelligence tool:** a type of application software designed to retrieve, analyze, transform and report data for business intelligence. The tool can be used for querying and reporting, online analytical processing (OLAP), data mining, or dash-boarding, among others.

- **Implementation:** is the realization of a BI system; it is the building process and deployment of a BI system.

- **Open source tool:** a program or just a tool that performs a very specific task, in which its source code is openly published for use and/or modification from its original design, free of charge. These tools are typically created as a collaborative effort in which developers improve upon the code and share the changes within the community and are usually available at no charge under a license defined by the Open Source Initiative (Webopedia, 2015). These tools have been applied because there is no cost associated with using or modifying them.

- **University:** is an institution of higher (or tertiary) education and research; it is where students are granted academic degrees in a variety of disciplines. University has been considered as one of the organizations where BI is applied.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This section surveyed the past studies on BI in general, Hadoop and R, BI in Higher Learning Institutions, BI methodologies/implementation frameworks, open source BI tools, case implementation of BI in Higher Learning Institutions, and the BI implementation approach proposed in this project.

2.2 Introduction to Business Intelligence

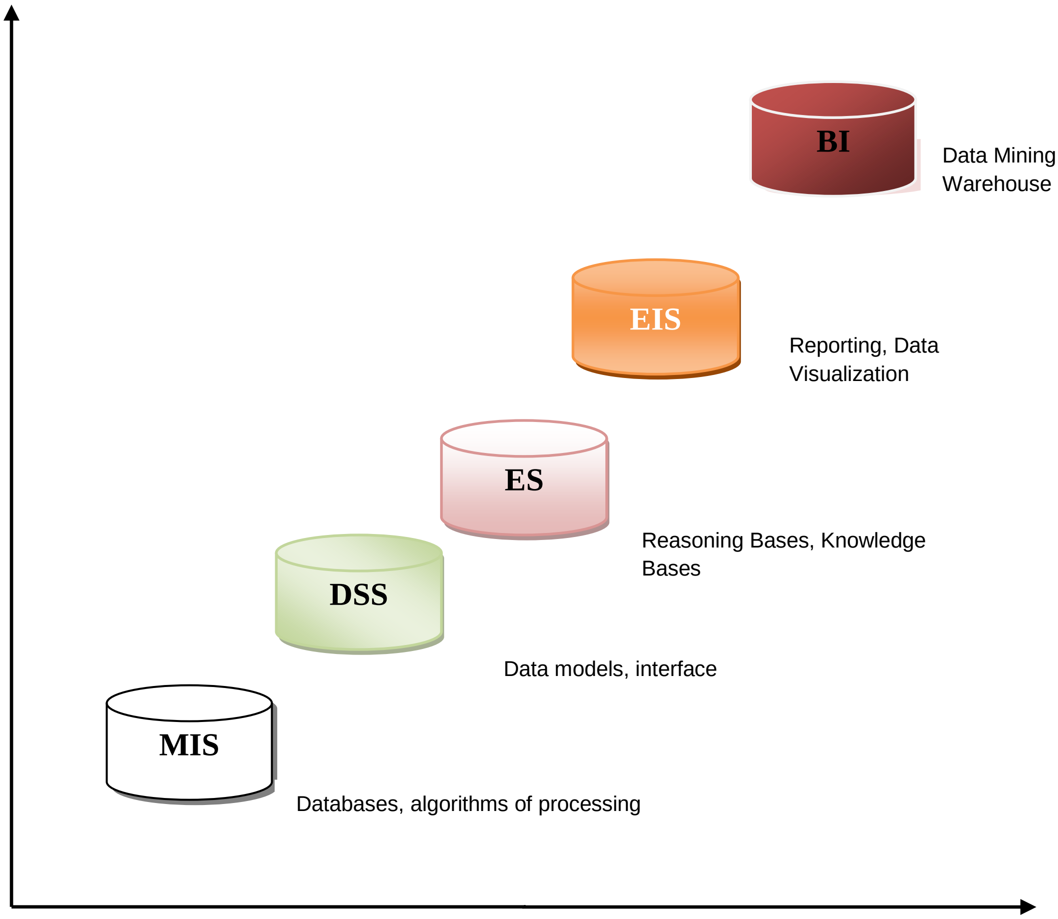

BI
Data Mining, OLAP, Warehouse
EIS
Reporting, Data Visualization
ES
Reasoning Bases, Knowledge Bases
DSS
Data models, interface
MIS
Databases, algorithms of processing

Figure 1: Development of Management Information Systems

Source: Olszak & Ziemba (2007)

How did Business Intelligence come into existence? This section brings into perspective BI and how it fits into the current management information systems.

For a long time management information systems (MIS) have been used by organizations to support different tasks performed by these organizations (Olszak & Ziemba, 2007) specifically mentioning, generation of reports. Olszak & Oziemba (2007) further observe that Information Technology (IT) systems have undergone significant depreciation up to a state where these existing information systems (that is MIS, DSS, ES, EIS) have not met decision makers' expectations, such as: making decisions under time pressure; monitoring competition; possessing such information on their organizations that includes different points of view; and conducting constant analyses of numerous data and considering different variants of organization performance.

It is no doubt that systems supporting decision-making have been evolving since the introduction of computers to commercial enterprises in the mid-1950s when these were used for repetitive data processing data processing (DP) (Dawson & Van Belle, 2013). As the use of computers evolved, transaction processing came to denote the repetitive processing of business events and storing the associated data. Managers soon realized that summarized transactional data had value with respect to decision-making. In the 1970s, the first versions of analytical software packages, referred to as management information systems (MIS), appeared on the market, these systems primarily supported structured decisions (Mallach, 2000) .

The 1980s saw the release of spreadsheet software that continues to be used widely, and by the mid-1980s and early 1990s, executive information systems (EIS) were introduced and rapidly grew in popularity (Tabman & Chi, 1995). These systems promised to provide the top management with easy access to both internal and external information relevant to their decision-making needs. The “easy access” was due to user-friendly interfaces and powerful analytical functionalities. Similar factors accounted for the popularity of decision support systems (DSS) (Carlsson et al., 2002) that included, among others, exception, reporting and an integrated data repository.

The most recent development of systems that support organizational decisions is Business Intelligence (BI). Dresner of Gartner Research is credited with first using the term BI in 1989 to denote "*a broad category of software and solutions for gathering, consolidating, analyzing and providing access to data in a way that lets enterprise users make better business decisions*" (Gartner Research, 2009). Today's business and social environments are complex, hyper-competitive, and highly dynamic. When decisions have to be made quickly and under uncertainty in such a context, the selection of an action plan must be based on reliable data, accurate predictions, and evaluations of the potential consequences. Business intelligence (BI) tools provide fundamental support in this direction. For instance, in medium and large companies, BI tools lean on an integrated, consistent, and certified repository of information called a data warehouse (DW), which is periodically fed with operational data. Information is stored in the DW in the form of multidimensional cubes that are interactively queried by decision makers according to the OLAP paradigm (Golfarelli & Rizzi, 2009).

BI differs from MIS (i.e. DSS, EIS, and ES) in, first of all, their wider thematic range, multivariate analysis, semi-structured data originating from different sources and multidimensional data presentation (Gray, 2003). Baars & Kemper (2008) further emphasize the incorporation of unstructured data into BI platforms for improved reports. Considering the historical evolution of decision support systems, Frolick and Ariyachandra (2006) observe that older decision support systems and executive information systems were application-oriented whereas business intelligence systems are now data-oriented; centered upon data warehousing, they provide the analytical tools required to integrate and analyze organizational data. Transactional systems focus on the fast and efficient processing of transactions, while business intelligence systems focus on providing quick access to information for analysis and reporting (Popovic, Coelho, & Jaklic, 2009). "*While business intelligence systems support decision-making, adapt to the business and anticipate events, transactional systems concentrate on automating processes, structuring the business and reacting to events*." (Popovic, Coelho, & Jaklic, 2009).

It is assumed that BI may support decision making on all levels of management regardless of the level of their structuralization (Olszak, & Ziemba, 2003), for example on the strategic level, BI makes it possible to set objectives precisely and to follow realization of such established objectives. BI allows for performing different comparative reports, for example, on historical results, profitability of particular offers, and effectiveness of distribution channels along with carrying out simulations of development or forecasting future results on the basis of some assumptions. "*Business Intelligence (BI) as a concept and technology has significant potential in transforming data from distributed and heterogeneous sources into an integrated enterprise view for supporting organizational decision-making, management and strategic planning*." (Duan, Cao, One, & Woolley, 2013).

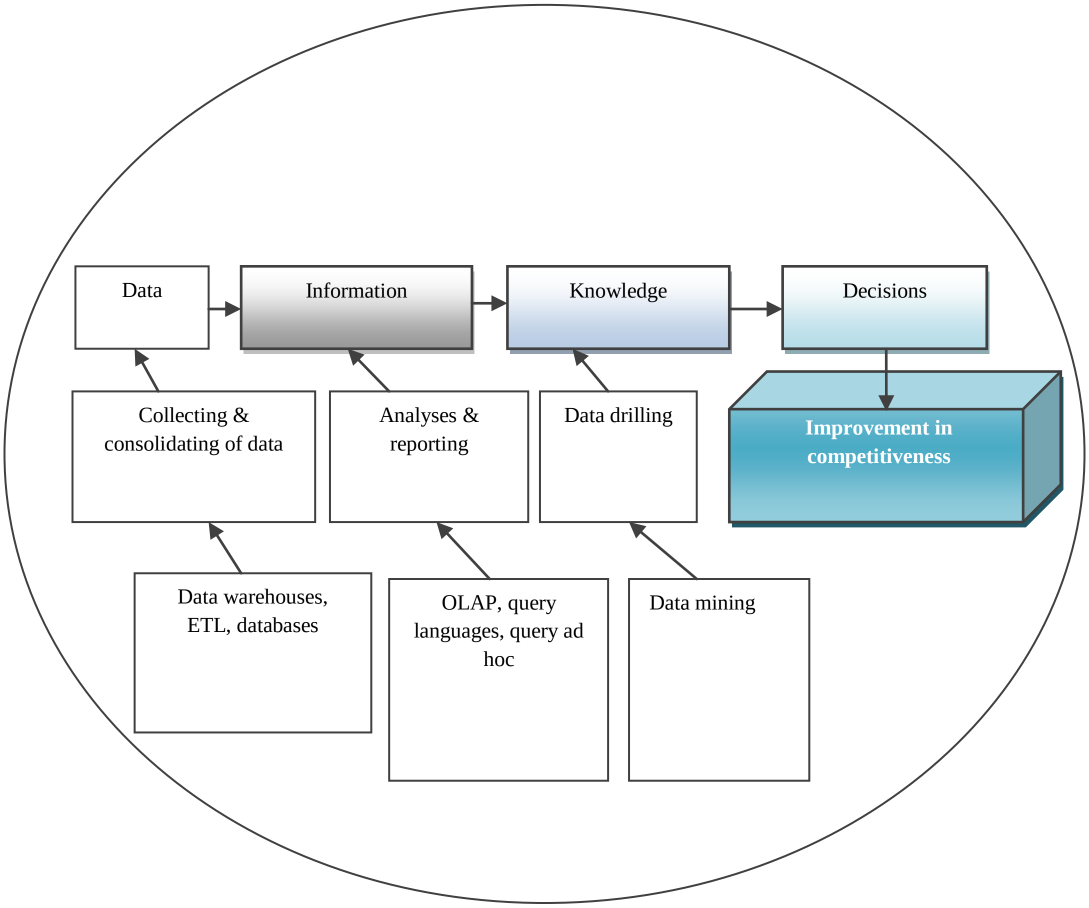

Data
Information
Knowledge
Decisions
Collecting & consolidating of data
Analyses & reporting
Data drilling
Improvement in competitiveness
Data warehouses, ETL, databases
OLAP, query languages, query ad hoc
Data mining

Figure 2: Role of Business Intelligence in Decision Making

Source: Olszak and Ziemba (2007)

2.3 The BI Three-Tier Computing Architecture

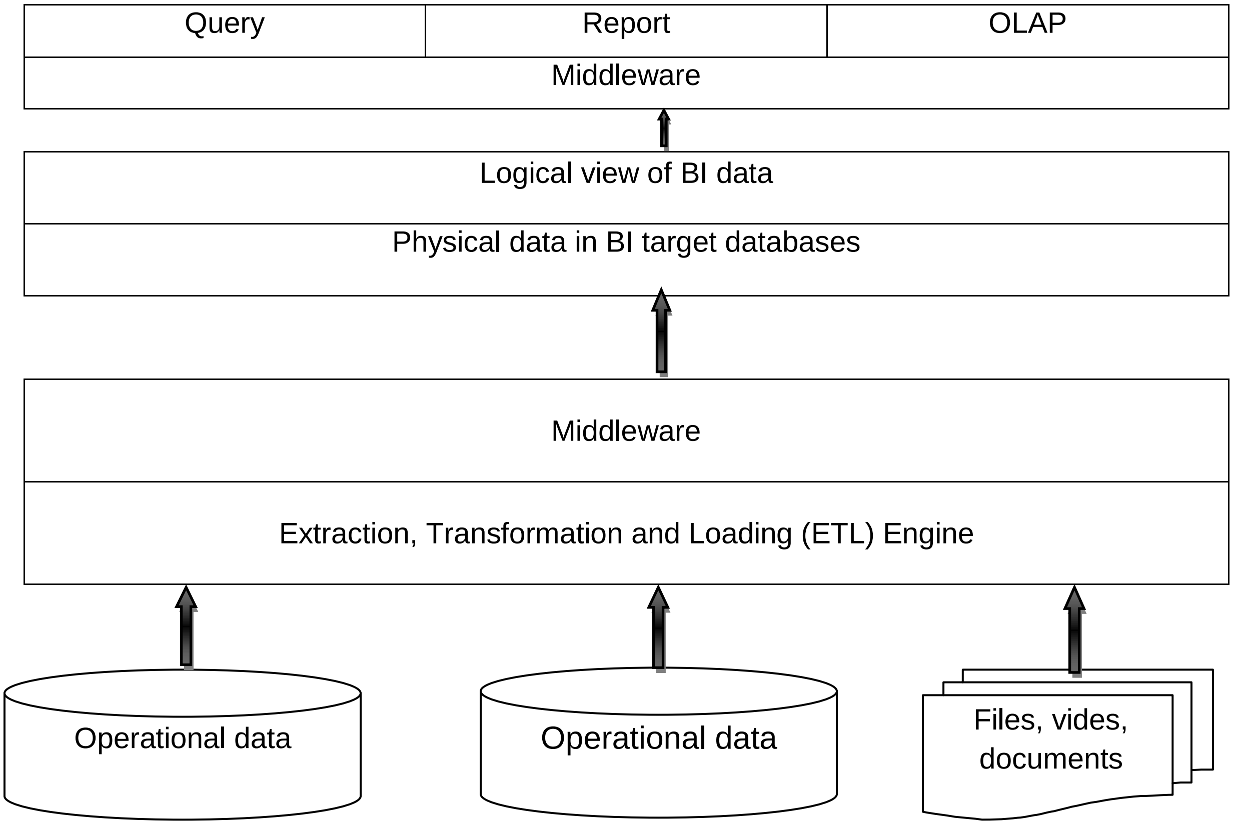


Figure 3: BI Three-Tier Computing Architecture

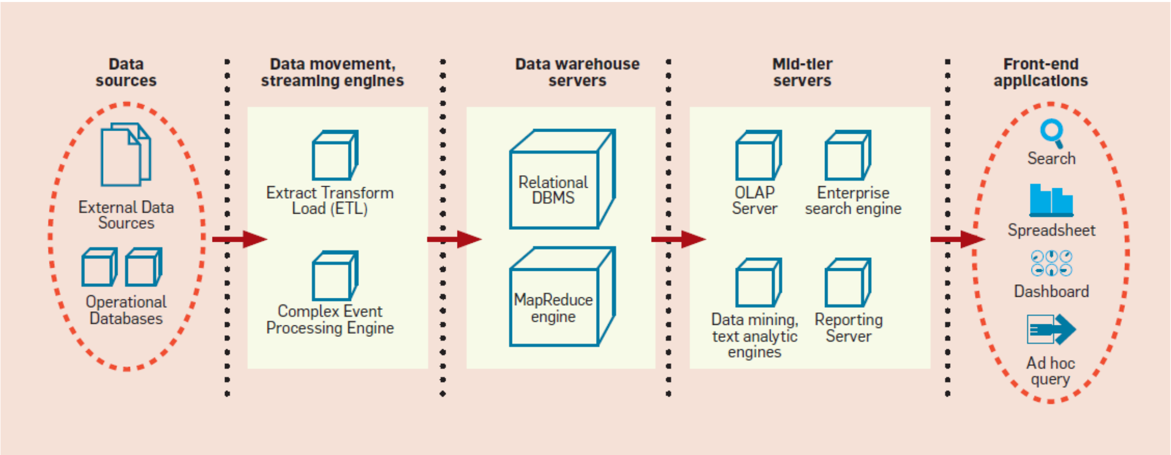


Figure 4: Traditional Business Intelligence Architecture

Source: Chaudhuri , Dayal & Narasayya (2011)

2.4 Hadoop Framework

Hadoop is an Apache open source framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models, designed to scale up from single servers to thousands of machines, each offering local computation and storage (Apache Hadoop, 2015). The framework relieves the hardware from delivering high-availability since the library itself is designed to detect and handle failures at the application layer; this ensures that the application is fault-tolerant. Basically, Apache Hadoop is a framework of open-source software for large-scale and storage processing on sets of data involving commodity hardware clusters.

Hadoop is composed of two main components:

1. Hadoop Distributed File System (HDFS)
2. MapReduce

2.4.1 Hadoop Distributed File System

According to Borthakur, (2008), the Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware having many similarities with the existing distributed file systems, however, it is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets.

There are two main types of machines in HDFS cluster namely *Namenode* and *Datanode*. However, a third one is usually called Secondary *Namenode*. Namenode is like the master machine; it controls all the metadata for the cluster, for example, what blocks make up a file and what datanodes those blocks are stored. Datanode is where HDFS actually stores the data blocks or chunks. Secondary namenode is not necessarily a backup of namenode, but, is a separate service that keeps a copy of both edit logs and file system image, merging them time to time in order to keep the size reasonable.

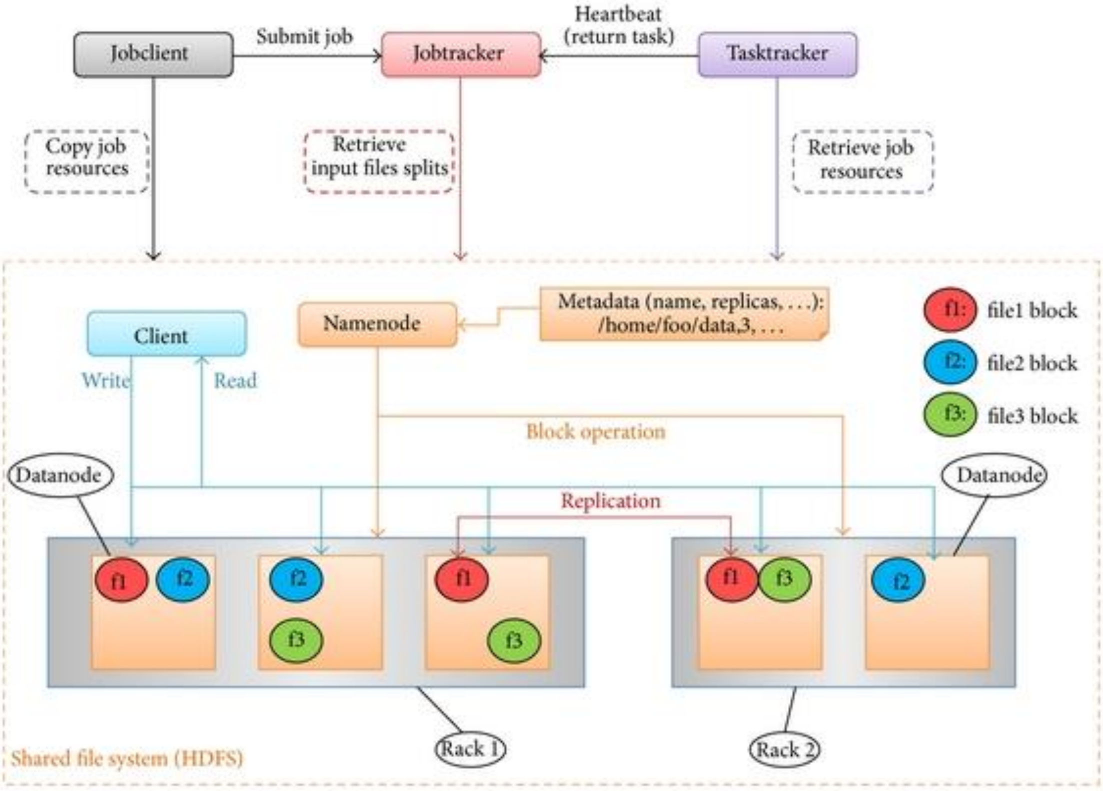


Figure 5: The Architecture of Hadoop Cluster and HDFS.

Source: Che-Lun and Guan-Jie (2014)

2.4.2 MapReduce

MapReduce is a programming model and an associated implementation for processing and generating large datasets that is amenable to a broad variety of real-world tasks (Dean & Sanjay, 2008). It can also be termed as a software framework for easily writing applications which process large amounts of data in parallel on large clusters (thousands of nodes) of commodity hardware in a reliable and fault-tolerant manner.

A Hadoop installation consists of a single master node and many worker nodes. The master, *JobTracker*, accepts jobs from clients, divide those jobs into tasks, then assigns those tasks to be executed by the worker nodes (Condie, Conway, Alvaro, & Hellerstein, 2010).

Execution of tasks assigned to worker nodes is managed by a *TaskTracker* process and each TaskTracker has a fixed number of slots for executing tasks, which by default, is usually two maps and two reduces.

2.4.3 Hadoop Ecosystem

Hadoop provides a reliable distributed storage through HDFS and an analysis system by MapReduce and was designed to scale up from a few servers to hundreds or thousands of computers, having a high degree of fault tolerance (Oancea & Dragoescu, 2014). Hadoop is now a de-facto standard in big data processing and storage, it provides unlimited scalability and is supported by major vendors in the software industry.

Hadoop includes an ecosystem of other projects built on top of HDFS and MapReduce in helping achieving certain operations on the platform. Some selected popular components include:

- **Hive:** This is a data warehouse system for Hadoop facilitating easy data summarization, ad-hoc queries and analysis of large datasets stored in HDFS. Hive has an SQL-like query language called Hive Query Language (HiveQL), which is used to issue query commands to Hadoop. Queries are executed in batch mode using MapReduce.
- **HBase:** This is basically a Hadoop Database. It is a distributed, column-oriented database that uses HDFS for underlying storage. HBase supports both batch operations using MapReduce and random queries.
- **Pig:** This is a high-level language used to analyze large datasets with its own language syntax for expressing data analysis programs and an infrastructure for analyzing these programs.
- **Mahout:** This is an extensive Java library for machine learning algorithms. It is used both in MapReduce and standalone programs. It can be applied in classification, prediction, clustering and recommender systems.
- **Sqoop:** This is a tool used in transferring large amounts of data to between Hadoop and Relational Database Management Systems (RDBMS). It is an abbreviation for **SQ**L to Had**oop.**

- **ZooKeeper:** is a centralized service to maintain configuration information, naming, providing distributed synchronization, and group services which are key to a variety of distributed systems.
- **HUE:** This is an abbreviation for Hadoop User Experience in. It is a web interface for Apache Hadoop that makes common tasks, for example, running MapReduce jobs, browsing HDFS and creating Apache Oozie workflows easier.
- **Impala:** This is a project still under Hadoop but created by Cloudera, Hadoop vendor, to provide low-latency queries over Hive. Impala queries do not execute MapReduce jobs hence faster that Hive queries.

2.5 R Statistical Software

R is a language and environment for statistical computing and graphics, but is also a GNU project similar to the S language and environment which was developed at Bell Laboratories by John Chambers and colleagues (The R Project for Statistical Computing, 2015). It is open source statistical software with rich library of packages for data analysis and huge community base. It can be considered as a suite of software and programming language for the purpose of data visualization, statistical computations and analysis of data.

R fall short when it comes to analyzing large data since the data size that can be manipulated by R is dictated by the random access memory (RAM) of the computer hosting R package. Hadoop and R are quite complementary in terms of visualization and analytics of big data.

2.6 Ways Hadoop and R Work Together

Generally, integration with Hadoop can be done using RHadoop, rJava, RImpala, JDBC and ODBC library packages. Hence, it is possible to perform data analysis using R inside Hadoop. The possible different ways of using Hadoop and R together are as follows:

- **Hadoop Streaming**: This is a utility that lets users run and develop the Map Reduce program in languages apart from Java from which Hadoop is programmed. It was developed by David Rosenberg. Hadoop streaming are utilities available as R scripts that make it easy to use for R users.

- **ORCH**: This can be used on the non-Oracle Hadoop clusters or on the Oracle Big Data Appliance. ORCH is a Hadoop Oracle R connector.
- **RHIPE**: These are techniques designed for analyzing large sets of data. RHIPE stands for R and Hadoop Integrated Programming Environment.
- **RHadoop**: Provided by Revolution Analytics, RHadoop is a great solution for open source Hadoop and R. RHadoop is bundles with 4 primary packages of R to analyze and manage Hadoop Framework data which are *rhdfs*, *rhbase*, *rmr2* and *plyrmr*.

2.7 Business Intelligence in Higher Learning Institutions

In order to promote students' success, learning institutions require the implementation of processes and mechanisms that allows the close monitoring of the students' academic activities (Piedade & Santos, 2010). Piedade & Santos (2010) further observe that even though essential, the activities involved in this complex process do not take place in many higher education institutions due to the lack of appropriate practices and an adequate technological support to sustain these practices.

Dimokas, Mittas, Nanopoulos, & Angelis (2008), observe that universities are encountering growing demands by legislators and communities who are clamoring for valuable information about student achievement and university system accountability. The existing management information systems in universities are often designed for specific management applications, and there are still many problems and shortcomings on data analysis and decision support (Xuejian & Li, 2011).

Piedade & Yasmina (2009) claim that closely monitoring of the students' academic activities, the evaluation of their academic success and the approximation to their day-by-day academic activities are key factors in promotion of the students' academic success in higher education institutions. To make possible the implementation of monitoring processes and activities, it is essential to acquire knowledge about the students and their academic behaviours. This knowledge supports the decision-making associated with teaching-learning process, enhancing an effective institution-student relationship. The knowledge can be provided by BI systems which consolidate all student data and provide a holistic view.

2.8 Business Intelligence Roadmap

Almost every kind of engineering project; be it structural engineering or software engineering project, goes through six stages between inception and implementation (Moss & Atre, 2003). Business Intelligence systems development, being one of the projects, follows broadly these stages.

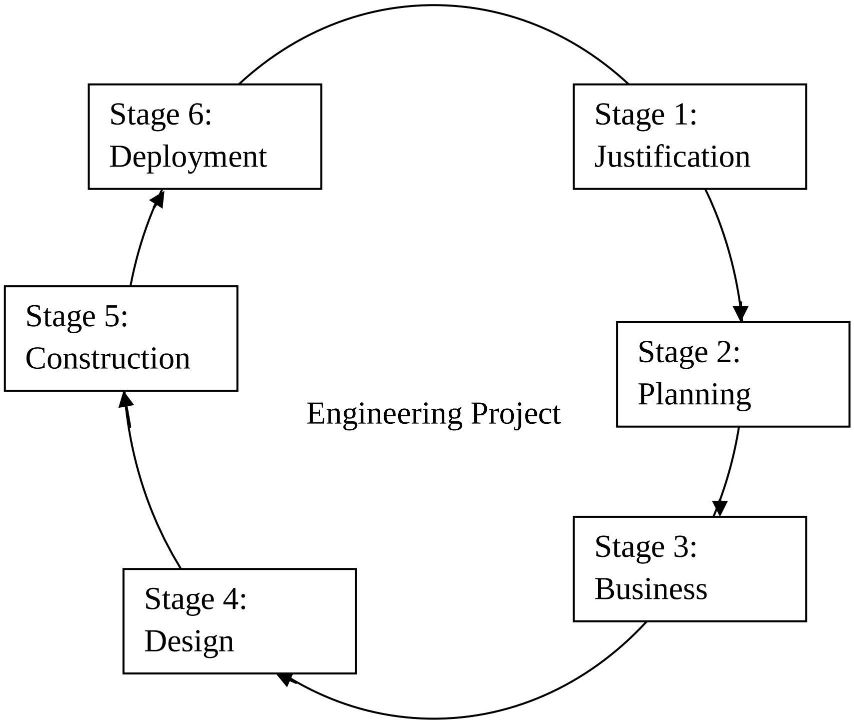


Figure 6: Project Engineering Stages

Source: Moss & Atre (2003)

These six (6) broad stages are broken down into sixteen (16) steps by Moss & Atre (2003) to achieve BI engineering stages and step development.

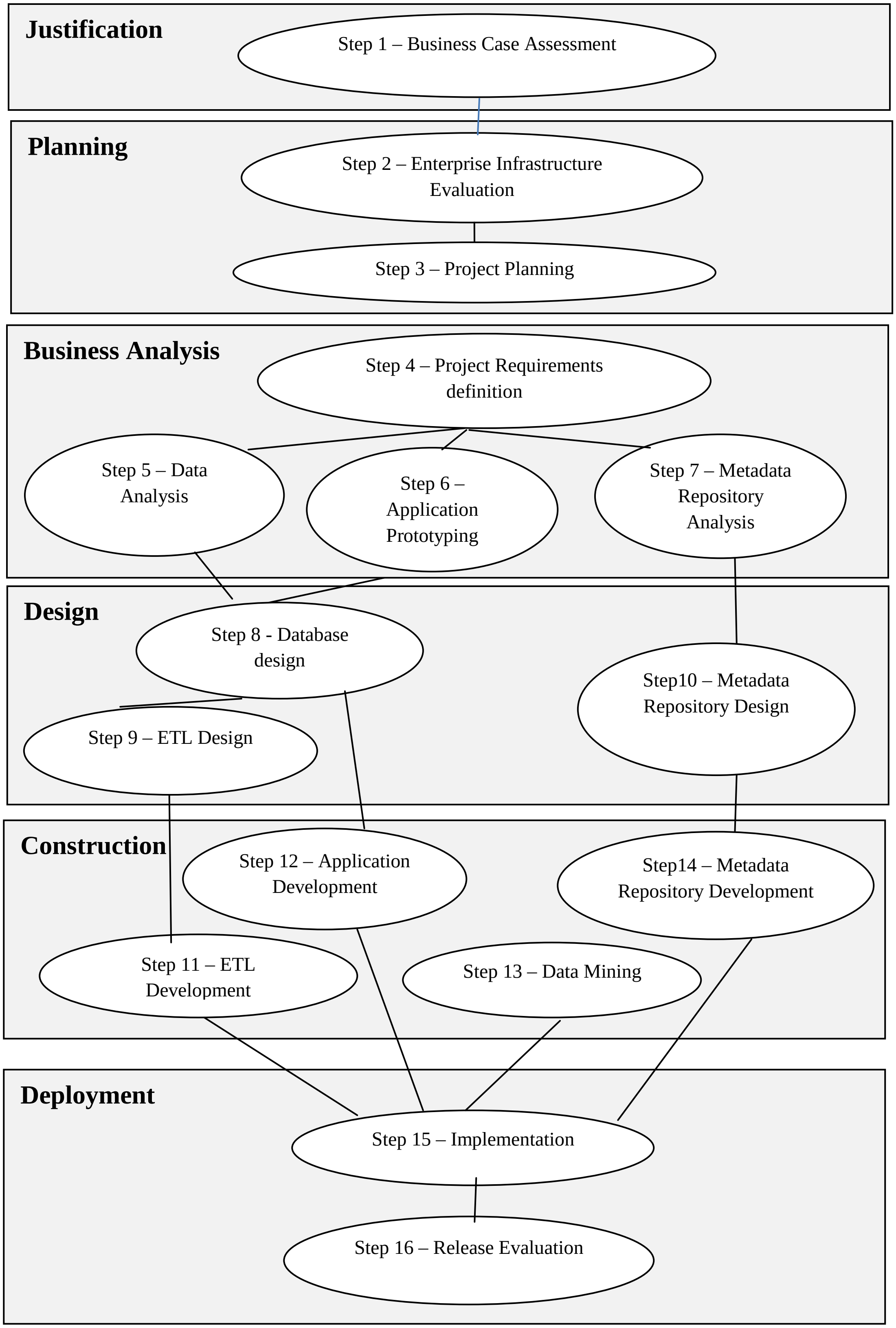

Justification
Step 1 – Business Case Assessment
Planning
Step 2 – Enterprise Infrastructure Evaluation
Step 3 – Project Planning
Business Analysis
Step 4 – Project Requirements definition
Step 5 – Data Analysis
Step 6 – Application Prototyping
Step 7 – Metadata Repository Analysis
Design
Step 8 - Database design
Step10 – Metadata Repository Design
Step 9 – ETL Design
Construction
Step 12 – Application Development
Step14 – Metadata Repository Development
Step 11 – ETL Development
Step 13 – Data Mining
Deployment
Step 15 – Implementation
Step 16 – Release Evaluation

Figure 7: BI Engineering Stages and Step Development

Source: Moss & Atre (2003)

2.9 Business Intelligence Implementation in Higher Learning Institutions

Business Intelligence is a new phenomenon in the academic environment especially in Kenya. However, some universities have taken a leap in incorporating BI in their management, processes and decision-making. While researching on BI implementation, the following institutions of higher learning were identified. The study reviewed how BI was implemented in these various institutions of higher learning globally. Special focus is pegged on a public University in Kenya where a researcher implemented single multi-dimensional cube using Microsoft Excel.

2.9.1 Implementing Data Warehousing and Business Intelligence at McMaster University Using the SAS Intelligence Value Chain

McMaster University (2005) claims that like many other organizations, McMaster's transactional applications do not store data in easily accessible data models that can be transformed into comprehensive, meaningful information to support evidenced-based decision-making. This leads to the creation of the research funding data mart and deployment of Business Intelligence tools was a critical first step in creating a "University without boundaries". The continued development of the Data Warehouse and increased deployment of query and reporting tools through the decision-support portal will provide the campus community with a greater capacity for data analysis, interpretation, flexible reporting to support decision-making and facilitate greater accountability in an environment of devolved authority.

Belaire, Matthews, McInnis, Scime, & Weisensee (2005) review incremental development of an enterprise-wide data warehouse at McMaster University, along with the deployment of end-user, web-based query and reporting tools using SAS Value Chain Analytics.

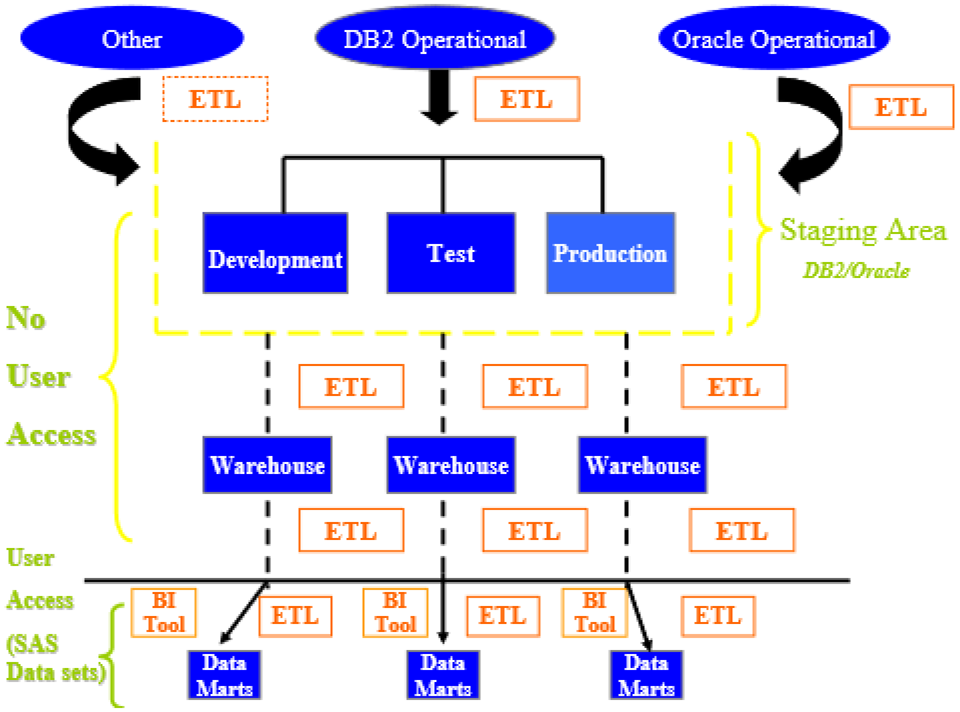


Figure 8: McMaster University BI Architecture

Source: McMaster University (2005)

2.9.2 Business Intelligence Implementation at Maastricht University

This was provided by Henny Claessens during International HERUG conference 2008 on 28 March 2008 in the University of Cape Town South Africa.

The major advantages the University experienced after the BI implementation include:

- Flexible reporting capabilities due to multiple sources and multidimensional business analytics Complex reporting capabilities, in ‘close to real-time’ mode
- One version of the truth due to data warehouse and coherent reporting the same numbers(counting/accounting issue)
- Less effort required from end users since due to improved user friendly interfaces, web based reporting, and self-explanatory reports

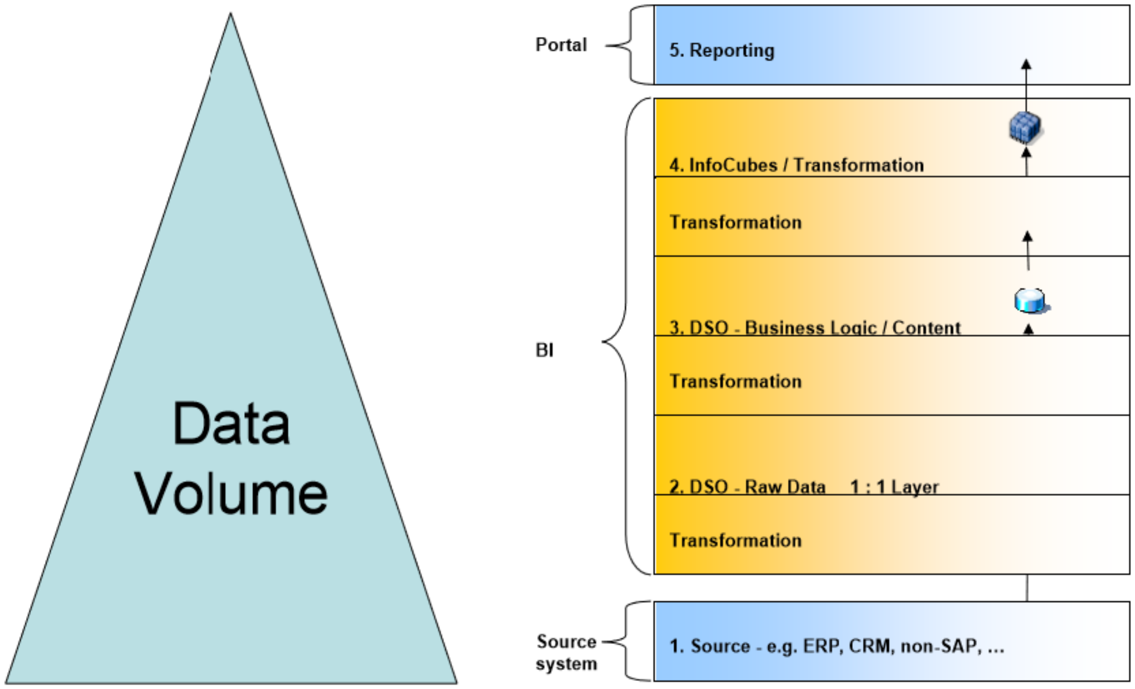

Data Volume
Portal
5. Reporting
BI
4. InfoCubes / Transformation
Transformation
3. DSO - Business Logic / Content
Transformation
2. DSO - Raw Data 1 : 1 Layer
Transformation
Source system
1. Source - e.g. ERP, CRM, non-SAP, ...

Figure 9: Maastricht University SAP BI Architecture

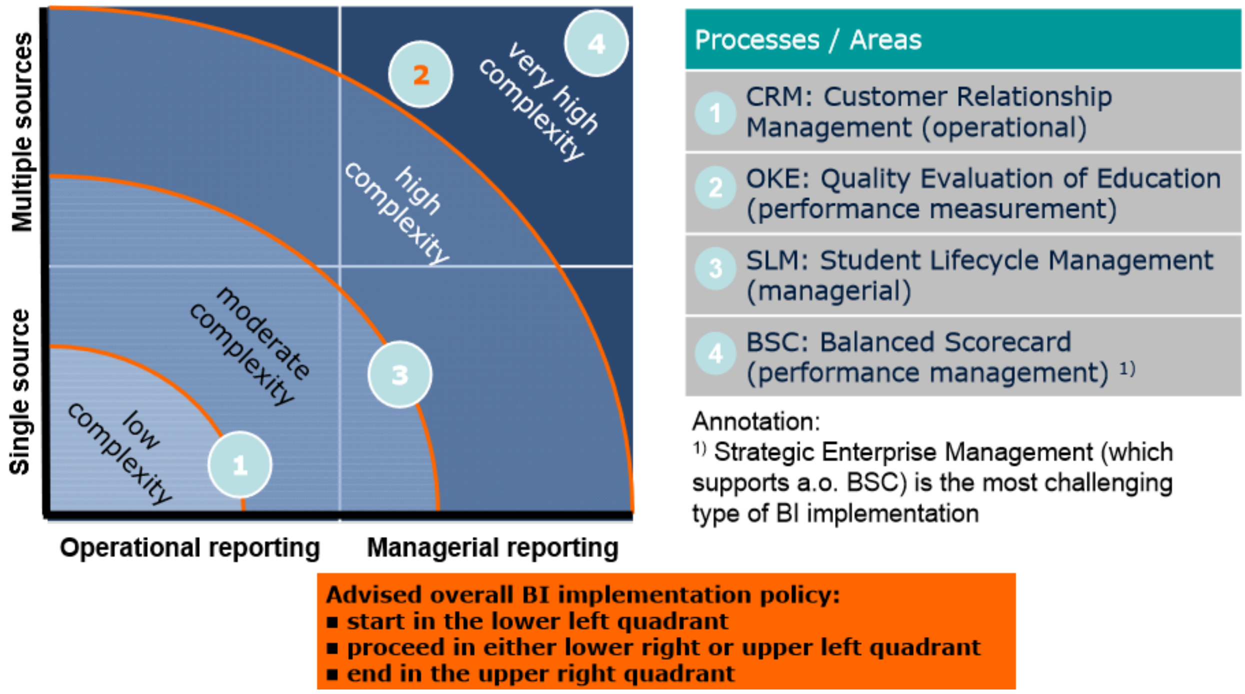
Source: Claessens (2008)

Implementation approach

The implementation approach used was Incremental. They believed in thinking big but starting small. They were able to focus on the overall BI strategy by breaking it down into smaller increments:

- Increment1: BI for ‘CRM’ (Student Recruitment, Student Admissions)
- Increment2: BI for‘ Quality Evaluation of Education’
- Increment3: BI for ‘SLM’

‘BI Quadrant’ Implementation Policy



Processes / Areas
1 CRM: Customer Relationship Management (operational)
2 OKE: Quality Evaluation of Education (performance measurement)
3 SLM: Student Lifecycle Management (managerial)
4 BSC: Balanced Scorecard (performance management) [1)]

Annotation:
[1)] Strategic Enterprise Management (which supports a.o. BSC) is the most challenging type of BI implementation

Advised overall BI implementation policy:
- **start in the lower left quadrant**
- **proceed in either lower right or upper left quadrant**
- **end in the upper right quadrant**

Figure 10: BI Quadrant Implementation Policy

2.9.3 Business Intelligence at the University of Minnesota

University of Minnesota required BI system for:

- Student recruitment and retention
- Course scheduling and support
- Tracking Purchasing, that is, analysis unit spending by item/ category
- Employee satisfaction

The BI Tools used were:

Oracle Business Intelligence Enterprise Edition (OBIEE) - Oracle BI Toolset providing capabilities that are to produce but easy to use.

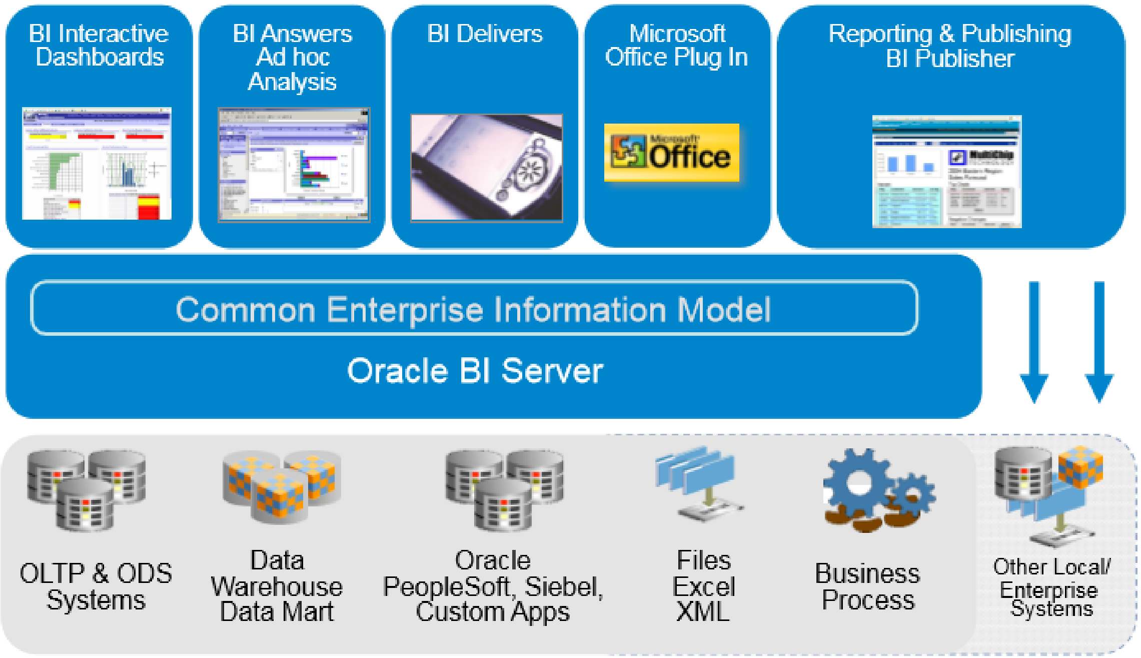

BI Interactive Dashboards
BI Answers Ad hoc Analysis
BI Delivers
Microsoft Office Plug In
Office
Reporting & Publishing BI Publisher
Common Enterprise Information Model
Oracle BI Server
OLTP & ODS Systems
Data Warehouse Data Mart
Oracle PeopleSoft, Siebel, Custom Apps
Files Excel XML
Business Process
Other Local/ Enterprise Systems

Figure 11: University of Minnesota BI Architecture

2.9.4 Data Warehousing: A Case of Public University in Kenya

Amurgat (2012), in an attempt to determine the level of interdependency between top decision-makers and information technology (IT) personnel of a Kenyan public university in accessing and analyzing and reporting data, managed to implement a single multi-dimensional cube using Microsoft Excel. The researcher found out that Excel could easily be used as a BI tool in the institution.

2.10 Open Source BI Tools

Back then in 2009, closed and commercial tools had been dominating the BI market (Golfareli, 2009) with only limited adoption of open source tools. However lately, there has been an increasing popularity of open source initiative in BI; Wise (2012) claims that open source BI solutions have many benefits over traditional proprietary software, from offering lower initial costs to more flexible support and integration options.

A look at different Business Intelligence delivery methods brings into perspective various solution types and how they fit within the overall BI market.

Table 1: BI Delivery Type Breakdown

Source: Wise (2012)

BI Type	**Definition**	**Solution Parameters**
Traditional	Business Intelligence is installed and developed at the customer site, with the general purpose of reporting and analytics using historical data sets.	Organizations can use one BI component or many. This may include a data warehouse, interactive reports, analytics, and/or dashboards to provide a diverse access point to information analysis
Software as a Service (SaaS)	BI offerings or components that are hosted and offered as a service to organizations through online access.	Expands all areas of BI – with data warehouse solutions being called data as a service (DaaS); most offerings providing dashboards and analytics are targeted to specific industries or business areas.
Cloud	Similar to SaaS based on the fact that solutions/data are hosted externally to the organization. In many cases, organizations develop and maintain their own BI applications.	Many organizations choose to have portions of their data hosted within a public or private cloud. BI vendors are beginning to provide this option to their customers.
Operational	Similar to traditional BI in terms of delivery but focuses on real-time or continuous business visibility.	Operational BI (OBI) requires a specific infrastructure that enables continual data updates to feed into front-end applications
Open Source	Similar to traditional BI delivery and development but uses free source code as the basis for development.	Both community (free) and commercial versions exist and span all components of a BI environment.

2.10.1 Open Source versus Commercial Tools

This section brings out the merits and demerits of using open source and commercial BI tools.

The most relevant benefit of using open source tools is that they are free and allow access to the source code, with the possible modification of the various modules. Again, unlike the open source tools, commercial tools are paid and often represent an extra expense, unbearable for the vast majority of organizations especially Universities. One of the advantages of open source platforms are that if they do not serve the needs of an organization, then they can be replaced by other platforms without a cost; but this does not happen with commercial platforms (Madureira, 2012). The open source tools generally require lower system requirements than commercial applications and this is justified in assuming that those who cannot invest in software may not invest in hardware.

The demerits of open source against commercial platforms are elaborated. The open source projects are in constant development, which means that some projects have not yet reached their desired maturity and the fact that open source tools are developed by a community of contributors raises some concerns about product quality (Tereso & Bernardino, 2011). Bernardino & Tereso (2013) observe that cases of lack of development of appropriate methods and quality standards can be a reality in a community of open source development; stating that the open source tools are usually seen as less reliable, given that the testing process is typically narrow in scope, lack of comprehensive documentation, the continuity of the open source projects is not always guaranteed.

Some of the demerits of commercial BI platforms are: high acquisition costs, the requirement to be connected to the sellers, typically require more powerful hardware and the difficulty of transition to other platforms, taking into account the costs and terms of the contract (Bernardino & Tereso, 2013).

2.10.2 Selected Commercial BI Tools

This section describes some of the selected commercial BI tools. The four major commercial business intelligence platforms, according to the study of Gartner (2011) are IBM Cognos, Microsoft BI, MicroStrategy and Oracle BI. SAS will also be considered.

2.10.2.1 IBM Cognos

IBM Cognos resulted from the acquisition of Cognos Company by IBM and is its business intelligence platform. IBM Cognos BI platform allows: query and reporting, OLAP analysis, scorecarding, dashboarding, real-time monitoring, statistics, planning and budgeting, extending BI, collaborative BI and importing/exporting data. Cognos has also a mobile version, where we can use its benefits anywhere at any time online or offline. The mobile version was supported by: Apple iPhone and iPad; RIM BlackBerry smart phones and PlayBook. The mobile version has support for devices using the Android, Symbian and Windows Mobile operating systems.

2.10.2.2 Microsoft BI

According to Gartner (2011), Microsoft emerged as the leading tool of Business Intelligence tools market. In the business intelligence processes Microsoft uses two of its solutions packages: Microsoft Office and Microsoft SQL Server. SQL Server supports features of Data Warehouse, Data Marts and Operational Data while data integration is supported by SQL Integration Services; SQL Analysis Services performs data analysis and SQL Reporting Services supports operations with reports. Excel supports the features of end-user analysis tool and the business intelligence portal uses SharePoint. The business scorecards, analytics and planning tools are supported by PerformancePoint. This platform allows: querying, analysis, reporting, data integration, synchronization, searches, cloud storage and importing/exporting data. Webalo Inc developed an application that supports the Microsoft BI, Mobile Webalo Dashboard. This application functions as a hosted service. This application allows transforming the SQL Server, Excel, Performance-Point and SharePoint contents in Dashboards. This service has the benefit not be developed for a specific device but for a range of devices that support operating systems. The mobile devices compatible are: RIM BlackBerry, Windows Mobile 2003 for Pocket PC, Windows Mobile 5, Windows Mobile 6 and Java MIDP 2.0 smart phones.

2.10.2.3 MicroStrategy

According to Gartner (2011), MicroStrategy holds the third podium position of commercial business intelligence platforms market leaders. Its version 9 was available in optimized solutions for mobile environments, including iPad, iPhone and BlackBerry. The main features of the MicroStrategy are: querying, reporting, OLAP analysis (allows the analysis and construction of an intelligent data cubes, MDX connection and incorporates a ROLAP server), dashboarding, data mining and importing/exporting data. Its mobile version is developed around the Apple applications (iPhone and iPad) but is supported by other devices as BlackBerry smart phones. MicroStrategy provides a free mobile suite.

2.10.2.4 Oracle BI

According to Gartner (2011), Oracle BI holds the second podium position of commercial business intelligence platforms market leaders. Oracle BI is available in the market as package commonly known as Oracle Business Intelligence Enterprise Edition (OBIEE). OBIEE offers a range of solutions necessary to create and view reports. Oracle Business Intelligence Enterprise Edition allows: to create interactive dashboards, reporting and publishing, answers (ad-hoc analysis), delivers (proactive detection and alerts), disconnected analytics, Microsoft Office Plug-in and Web Services. The OBIEE provides a mobile version supporting BlackBerry, iPhone, HTC, Android, Windows Mobile, Symbian and Palm.

2.10.2.5 SAS

Statistical Analysis Software (SAS) is a software suite developed by SAS Institute for advanced analytics, business intelligence, data management, and predictive analytics. In terms of advanced analytics, it is the largest market-share holder. SAS programs have two steps; a DATA step, which retrieves and manipulates data, usually creating a SAS data set, and a PROC step, which analyzes the data.

The SAS software suite has more than 200 components. Some of the SAS components include:

- Base SAS - Basic procedures and data management
- SAS/STAT - Statistical analysis
- SAS/GRAPH - Graphics and presentation
- SAS/OR - Operations research
- SAS/ETS - Econometrics and Time Series Analysis
- SAS/IML - Interactive matrix language
- SAS/AF - Applications facility
- SAS/QC - Quality control
- SAS/INSIGHT - Data mining
- SAS/PH - Clinical trial analysis
- Enterprise Miner - data mining

2.10.3 Selected Open Source BI Tools

The most common open source BI platforms are Actuate, JasperSoft, OpenBI, Palo, Pentaho, SpagoBI and Vanilla (Bernardino & Figueiredo, 2014). This part presents a brief description of the potential of these BI platforms. The analysis platforms have one feature in common, that is, all have a GNU General Public License (GNU GPL).

Selective analysis the features of each of the four major open source business intelligence platforms are independently elaborated. The open source business intelligence tools have followed the evolution of the commercial tools. Developed by non-profit open source communities, these tools present good business intelligence features.

2.10.1 JasperSoft

JasperSoft is an open source business intelligence platform developed in Java and Perl programming languages. It runs on both Windows and Linux environments. JasperSoft has two proprietary versions, Community and Enterprise and Professional. The Community is a free version that is available in three individual modules: Jasper Reports Server, JasperSoft

OLAP and ETL JasperSoft. The Jasper Reports Server module provides the creation, structuring and displaying BI reports; JasperSoft OLAP module is the tool that allows linking of data and parameterization of the different views of the data cube, while ETL JasperSoft is the module that allows the tool creating the structure of production, processing and loading data.

The set of modules that make up the community version of the tool features available for creating and viewing reports, various graphics, dashboards, OLAP and ETL processes, analysis of geo-referencing, export data and ad-hoc queries.

2.10.2 Pentaho

Pentaho was developed in Java programming language and can run on Windows, Unix and Linux platforms. It is also available in two versions, Community and Enterprise. Like other platforms, the Community version is free and the Enterprise version is a paid solution with more features. Pentaho incorporates the Kettle application for ETL processes, Mondrian and Ramsetcube in OLAP processes, and Weka for properties of data mining. Pentaho allows creating reports, creating charts, dashboards, data manipulation processes through ETL and OLAP, data mining processes, KPI's, exports of data and ad-hoc queries. Pentaho uses MDX to access and manipulate data from databases. Pentaho provides the ad hoc analytical reporting for iPad. The mobile version allows creating the reports directly from the iPad. The touch-enabled technology provides more information about reports from user.

10.2.3 SpagoBI

SpagoBI is an open source business intelligence platform also developed in Java. It is available for Linux, Windows and UNIX operating systems. However, unlike other platforms, SpagoBI is available in a single Community version. This makes it the most comprehensive free platform, because has all the main features of business intelligence and provides them as a single release covering its users with a good solution that offers all its capabilities in a package completely free. SpagoBI contains features for creating and exporting reports, charting several, development of dashboards, data integration using ETL processes, analyzes, processes using OLAP, data mining processes, implementation of filters in queries, export data, ad-hoc queries and features GEO/GIS. The SpagoBI provide a mobile version for platforms Apple and Android support. This version allows remote connection with SpagoBI Server. This mobile version include: user authentication; documents selection, according to end user's role, and consequent controlled data download; data update on request; and periodical and automatic check to identify possible alarms and notifications.

10.2.4 Vanilla

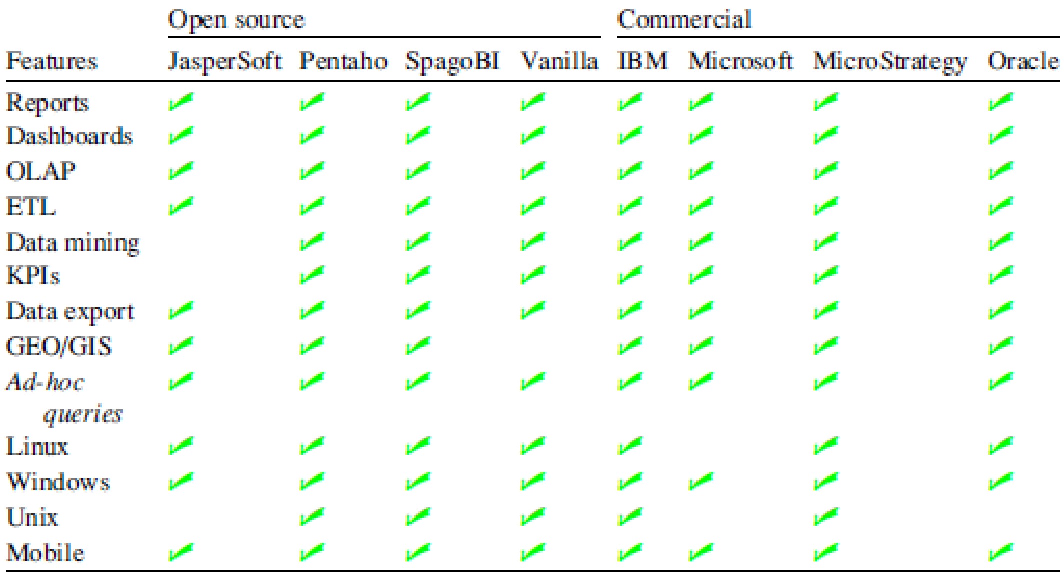
Vanilla is an open source business intelligence platform developed in PHP language and is available in a single version, community. Vanilla is a complete platform since it integrates all main functionalities of business intelligence. The platform allows to create reports, perform analysis, generate and analyze data tables, charting several, dashboards, ad-hoc queries, integrate OLAP and ETL processes, define KPIs, data export and use of procedures of Data Mining. BIRT and iReports are used for the purpose of metadata integration. Vanilla allows filing several separate projects, and an historical book. The platform is available in a version for mobile operating system Android. Vanilla is available for Linux, Windows and UNIX operating systems. Vanilla was the first open source tool to provide a mobile version, in version 3.4, where is possible to browse reports and run dynamic reports from smart phones. It is also possible to browse OLAP cubes.

2.10.4 Comparison between Commercial and Open Source BI Tools

In this section we intend to illustrate a comparative table of the different features offered by each platform. Reports are information elements such as information tables and several charts. Dashboards are graphical elements that allow view the data in graphical mode. The On-Line Analytical Processing (OLAP) is the process that allows the visualization of information in data cubes. ETL is the process in charge of data extract, data transformation and data load. Data mining is a complex process of extract information by data sequences. The KPIs are the key performance indicators that usually arise associated with the data represented. The data export allows export data to Excel, CSV files etc. GEO/GIS describes any information system that displays geographic information for informing decision-making. Ad-hoc queries are the typical queries made by decision makers where any field can be queried at any time. The results of this work on the eight BI platforms are shown in table below.

Table 2: Comparison between Open Source and Commercial BI Tools

Source: Bernardino and Tereso (2012)



	Open source				Commercial			
Features	JasperSoft	Pentaho	SpagoBI	Vanilla	IBM	Microsoft	MicroStrategy	Oracle
Reports	✓	✓	✓	✓	✓	✓	✓	✓
Dashboards	✓	✓	✓	✓	✓	✓	✓	✓
OLAP	✓	✓	✓	✓	✓	✓	✓	✓
ETL	✓	✓	✓	✓	✓	✓	✓	✓
Data mining		✓	✓	✓	✓	✓	✓	✓
KPIs		✓	✓	✓	✓	✓	✓	✓
Data export	✓	✓	✓	✓	✓	✓	✓	✓
GEO/GIS	✓	✓	✓		✓	✓	✓	✓
Ad-hoc queries	✓	✓	✓	✓	✓	✓	✓	✓
Linux	✓	✓	✓	✓	✓		✓	✓
Windows	✓	✓	✓	✓	✓	✓	✓	✓
Unix		✓	✓	✓	✓		✓	
Mobile	✓	✓	✓	✓	✓	✓	✓	✓

2.11 Conceptual Framework

This project was aimed at developing a Business Intelligence system in a university setup. The proposed conceptual framework adopted was geared towards addressing the gaps identified. With the high cost of current commercial BI solutions and complexity of the implementation process, the BI system was expected to have the following features:

Open source: the prototype was developed using open source tools. These tools included Hadoop components; Hadoop Distributed File System (HDFS), Apache Hive, Impala, Apache Solr, Apache Pig, Apache Mahout and Hadoop User Experience (HUE), Python, R Statistical Software, and Python Django framework.

Data warehouse/ BI Methodology: This project adopted Ralph's Kimball Dimensional Modeling. This is a bottom-up design approach where data marts facilitating reports and analysis are created first; after which these data marts are combined to create a broader enterprise data warehouse.

Reporting: Different reporting facilities would be provided for example standard reports, ad hoc reports, drill down, slicing and dicing.

Dashboards/Scoreboards: Dashboard would show the current status of the University based on the set targets. It would reveal whether a University is geared towards the strategic objective or not. Scoreboard would indicate how far the University has gone based on the set targets.

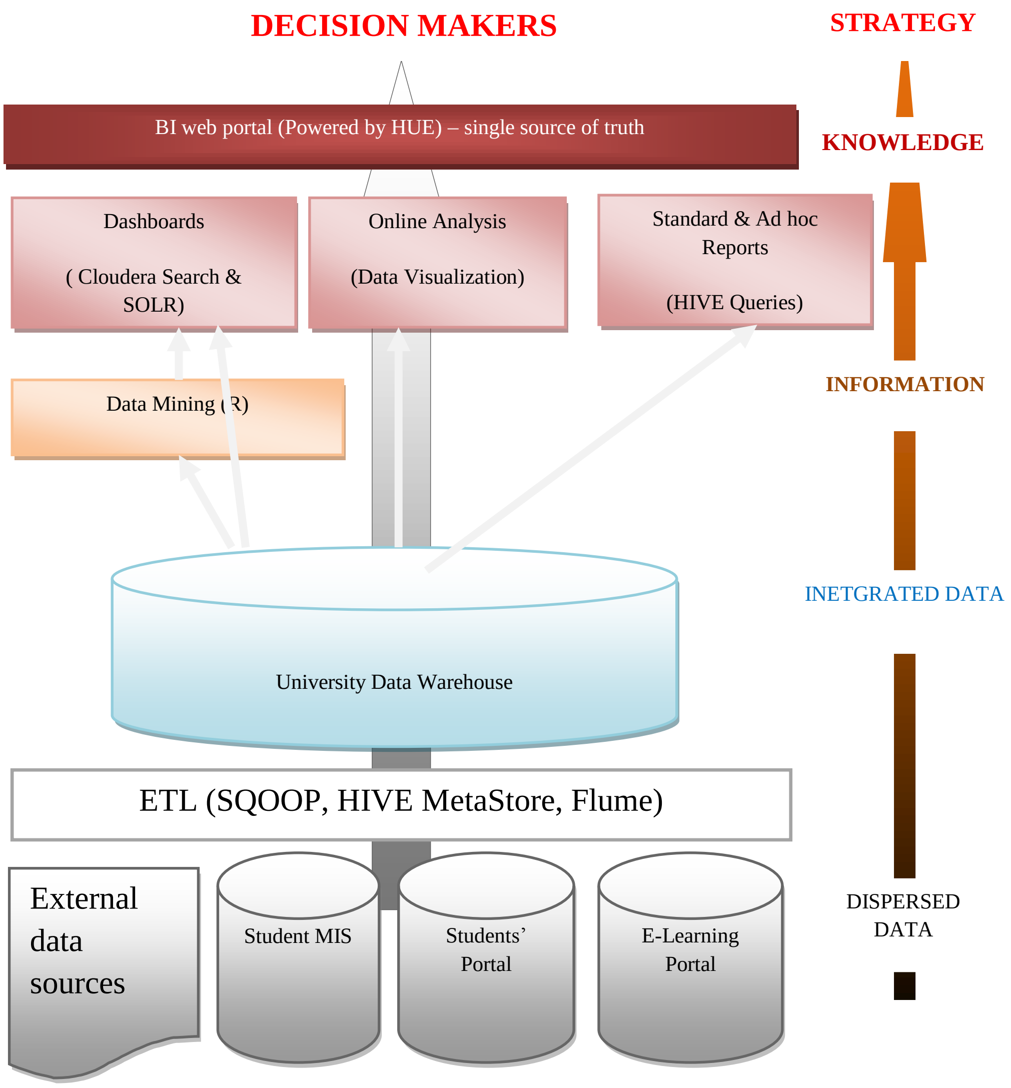


Figure 12: The Proposed Business Intelligence System Architecture

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This section describes step-by-step approach that was used to implement the BI system.

3.2 Study Area

This study was focused around the Technical University of Kenya. The methodology of building a data warehouse is elaborated.

3.3 Kimball's Bottom-up Dimension Modeling

The approach to designing a data warehouse/ BI depends on the business objectives of an organization, nature of business, time and cost involved, and the level of dependencies between various functions (Kimball & Ross, 2011). In a university setup, it is easy to create data marts for different departments first then combine the individual data marts to form the enterprise data warehouse.

3.4 Business Requirements Gathering

In order to capture business requirements, operational managers of one university in Kenya, The Technical University of Kenya, were interviewed. These managers were Academic Registrar, Examination Officer and Admissions Officer. The interview aimed to assess the level of reporting, awareness of data warehouse role in the academic intuition and the challenges that would be encountered in the implementation of BI in the institution. The following were identified as business questions they were unable to answer:

1. What will be the number of students enrolled in the coming academic years?
2. Which students are at the risk of being discontinued due to performance?
3. Who are the fee defaulters and are still continuing with learning?
4. Which department is the heaviest consumer of examinations materials?

In order to address the above questions, typical university processes relating to students only were identified as follows:

- ***Application:*** is the process of advertising of courses offered and choosing of courses by new entrants (applicants) for a university.
- ***Selection:*** is the process of short listing of applicants who qualify for courses chosen or applied for.
- ***Admission:*** is the process of admitting selected applicants into a university.
- ***Deferment:*** is the process of allowing already admitted students to defer or postpone their study.
- ***Transfer:*** this is the process of changing a course during study.
- ***Registration:*** is the process of registering for study during each given academic calendar enabling a university to know the total number of active students for each course at any given period.
- ***Fee Invoicing:*** is the process of assigning fees to be paid by students based on course, year of study and university calendar.
- ***Fee Payment:*** is the process of students paying for fees invoiced at any given academic period.
- ***Class Attendance:*** is the process of attending lessons by students. Examinations are based on class attendance and contact hours.
- ***Library Resource Usage:*** this entails process of student borrowing book, journal, e-book, or any library material for academic use.
- ***Examination:*** this comprises of student sitting for examinations related to courses being pursued.
- ***Graduation:*** is the process of student being conferred with various qualifications for course pursued during a graduation ceremony.
- ***Alumni Registration:*** is the process of registering former students of a university into an alumni association.

3.5 Steps to Getting the Dimension Model

The three (3) steps that guided the modeling of the university warehouse are:

1. Build a Data Warehouse Bus Matrix
2. Diagram the Data Warehouse Bus Matrix
3. Fill in the attributes associated with the processes and dimensions

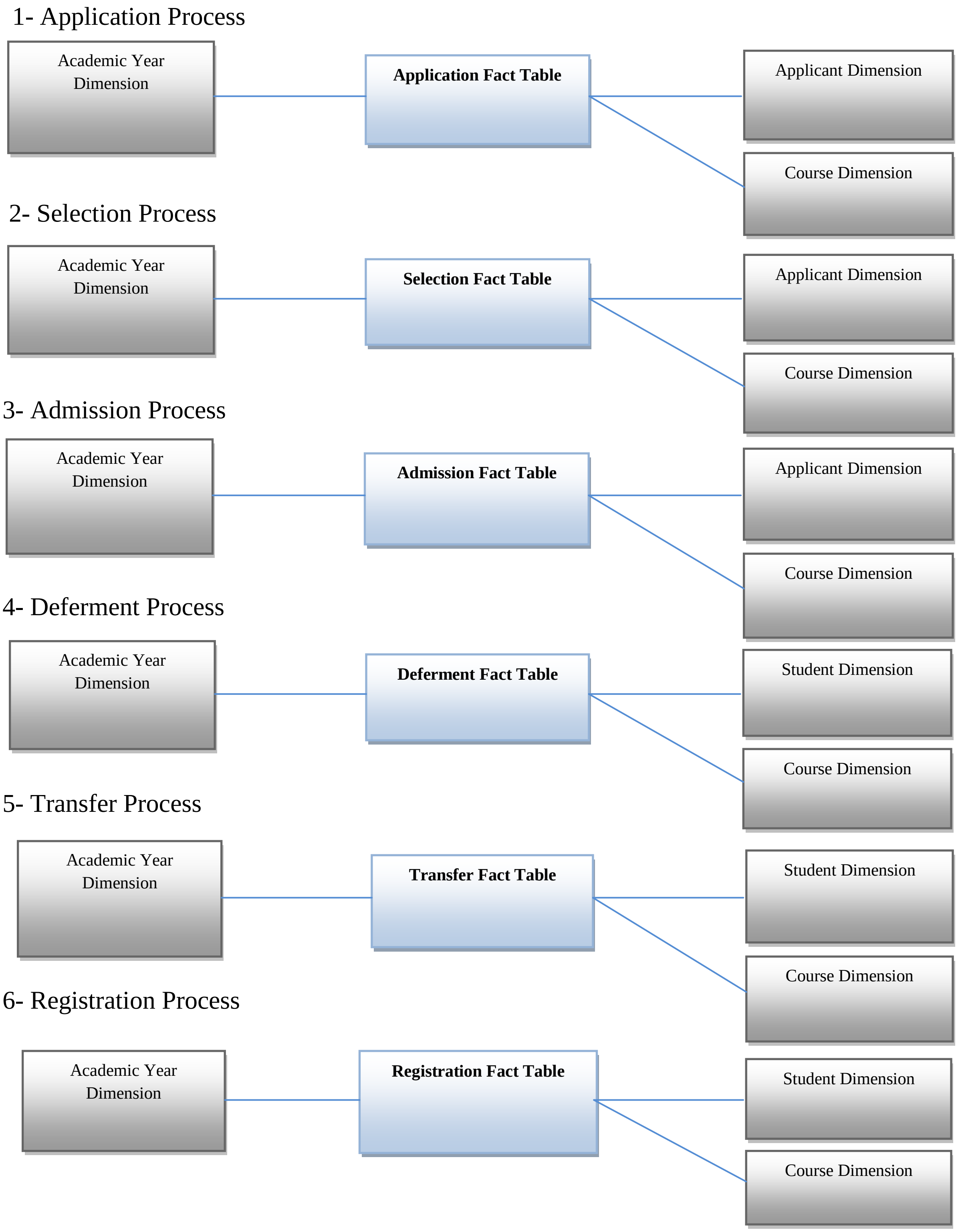
3.5.1 Data Warehouse Bus Matrix

This project revolved around processes pertaining to students in the University. The bus matrix identified major business processes that deal with student life cycle in a university in Kenya.

Table 3: Data Warehouse Bus Matrix

BUSINESS PROCESSES	DIMENSIONS				
	Academic Year	Course	Applicant	Student	Library Resource
1-Application	X	X	X		
2-Selection	X	X	X		
3-Admission	X	X	X		
4-Deferment	X	X		X	
5-Transfer	X	X		X	
6-Registration	X	X		X	
7-Fee Invoicing		X		X	
8-Fee Payment	X	X		X	
9-Class Attendance	X	X		X	
10-Library Usage	X	X		X	X
11-Examination	X	X		X	
12-Graduation	X			X	
13-Alumni Registration	X			X	

3.5.2 Diagram the Data Warehouse Bus Matrix



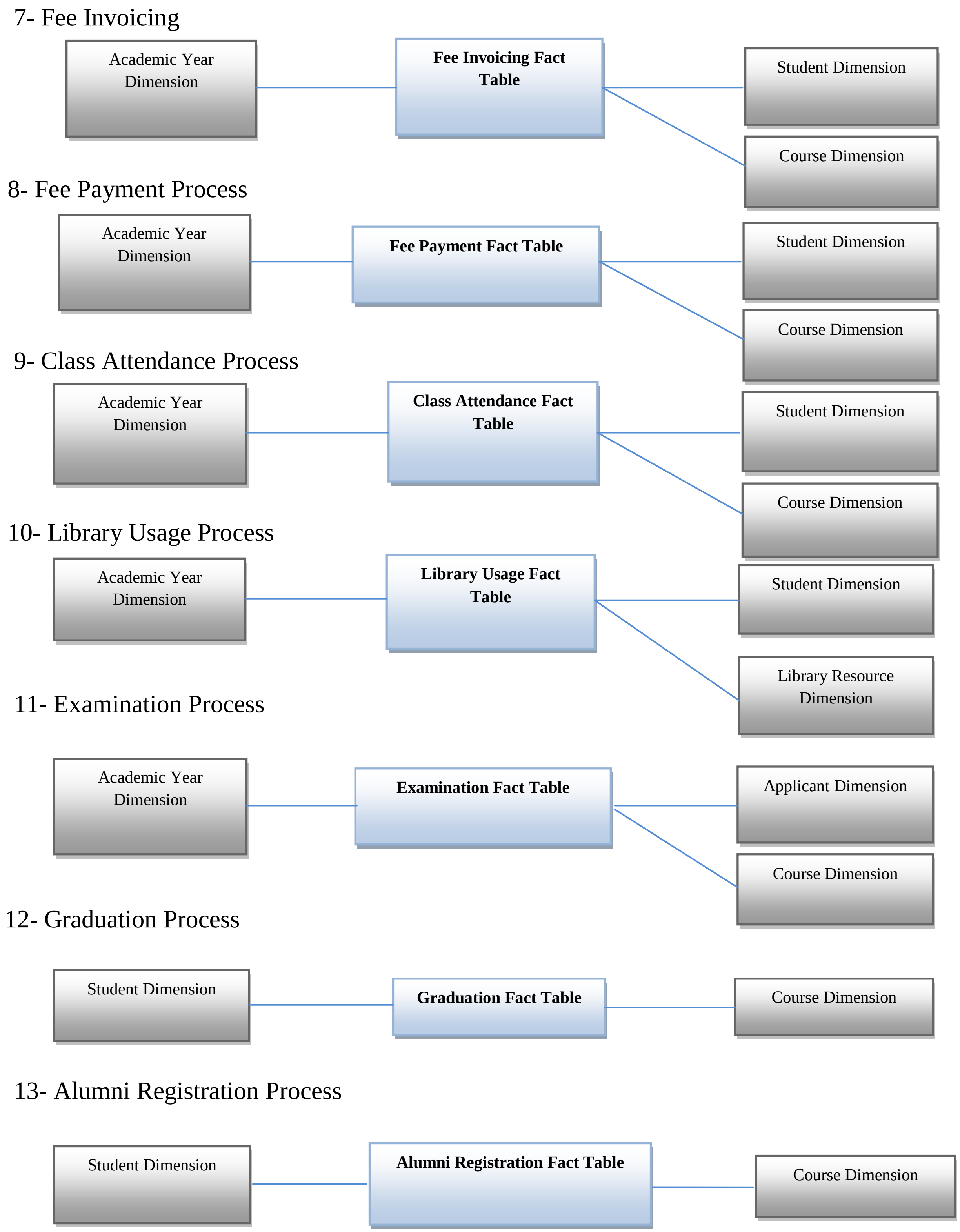


Figure 13: Diagram Showing Data Warehouse Bus Matrix

3.5.3 Attributes Associated with Fact and Dimension Tables

Academic Year Dimension
AcademicYearID(PK)
Name
StartDate
EndDate

Course Dimension
CourseID(PK)
Name
Department
School
Faculty
Campus

Applicant Dimension
ApplicantID(PK)
Name
Gender
County
Religion
Age
Nationality

Student Dimension
StudentID(PK)
YearOfStudy
SemesterOfstudy
CourseID(FK)
Name
Gender
County
Religion
Age
Nationality

Library Resource Dimension
ResourceID(PK)
Name
Category

Application Fact Table
ApplicationID(PK)
AcademicYearID(FK)
CourseID(FK)
ApplicantID(FK)
Date
ApplicationFee

Selection Fact Table
SelectionID(PK)
AcademicYearID(FK)
CourseID(FK)
ApplicantID(FK)
Date

Admission Fact Table
AdmissionID(PK)
AcademicYearID(FK)
CourseID(FK)
ApplicantID(FK)
Date

Deferment Fact Table
DefermentID(PK)
AcademicYearID(FK)
CourseID(FK)
StudentID(FK)
Date
ResumeDate

Transfer Fact Table
TransferID(PK)
AcademicYearID(FK)
CourseID(FK)
StudentID(FK)
Date
TransferCourseID

Registration Fact Table
AcademicYearID(PK)
Date
StudentID(FK)

Fee Invoice Fact Table
InvoiceID(PK)
AcademicYearID(FK)
StudentID(FK)
CourseID(FK)
Date
Amount

Fee Payment Fact Table
PaymentID(PK)
StudentID(FK)
CourseID(FK)
Date
Amount

Class Attendance Fact Table
AttendanceID(PK)
StudentID(FK)
CourseID(FK)
Date
Venue
AcademicYearID(PK)

Library Usage Fact Table
LibraryUsageID(PK)
AcademicYearID(FK)
StudentID(FK)
CourseID(FK)
ResourceID(PK)
Date

Examination Fact Table
ExaminationID(PK)
AcademicYearID(FK)
CourseID(FK)
StudentID(FK)
Date
Grade

Graduation Fact Table
GraduationID(PK)
StudentID(FK)
CourseID(FK)
Date

Alumni Fact Table
AlumniID(PK)
StudentID(FK)
CourseID(FK)
Date

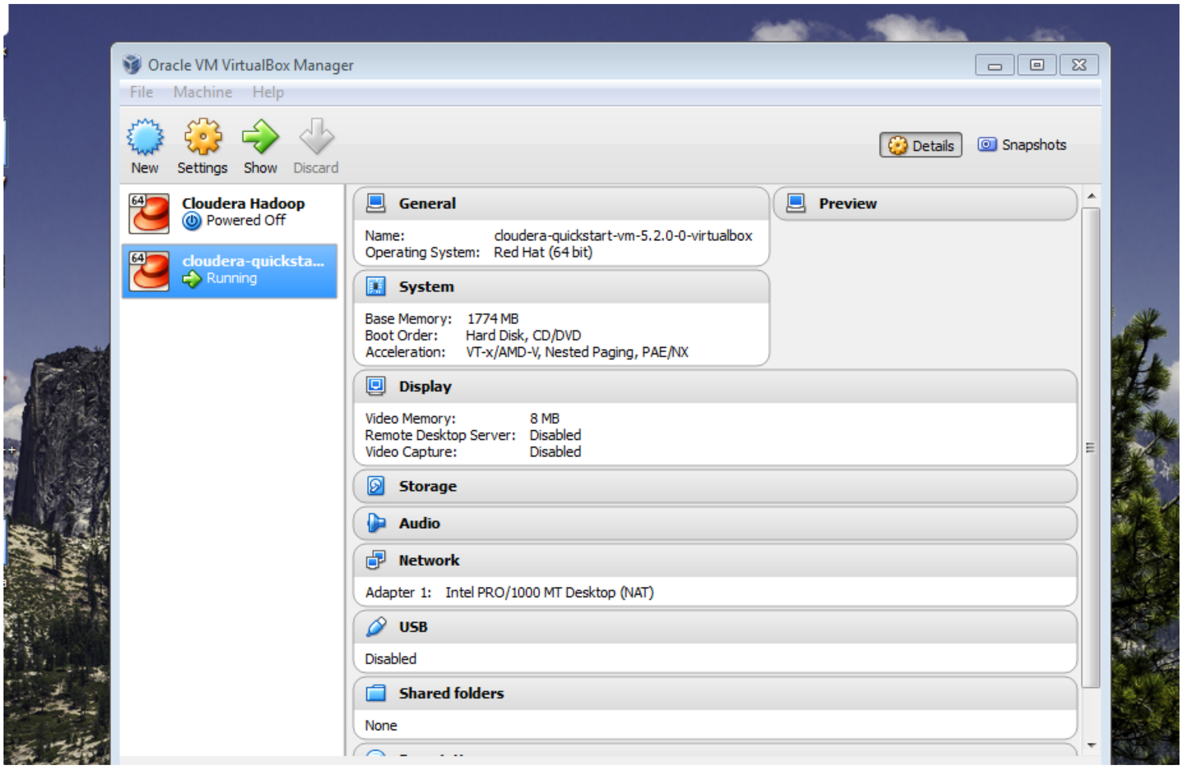
Figure 14: Attributes Associated with Fact and Dimension Tables

3.6 Implementation

Once the data warehouse had been modeled, implementation proceeded. This mainly, consisted of setting up of host machine (single-node cluster), Hadoop installation and configuration, setting up of programming environment, data warehouse development, dashboard creation and data analysis using R.

3.6.1 Setting up of the Host Machine

The initial step of the implementation was to set up the host machine. The project used Microsoft Windows 7 64 bit operating system on a core i5 processor with 4GB of RAM. The machine BIOS setup was configured to activate virtualization technology. This followed with an installation of virtual machine (VM) called Oracle's Virtual Box on top of Windows 7 operating system. It is this Virtual Box where the Hadoop cluster was to be configured.



Cloudera Hadoop
Powered Off

cloudera-quicksta...
Running

General

Name: cloudera-quickstart-vm-5.2.0-0-virtualbox
Operating System: Red Hat (64 bit)

System

Base Memory: 1774 MB
Boot Order: Hard Disk, CD/DVD
Acceleration: VT-x/AMD-V, Nested Paging, PAE/NX

Display

Video Memory: 8 MB
Remote Desktop Server: Disabled
Video Capture: Disabled

Storage

Audio

Network

Adapter 1: Intel PRO/1000 MT Desktop (NAT)

USB

Disabled

Shared folders

None

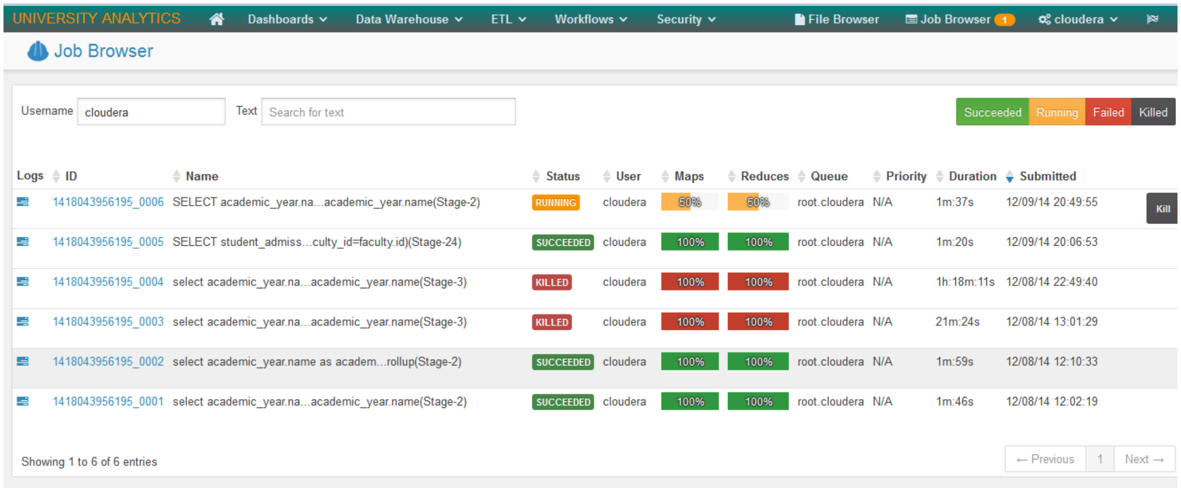
Figure 15: Oracle's Virtual Box

3.6.2 Hadoop Cluster Installation

Hadoop was installed on the Virtual Box using Cloudera Hadoop Distribution, which is one of the Hadoop vendors. The package was downloaded from the Cloudera website and installed. This is Hadoop version for pseudo-distributed single node cluster, meaning that it can run on one machine but use the processors of the machines as nodes to perform distributed processing. The version of the Cloudera Hadoop was 5.2.0.0 (Cloudera Quickstart VM 5.2.0.0).

3.6.3 Configuration of Hadoop Components

Hadoop distribution package is usually bundled with other components for example Hive, Pig, HBase, Apache Solr, ZooKeeper, Sqoop, Hue, among others. Once the Hadoop distribution was installed, its components were configured so as to be ready for use. Among the configuration required were the ports, external data sources, and file permissions.



Username: cloudera

Logs	ID	Name	Status	User	Maps	Reduces	Queue	Priority	Duration	Submitted
	1418043956195_0006	SELECT academic_year.na...academic_year.name(Stage-2)	RUNNING	cloudera	50%	50%	root.cloudera	N/A	1m:37s	12/09/14 20:49:55
	1418043956195_0005	SELECT student_admiss...culty_id=faculty.id)(Stage-24)	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	1m:20s	12/09/14 20:06:53
	1418043956195_0004	select academic_year.na...academic_year.name(Stage-3)	KILLED	cloudera	100%	100%	root.cloudera	N/A	1h:18m:11s	12/08/14 22:49:40
	1418043956195_0003	select academic_year.na...academic_year.name(Stage-3)	KILLED	cloudera	100%	100%	root.cloudera	N/A	21m:24s	12/08/14 13:01:29
	1418043956195_0002	select academic_year.name as academ...rollup(Stage-2)	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	1m:59s	12/08/14 12:10:33
	1418043956195_0001	select academic_year.na...academic_year.name(Stage-2)	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	1m:46s	12/08/14 12:02:19

Showing 1 to 6 of 6 entries

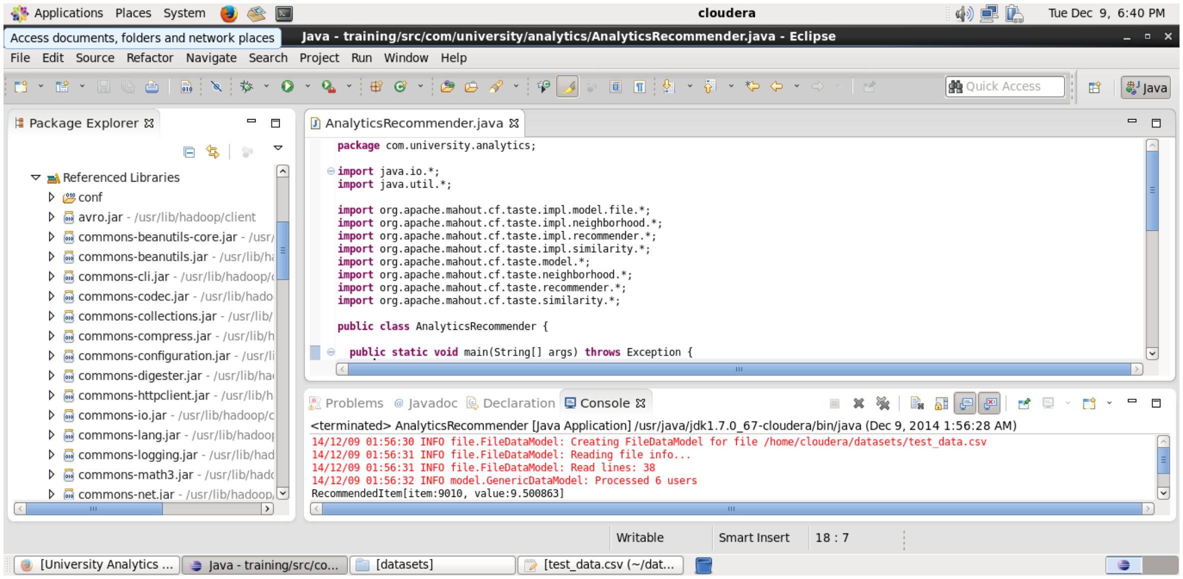
Figure 16: Job Browser for Hadoop

3.6.4 Installation of R and Hadoop Connectors

R statistical software was installed on the host machine and integrated to Hadoop. The packages for installing connectors to Hadoop cluster included rJava, RHadoop, RImpala, Java database connectivity (JDBC), and open database connectivity (ODBC). These packages allow R to access Hadoop HDFS. RHadoop was also downloaded as R packages and installed into the Hadoop ecosystem.

3.6.5 Configuration of Eclipse Integrated Development Environment

Eclipse Integrated Development Environment (IDE) was used to code and test the BI system. Apache Maven was used as the project management tool for building the source codes and other project artifacts. The IDE enabled testing of MapReduce jobs and also non-Hadoop Java programs.


Applications Places System
cloudera
Tue Dec 9, 6:40 PM
Access documents, folders and network places
Java - training/src/com/university/analytics/AnalyticsRecommender.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Java
Package Explorer
Referenced Libraries
conf
avro.jar - /usr/lib/hadoop/client
commons-beanutils-core.jar - /usr/
commons-beanutils.jar - /usr/lib/h
commons-cli.jar - /usr/lib/hadoop/
commons-codec.jar - /usr/lib/hado
commons-collections.jar - /usr/lib/
commons-compress.jar - /usr/lib/h
commons-configuration.jar - /usr/li
commons-digester.jar - /usr/lib/ha
commons-httpclient.jar - /usr/lib/h
commons-io.jar - /usr/lib/hadoop/c
commons-lang.jar - /usr/lib/hadoo
commons-logging.jar - /usr/lib/had
commons-math3.jar - /usr/lib/had
commons-net.jar - /usr/lib/hadoop
AnalyticsRecommender.java
Problems
Javadoc
Declaration
Console
Writable
Smart Insert
18 : 7
[University Analytics ...
Java - training/src/co...
[datasets]
[test_data.csv (~/dat...

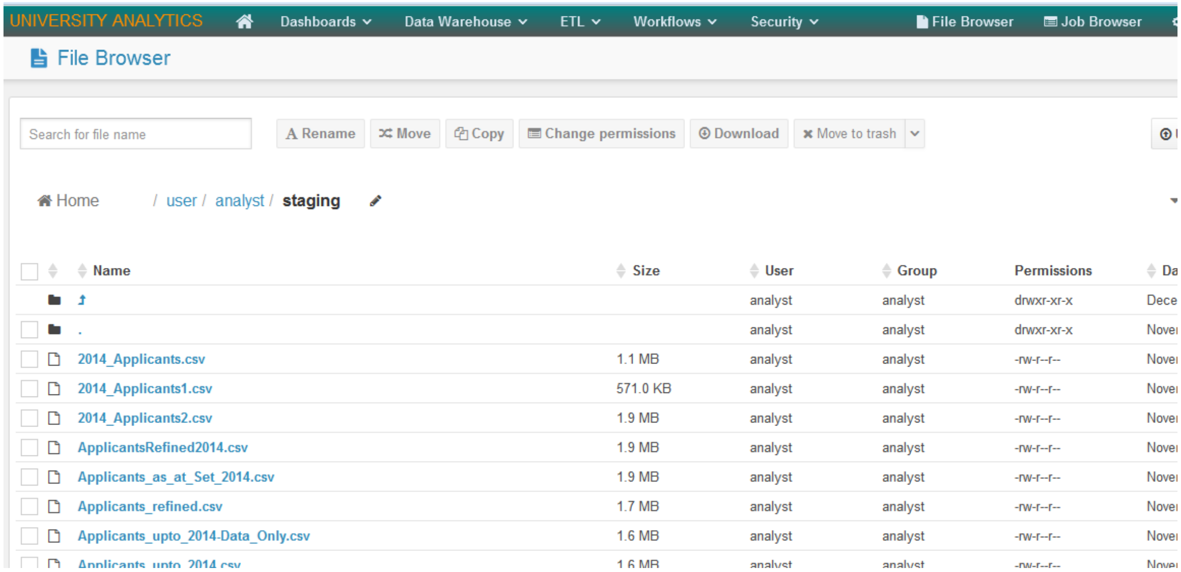
```
package com.university.analytics;

import java.io.*;
import java.util.*;

import org.apache.mahout.cf.taste.impl.model.file.*;
import org.apache.mahout.cf.taste.impl.neighborhood.*;
import org.apache.mahout.cf.taste.impl.recommender.*;
import org.apache.mahout.cf.taste.impl.similarity.*;
import org.apache.mahout.cf.taste.model.*;
import org.apache.mahout.cf.taste.neighborhood.*;
import org.apache.mahout.cf.taste.recommender.*;
import org.apache.mahout.cf.taste.similarity.*;

public class AnalyticsRecommender {

  public static void main(String[] args) throws Exception {
```

```
<terminated> AnalyticsRecommender [Java Application] /usr/java/jdk1.7.0_67-cloudera/bin/java (Dec 9, 2014 1:56:28 AM)
14/12/09 01:56:30 INFO file.FileDataModel: Creating FileDataModel for file /home/cloudera/datasets/test_data.csv
14/12/09 01:56:31 INFO file.FileDataModel: Reading file info...
14/12/09 01:56:31 INFO file.FileDataModel: Read lines: 38
14/12/09 01:56:32 INFO model.GenericDataModel: Processed 6 users
RecommendedItem[item:9010, value:9.500863]
```

Figure 17: Eclipse IDE Setup

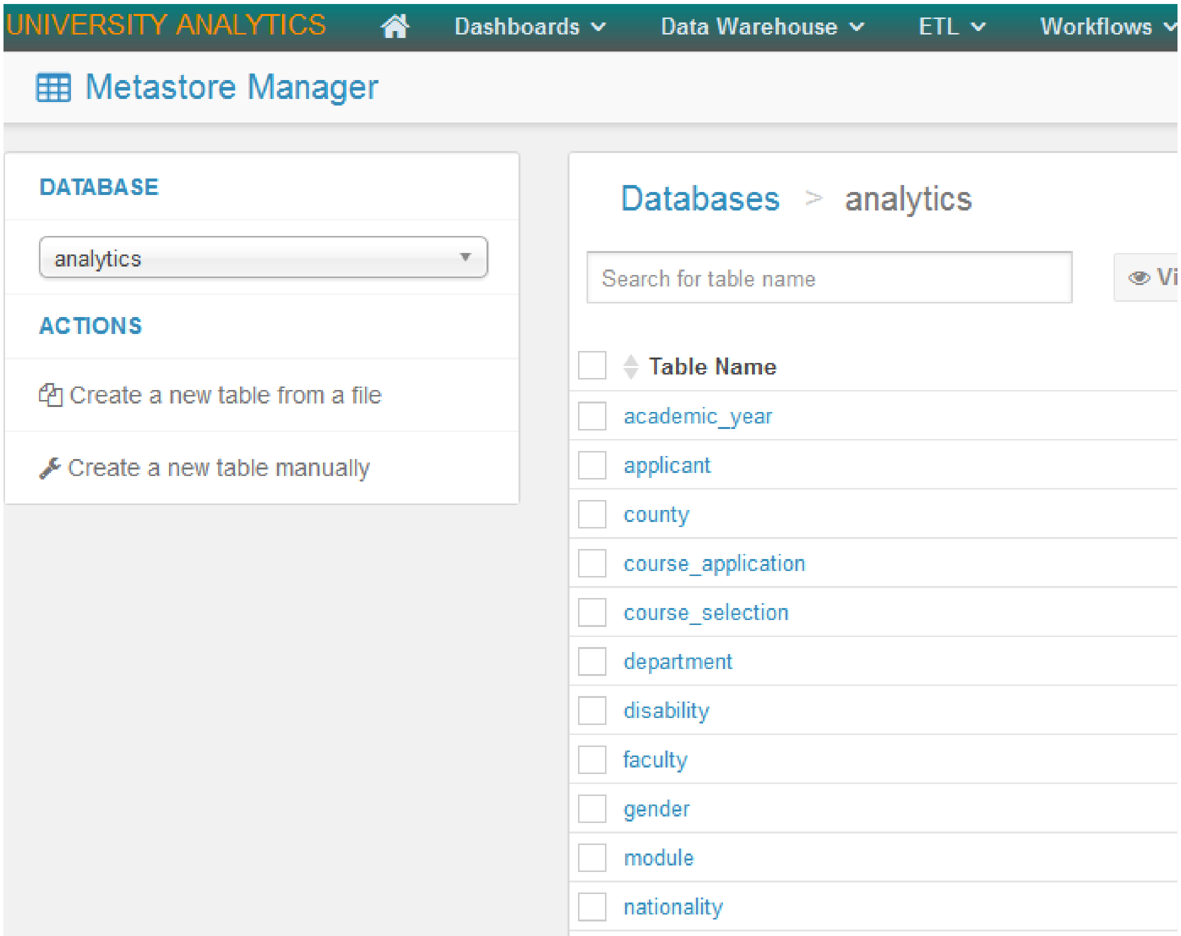
3.6.6 Data Warehouse Development

Data schema was developed using the Star Schema based on the Kimball's Dimension Modeling discussed above. Sample data were generated with help of existing data at the Technical University of Kenya. This was due to confidentially of the University data, avoiding using live data. Data was transferred to Hive data warehouse using Hive Table MetaStores. Pig scripts were also written for some extraction, transformation and loading procedures. Hadoop User Experience (HUE) was configured and used to develop web portal for accessing the data warehouse.



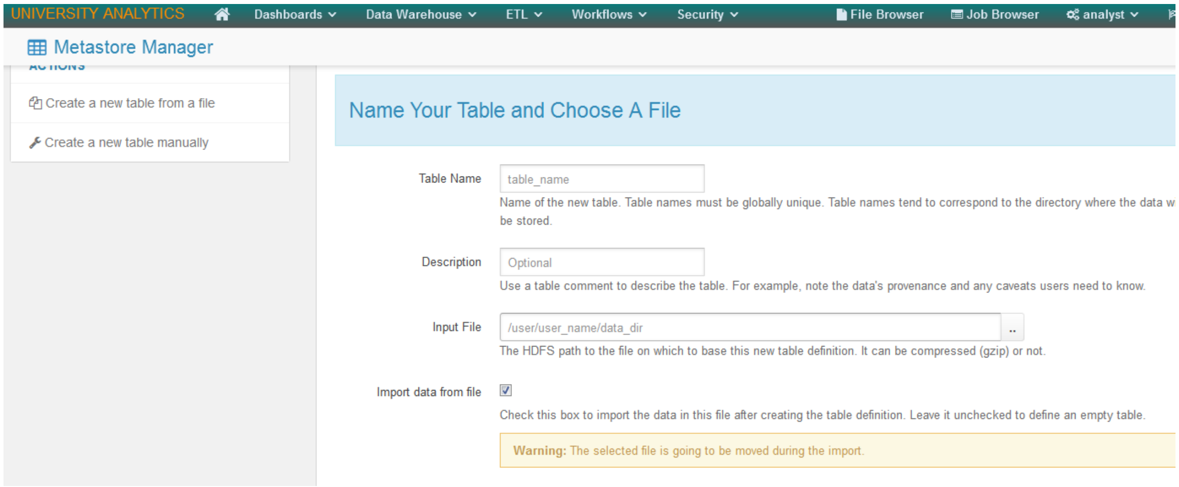
□		Name	Size	User	Group	Permissions	Da
		↥		analyst	analyst	drwxr-xr-x	Dece
□		.		analyst	analyst	drwxr-xr-x	Nove
□		2014_Applicants.csv	1.1 MB	analyst	analyst	-rw-r--r--	Nove
□		2014_Applicants1.csv	571.0 KB	analyst	analyst	-rw-r--r--	Nove
□		2014_Applicants2.csv	1.9 MB	analyst	analyst	-rw-r--r--	Nove
□		ApplicantsRefined2014.csv	1.9 MB	analyst	analyst	-rw-r--r--	Nove
□		Applicants_as_at_Set_2014.csv	1.9 MB	analyst	analyst	-rw-r--r--	Nove
□		Applicants_refined.csv	1.7 MB	analyst	analyst	-rw-r--r--	Nove
□		Applicants_upto_2014-Data_Only.csv	1.6 MB	analyst	analyst	-rw-r--r--	Nove
□		Applicants_upto_2014.csv	1.6 MB	analyst	analyst	-rw-r--r--	Nove

Figure 18: File browser powered by Hue



□	Table Name
□	academic_year
□	applicant
□	county
□	course_application
□	course_selection
□	department
□	disability
□	faculty
□	gender
□	module
□	nationality

Figure 19: Hive Metastore Manager for Extraction, Transformation and Loading



Name Your Table and Choose A File

Table Name: table_name

Name of the new table. Table names must be globally unique. Table names tend to correspond to the directory where the data w be stored.

Description: Optional

Use a table comment to describe the table. For example, note the data's provenance and any caveats users need to know.

Input File: /user/user_name/data_dir

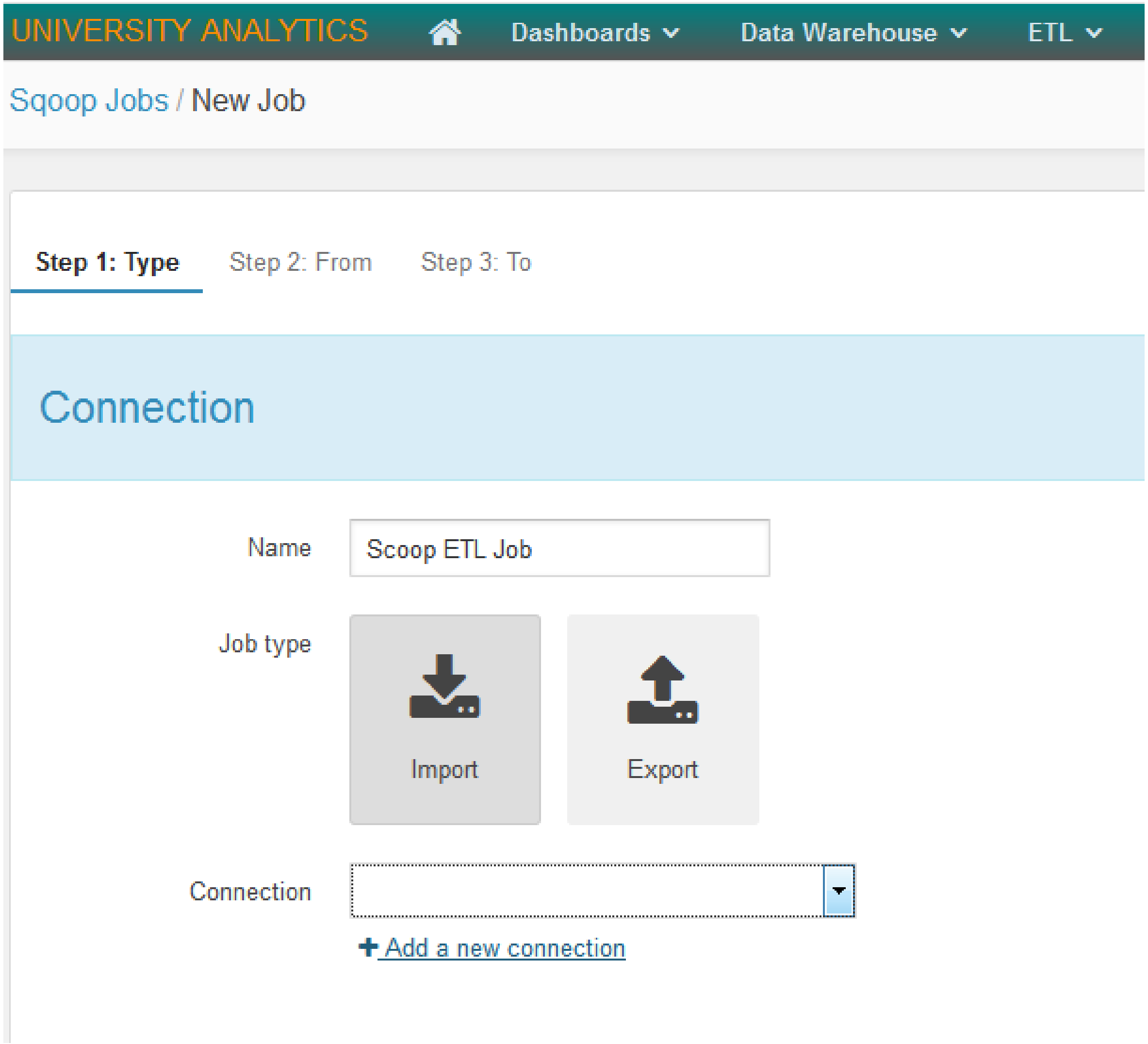
The HDFS path to the file on which to base this new table definition. It can be compressed (gzip) or not.

Import data from file ☑

Check this box to import the data in this file after creating the table definition. Leave it unchecked to define an empty table.

Warning: The selected file is going to be moved during the import.

Figure 20: Extraction, transformation and loading using Hive Metadata Tables



Name: Scoop ETL Job

Job type: Import

Connection:

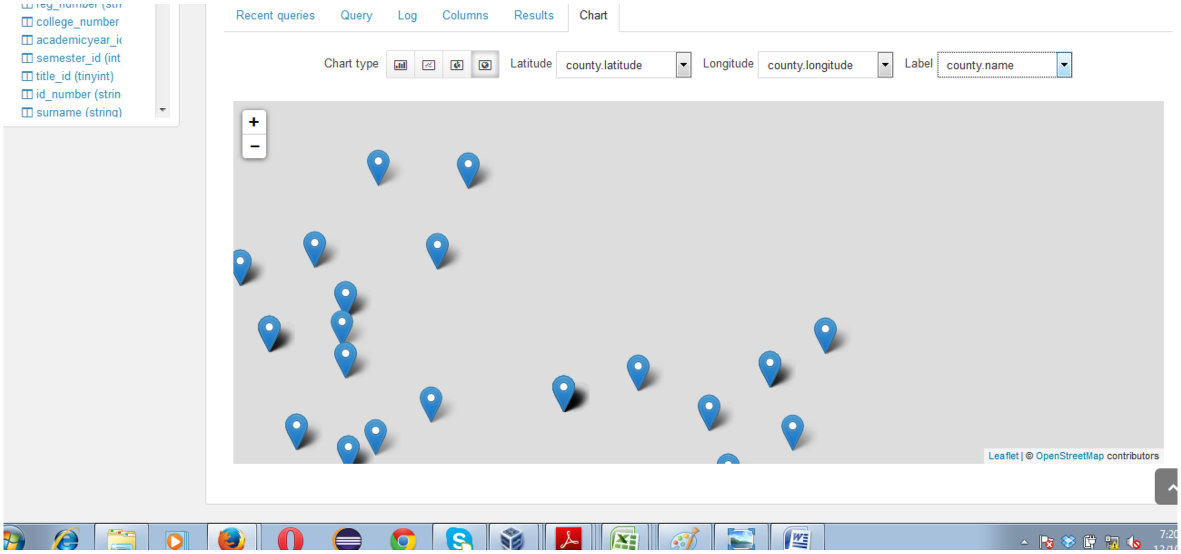
Figure 21: ETL Using Sqoop Job

3.6.7 Data Visualization and Dashboard Development

Different dashboards were created using Apache Solr and Cloudera Search. Analysis of data was carried out to generate the required reports and also answer business questions and appropriate dashboards generated using Apache Solr. Hive queries were also used to analyze and visualize data.

```
SELECT

student_admission.reg_number,academic_year.name,gender.name,county.name,
county.latitude,county.longitude,religion.name,nationality.name,module.name,
salutation.name,programme.name,department.name,school.name,faculty.name, disability.name

FROM

student_admission

JOIN academic_year ON (student_admission.academicyear_id=academic_year.id)

JOIN gender ON (student_admission.gender_id=gender.id)

JOIN county ON (student_admission.county_id=county.id)

JOIN religion ON (student_admission.religion_id=religion.id)

JOIN nationality ON (student_admission.nationality_id=nationality.id)

JOIN module ON (student_admission.module_id=module.id)

JOIN salutation ON (student_admission.title_id=salutation.id)

JOIN disability ON (student_admission.disability_id=disability.id)

JOIN programme ON (student_admission.programme_id=programme.code)

JOIN department ON (programme.department_id=department.id)

JOIN school ON (school.id=department.school_id)

JOIN faculty ON (school.faculty_id=faculty.id)
```

Figure 22: Sample Hive query to retrieve student admission data mart

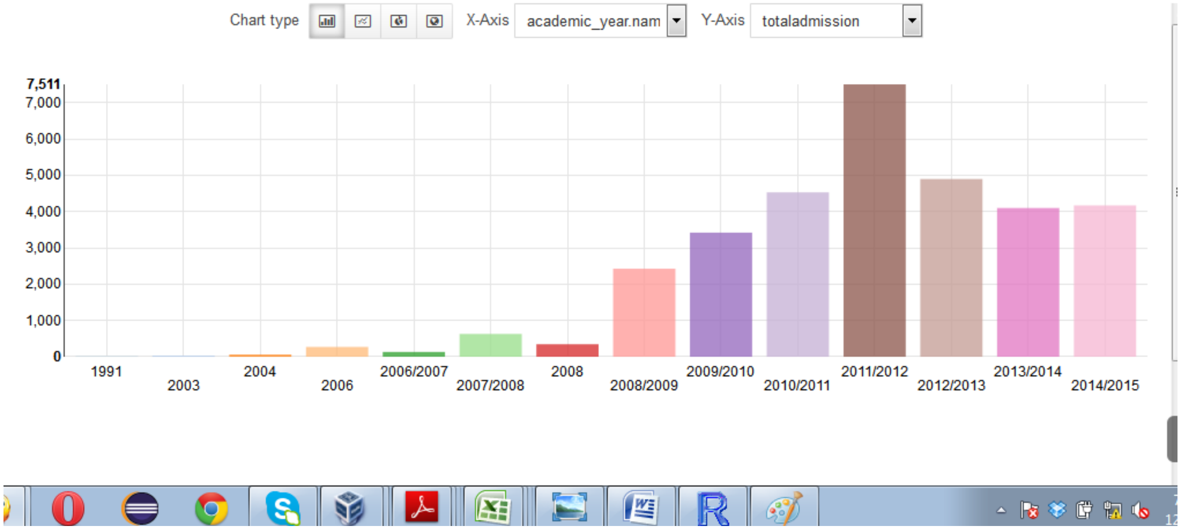


Latitude county.latitude

Longitude county.longitude

Label county.name

Figure 23: Data Visualization of student distribution globally using Hive data warehouse



X-Axis academic_year.nam

Y-Axis totaladmission

Figure 24: Student Admission Pattern

3.6.8 Data Analysis Using R Software

R was integrated with Hadoop and it was possible to transfer data between Hadoop cluster and R.

> Sys.setenv(JAVA_HOME="")

> library(RImpala)

```
Loading required package: rJava
> rimpala.init(libs='C:\\Users/webmaster1/Documents/lib')
[1] "Classpath added successfully"
> rimpala.connect();
[1] TRUE
> rimpala.usedatabase("analytics")
[1] TRUE
> rimpala.query(Q="show tables")
                 name
1       academic_year
2   admission_pattern
3           applicant
4              county
5  course_application
6    course_selection
7          department
8          disability
9             faculty
10             gender
11             module
12        nationality
13          programme
14     programme_type
```

```
15          religion
16        salutation
17            school
18          semester
19  student_admission
20      user_activity
> admission = rimpala.query("select * from analytics.admission_pattern")
> summary(admission)
 academic_yearname  totaladmission
 Length:14          Min.   :   5.0
 Class :character   1st Qu.: 167.2
 Mode  :character   Median :1527.5
                    Mean   :2324.5
                    3rd Qu.:4153.8
                    Max.   :7511.0
> plot(admission)
```

Figure 25: Sample R Code to Connect to Hadoop and Analyze Data

CHAPTER 4

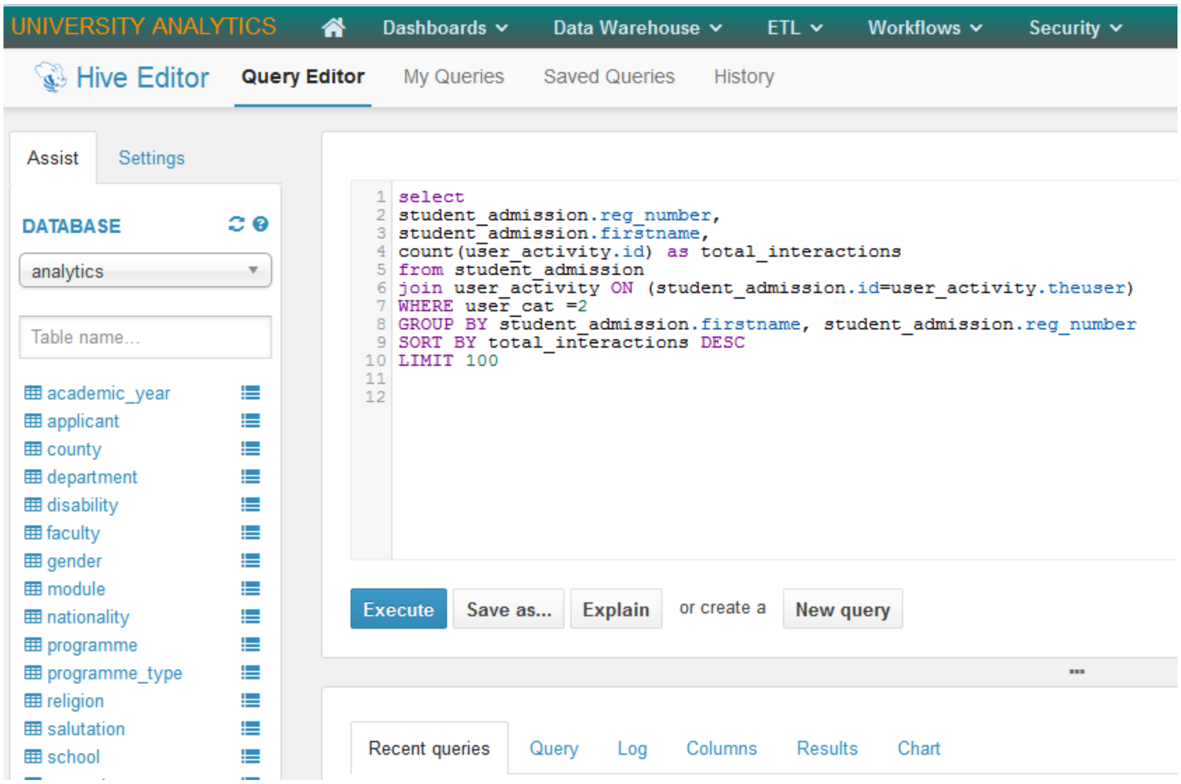
RESULTS AND DISCUSSIONS

4.1 Introduction

This chapter provides the results of implementing a Business Intelligence System for Universities in Kenya. Discussions based on these results are also elaborated.

4.2 Data Warehouse Design

With the above identified university processes, a data warehouse was designed and developed using Apache Hive, one of the Apache projects under Hadoop. The same data warehouse schema was also linked to Impala which is a high-latency Data warehouse component for Hadoop.



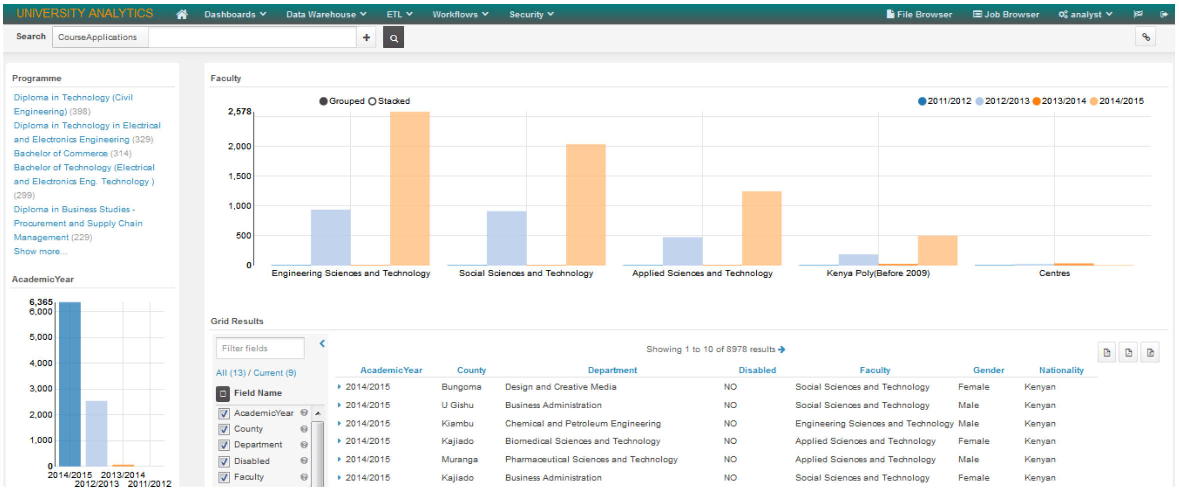
```
select
student_admission.reg_number,
student_admission.firstname,
count(user_activity.id) as total_interactions
from student_admission
join user_activity ON (student_admission.id=user_activity.theuser)
WHERE user_cat =2
GROUP BY student_admission.firstname, student_admission.reg_number
SORT BY total_interactions DESC
LIMIT 100
```

Figure 26: Data Warehouse View using Apache Hive

With this, it is possible to develop a fully working data warehouse using Apache Hive and Impala. Data analysis can be carried out on the data warehouse using ad-hoc queries and visualization. The data warehouse powered by Hive is accessed through Hue.

4.3 User Dashboards

Different dashboards were created for different users user needs. The dashboards included for course application, and student admission. These dashboards were interactive enough to allow user for ‘drill-down’ and ‘roll-ups’.



Showing 1 to 10 of 8978 results

AcademicYear	County	Department	Disabled	Faculty	Gender	Nationality
2014/2015	Bungoma	Design and Creative Media	NO	Social Sciences and Technology	Female	Kenyan
2014/2015	U Gishu	Business Administration	NO	Social Sciences and Technology	Male	Kenyan
2014/2015	Kiambu	Chemical and Petroleum Engineering	NO	Engineering Sciences and Technology	Male	Kenyan
2014/2015	Kajiado	Biomedical Sciences and Technology	NO	Applied Sciences and Technology	Female	Kenyan
2014/2015	Muranga	Pharmaceutical Sciences and Technology	NO	Applied Sciences and Technology	Male	Kenyan
2014/2015	Kajiado	Business Administration	NO	Social Sciences and Technology	Female	Kenyan

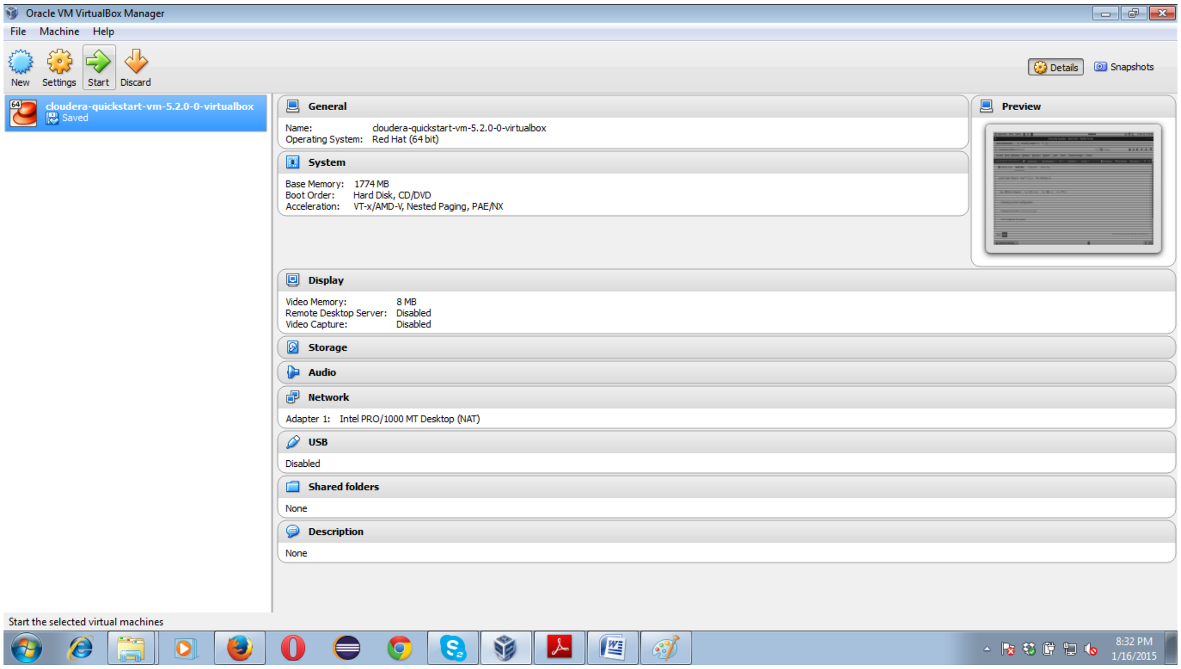
Figure 27: Dashboard for Course Applications

4.4 System User Interaction

This part describes how a user can extract data from different sources into a data warehouse and perform complex analytics.

4.4.1 Starting the Hadoop Cluster

In order to start the Hadoop cluster, open the Virtual Box application, then start Cloudera Virtual Machine (VM) already installed. Remember you can save state of your VM.


Oracle VM VirtualBox Manager
File
Machine
Help
New
Settings
Start
Discard
Details
Snapshots
cloudera-quickstart-vm-5.2.0-0-virtualbox
Saved
Preview
Start the selected virtual machines
8:32 PM
1/16/2015

General

Name: cloudera-quickstart-vm-5.2.0-0-virtualbox
Operating System: Red Hat (64 bit)

System

Base Memory: 1774 MB
Boot Order: Hard Disk, CD/DVD
Acceleration: VT-x/AMD-V, Nested Paging, PAE/NX

Display

Video Memory: 8 MB
Remote Desktop Server: Disabled
Video Capture: Disabled

Storage

Audio

Network

Adapter 1: Intel PRO/1000 MT Desktop (NAT)

USB

Disabled

Shared folders

None

Description

None

Figure 28: Starting the Hadoop Cluster

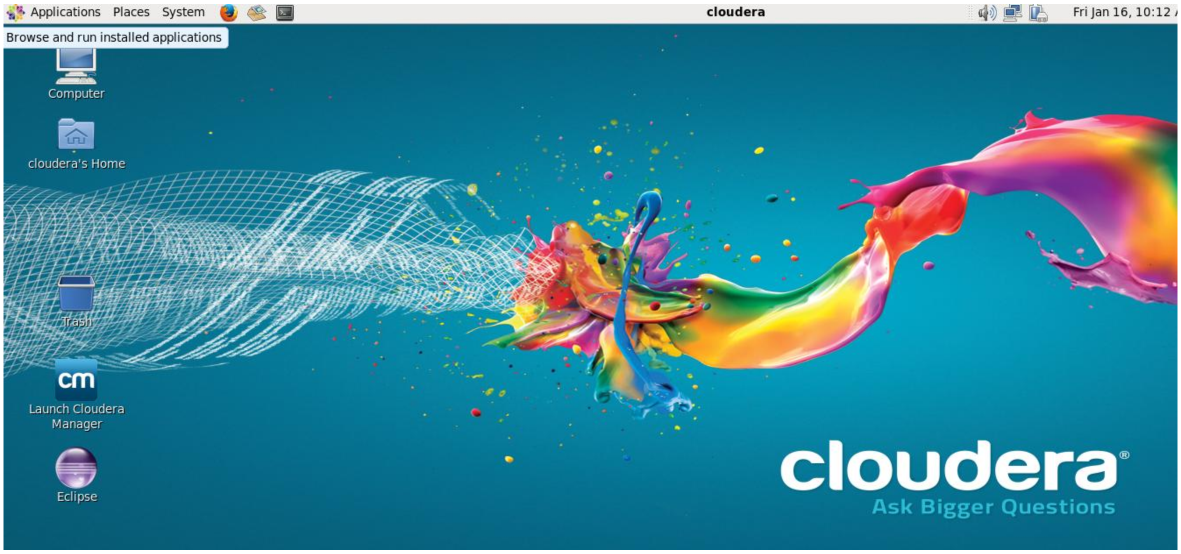


Figure 29: Cloudera VM

4.4.1 User Login

When the system is accessed remotely through the host machine, then login will be required. Otherwise, a user is logged in automatically when the cluster is started. The system can be accessed using the following links:

- Virtual Machine: http://quickstart.cloudera:8888/
- Host Machine: http://localhost:8888/

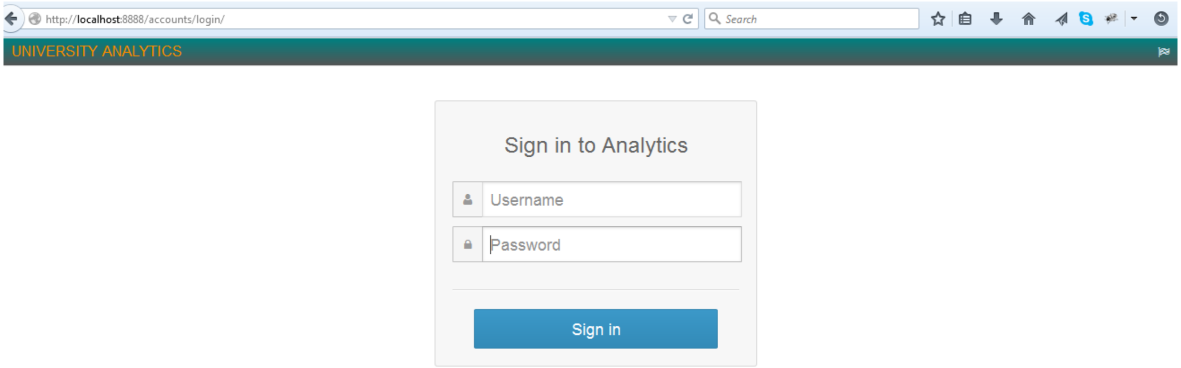
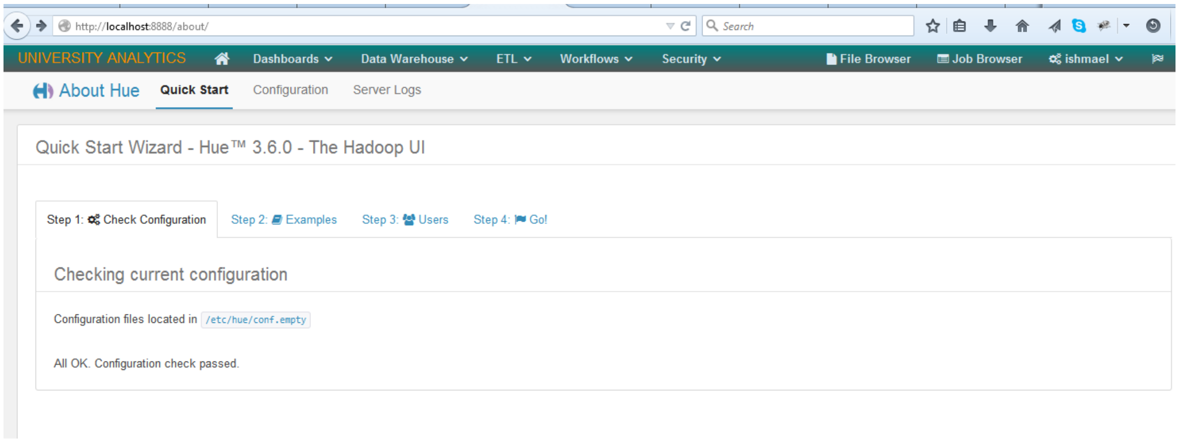


Figure 30: User Login Page

4.4.2 System Configuration

Once you log in, if you are super user, then you will be directed to system configuration setup page. In case of any errors, then this page will always raise flags.



Quick Start Wizard - Hue™ 3.6.0 - The Hadoop UI

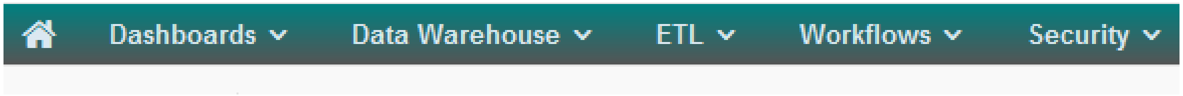
Checking current configuration

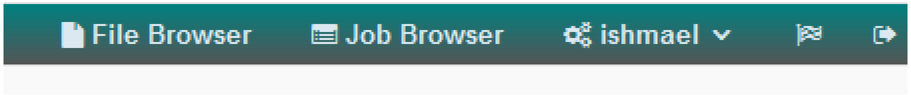
Configuration files located in `/etc/hue/conf.empty`

All OK. Configuration check passed.

Figure 31: System Configuration Status Page

4.4.3 Main System Links





The main system links are:

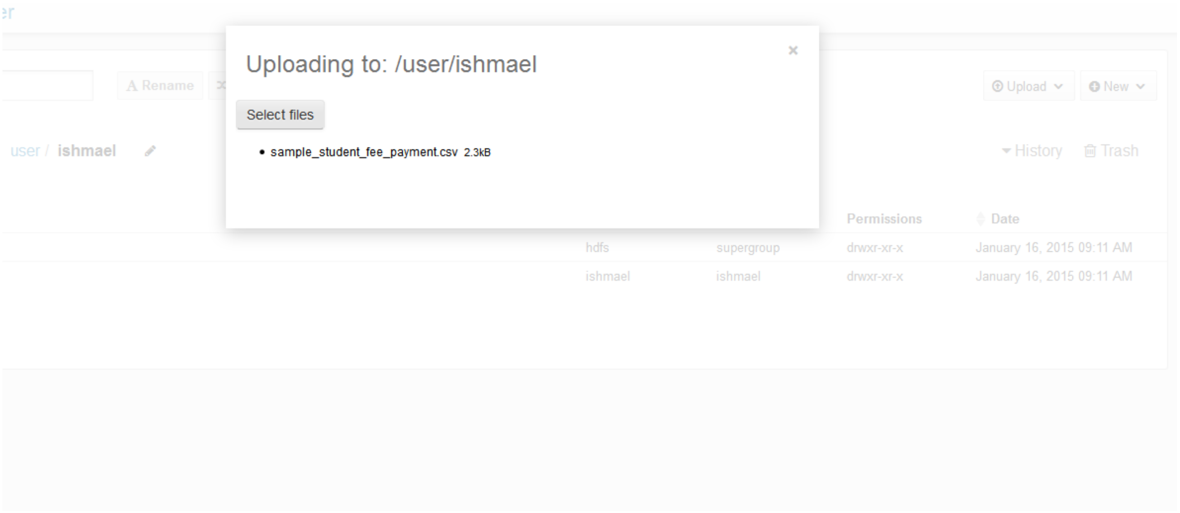
- **User Profile:** This links to profile of currently logged in user and also user management panel.
- **Job Browser:** This provides access to management of MapReduce jobs.
- **File Browser:** This enables access to HDFS; allowing uploading, deleting, renaming, moving, and copying of files and folders, amongst others.
- **Security:** This enables assignment of roles and privileges to users of the system.
- **Workflows:** This link provides access to Hadoop workflows, for example Pig jobs.

- **ETL:** Enables the extraction transformation and loading of data into the data warehouse. Tools adopted include Metastore Tables and Sqoop Transfer.
- **Data Warehouse:** This links to data warehouse of the system. Apache Hive and Impala have been used to implement the data warehouse. Queries on Hive are executed using MapReduce while Impala has in-built query engine.
- **Dashboards:** This is where users can access various dashboards. Users can also create new dashboards based on the business answers they may be seeking.

4.4.4 Extract and Load Data into Data Warehouse and Create Dashboards

NOTE: This step does not involve reading data from live sources like databases. To connect Hadoop to RDBMS, please use SQOOP tool.

Step 1: Data Preparation - Prepare your data in comma separated values (CSV) format. Upload this data into the Hadoop HDFS under the directory of your choice.



Uploading to: /user/ishmael

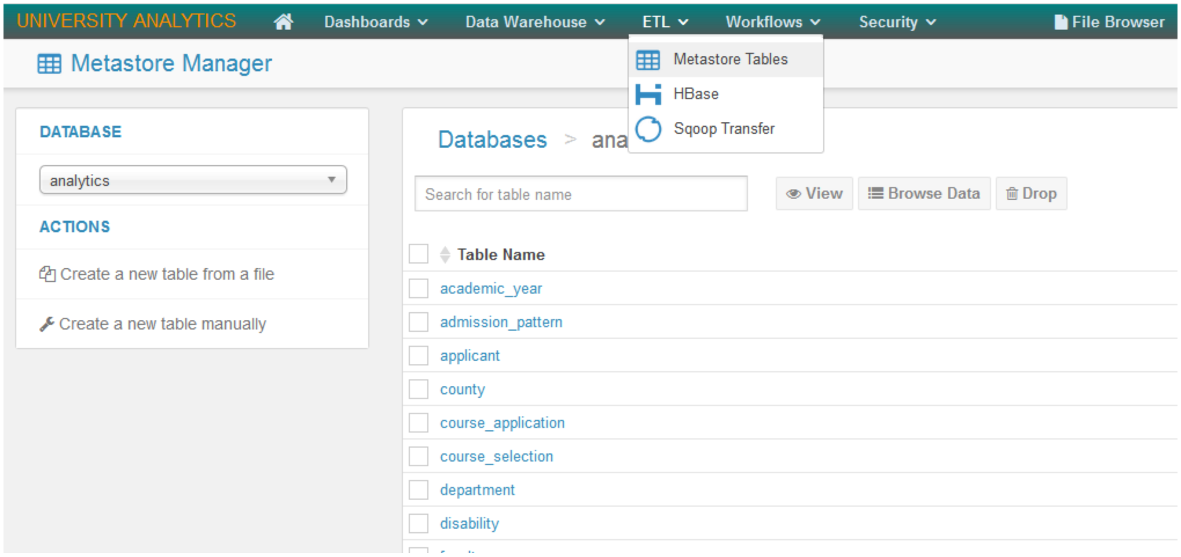
- sample_student_fee_payment.csv 2.3kB

user / ishmael

		Permissions	Date
hdfs	supergroup	drwxr-xr-x	January 16, 2015 09:11 AM
ishmael	ishmael	drwxr-xr-x	January 16, 2015 09:11 AM

Figure 32: Uploading Data Files to HDFS

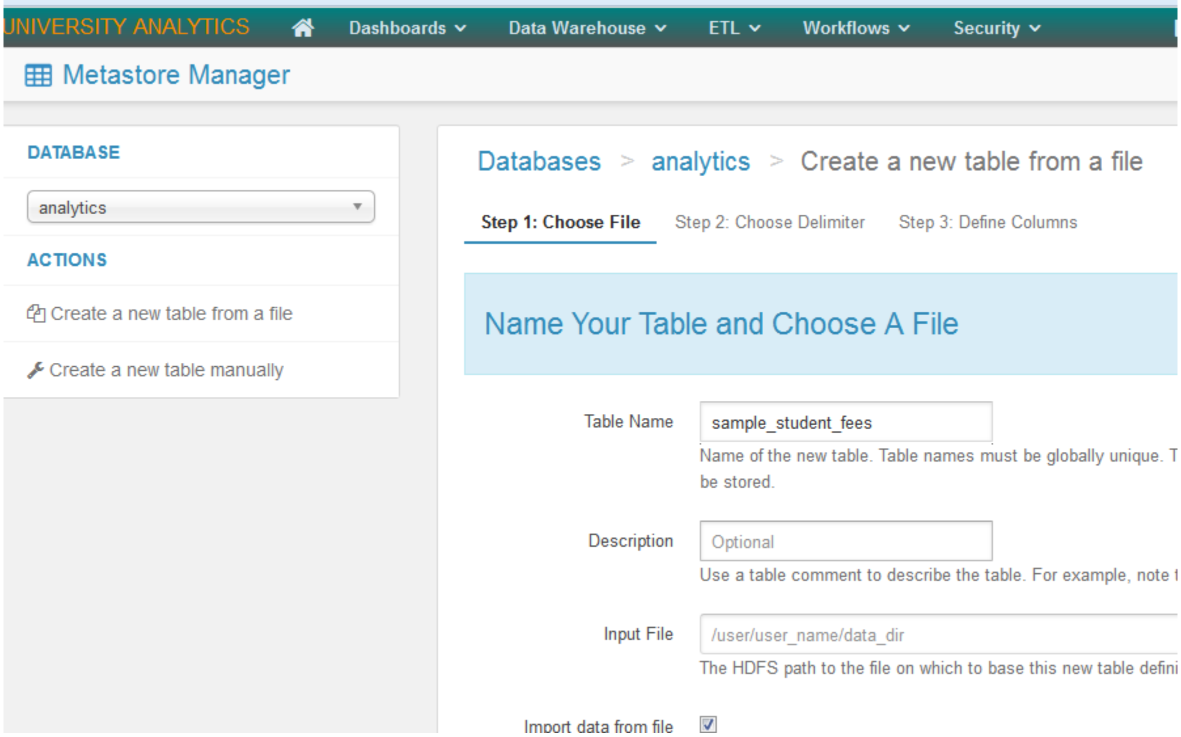
Step 2: Loading Data into Data Warehouse- Load the file into the data warehouse using Metastore Tables.



Database: analytics

Table Name
academic_year
admission_pattern
applicant
county
course_application
course_selection
department
disability

Figure 33: Uploading Data into Data Warehouse Part 1



Database: analytics

Name Your Table and Choose A File

Table Name: sample_student_fees

Name of the new table. Table names must be globally unique. T be stored.

Description: Optional

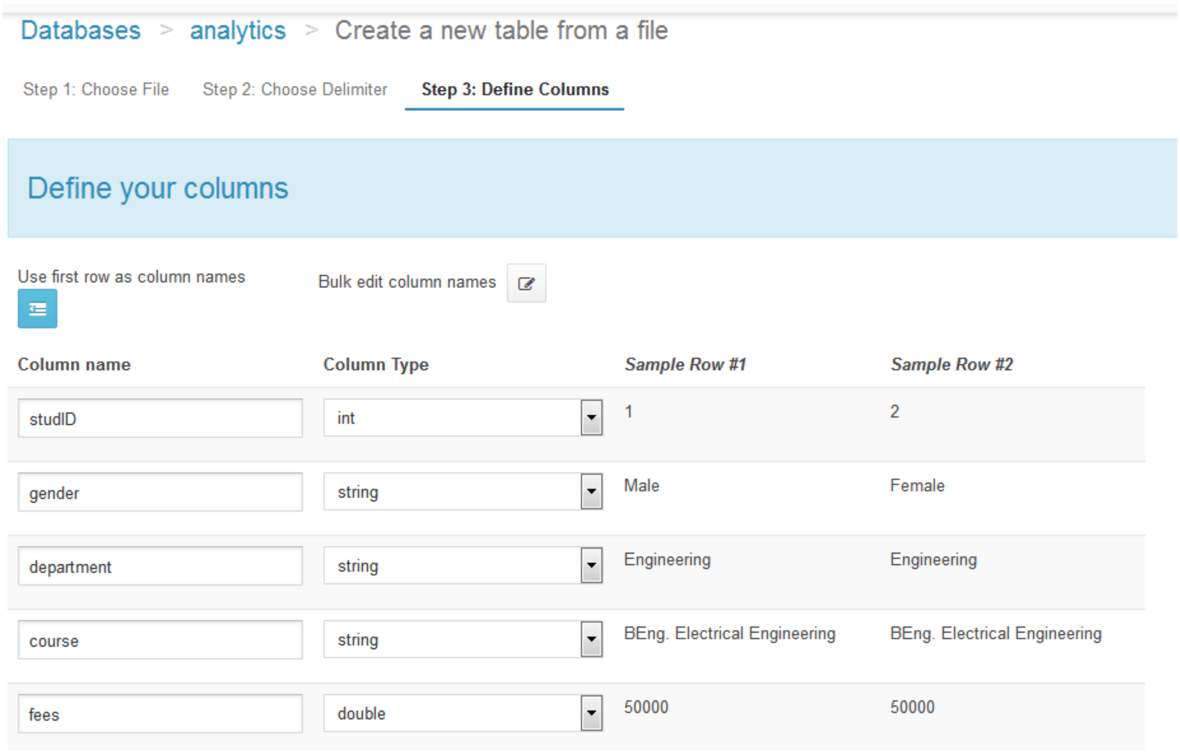
Use a table comment to describe the table. For example, note t

Input File: /user/user_name/data_dir

The HDFS path to the file on which to base this new table defini

Import data from file ☑

Figure 34: Uploading Data into Data Warehouse Part 2

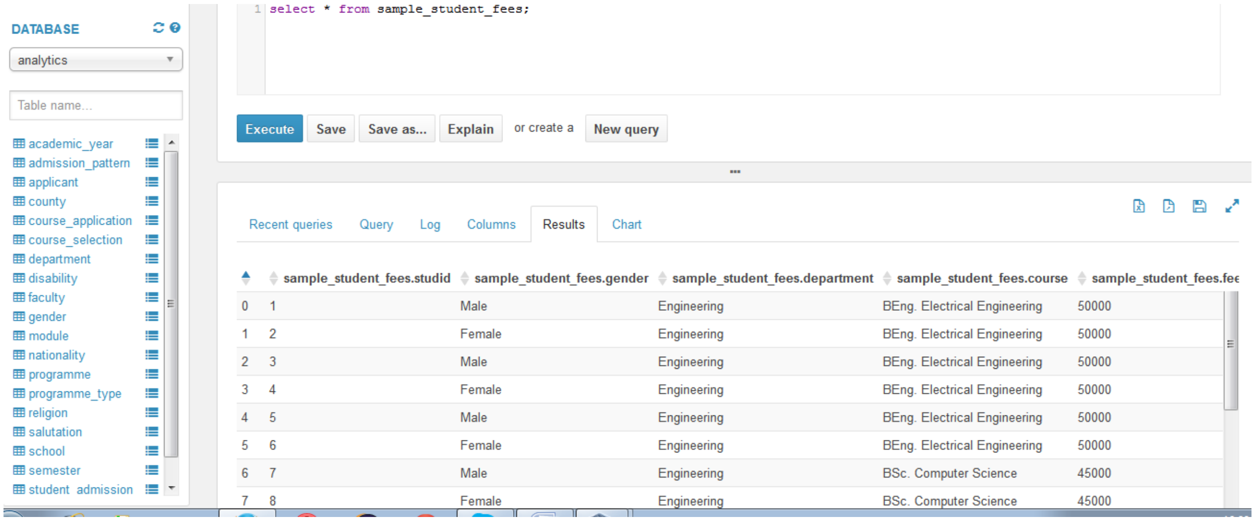


Column name	Column Type	*Sample Row #1*	*Sample Row #2*
studID	int	1	2
gender	string	Male	Female
department	string	Engineering	Engineering
course	string	BEng. Electrical Engineering	BEng. Electrical Engineering
fees	double	50000	50000

Figure 35: Uploading Data into Data Warehouse Part 3

Step 3: Data Analytics using Hive- Data Analysis using Hive. Using MapReduce queries on Hive various analyses can be carried out.

Queries can be written and submitted to MapReduce job as shown below.

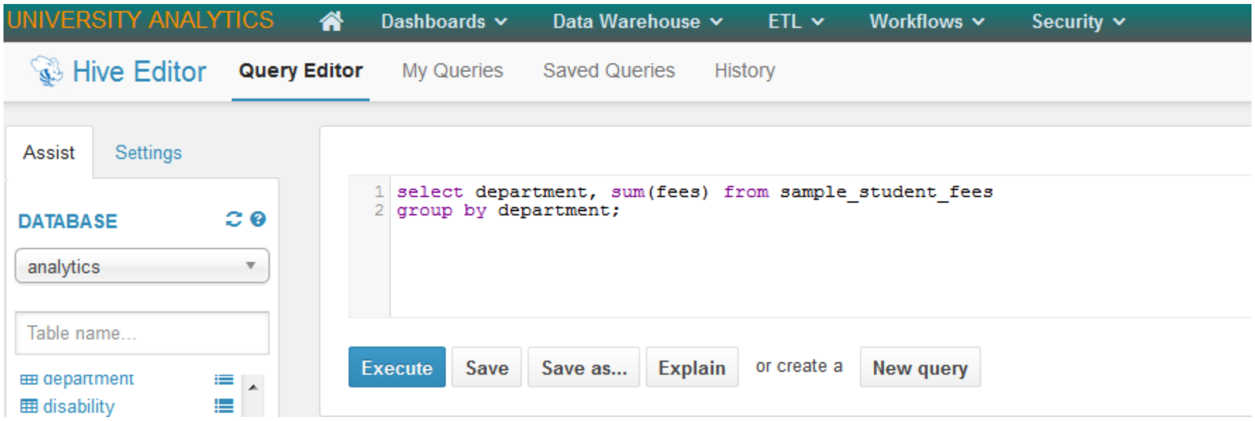


```
select * from sample_student_fees;
```

	sample_student_fees.studid	sample_student_fees.gender	sample_student_fees.department	sample_student_fees.course	sample_student_fees.fee
0	1	Male	Engineering	BEng. Electrical Engineering	50000
1	2	Female	Engineering	BEng. Electrical Engineering	50000
2	3	Male	Engineering	BEng. Electrical Engineering	50000
3	4	Female	Engineering	BEng. Electrical Engineering	50000
4	5	Male	Engineering	BEng. Electrical Engineering	50000
5	6	Female	Engineering	BEng. Electrical Engineering	50000
6	7	Male	Engineering	BSc. Computer Science	45000
7	8	Female	Engineering	BSc. Computer Science	45000

Figure 36: Listing of Student Sample Fees

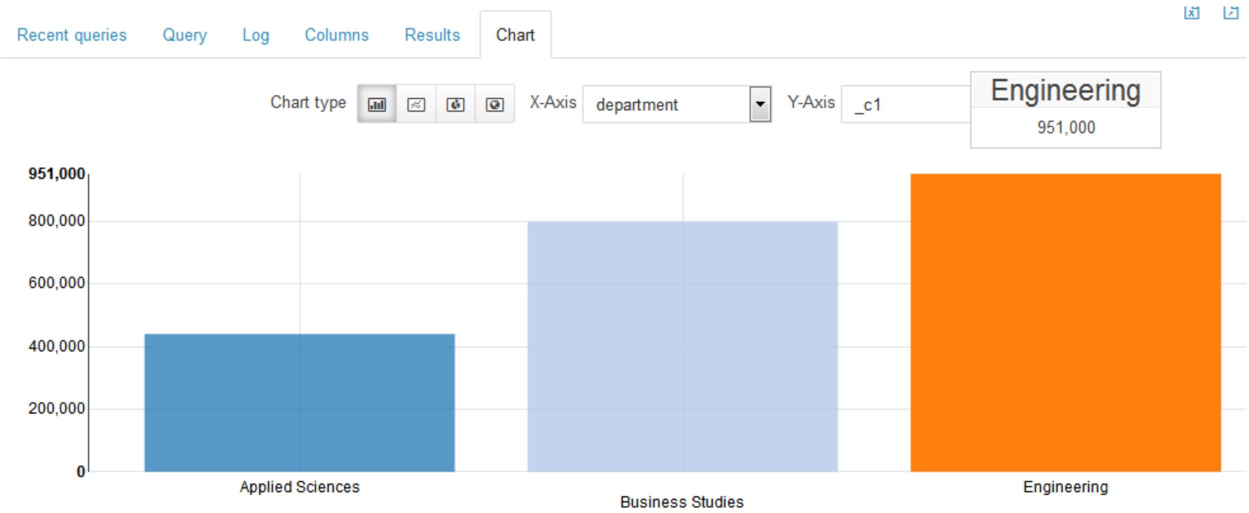
Graphical comparison can also be realized.



analytics

```
select department, sum(fees) from sample_student_fees
group by department;
```

Figure 37: Query Editor for Ad Hoc Reports



X-Axis department

Y-Axis _c1

Figure 38: Bar Graph Report for Sample Student Fees

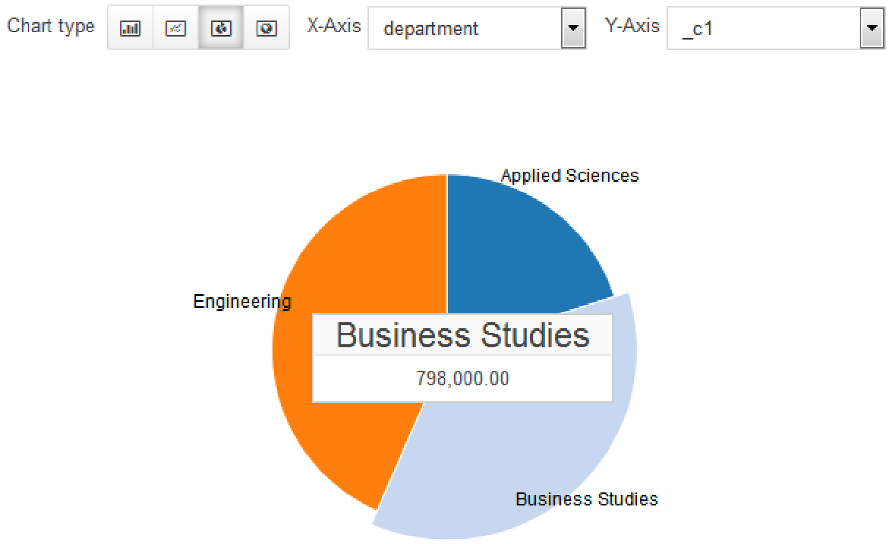


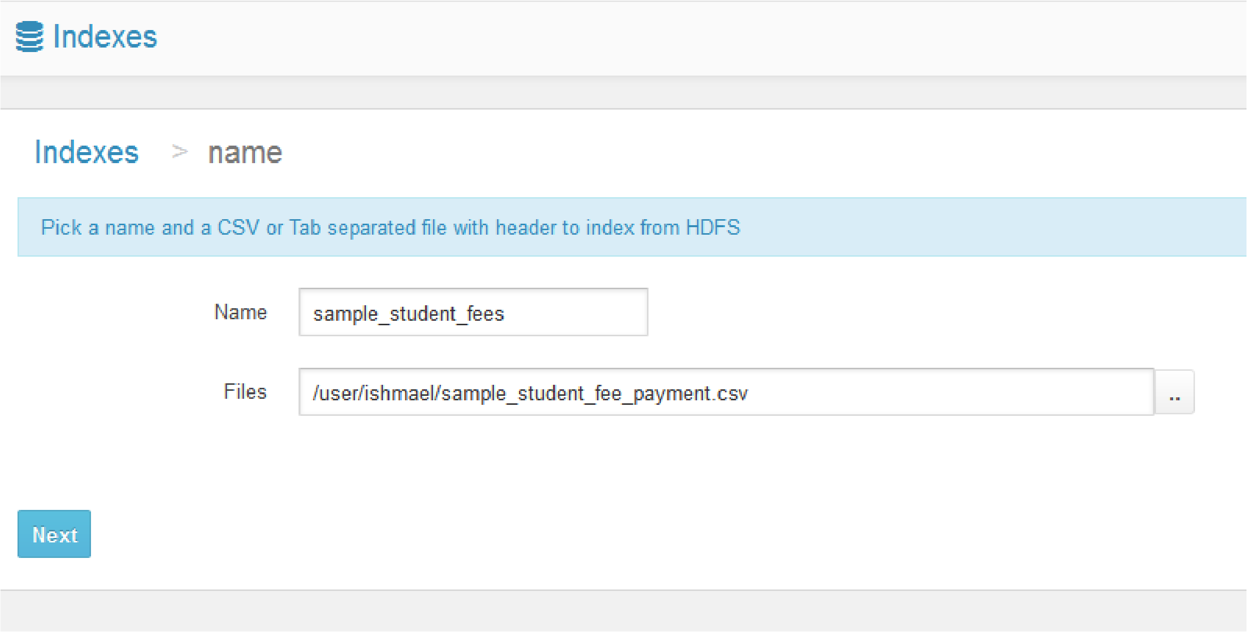
Chart type

X-Axis department

Y-Axis _c1

Figure 39: Pie Chart Report for Sample Student Fees

Step 4: Generation of Dashboard: The initial step is to create an index. This is achieved by choosing index name and selecting file with data.



Indexes

Indexes > name

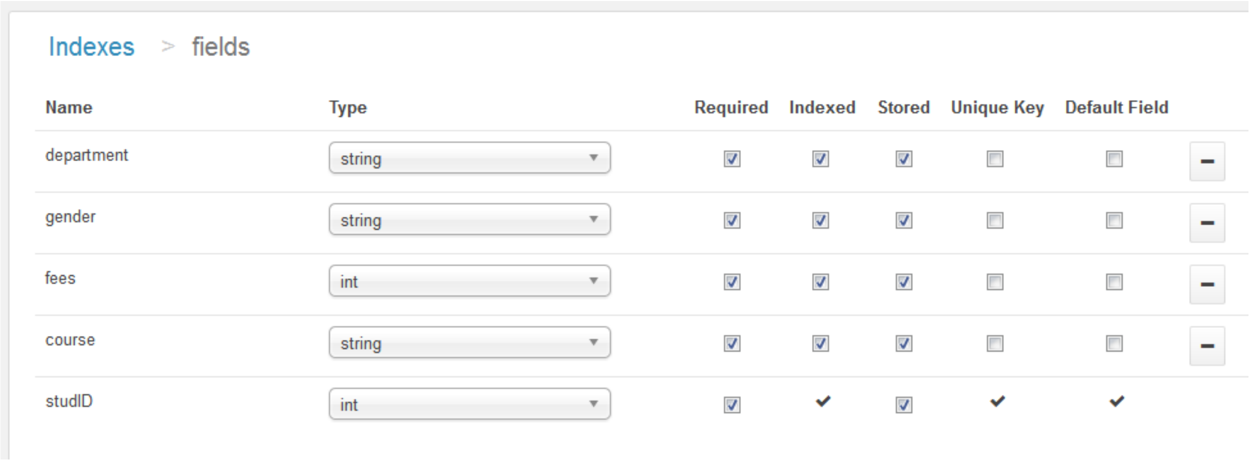
Pick a name and a CSV or Tab separated file with header to index from HDFS

Name: sample_student_fees

Files: /user/ishmael/sample_student_fee_payment.csv

Figure 40: Dashboard Generation Part 1

After picking the file, you set the index attributes.

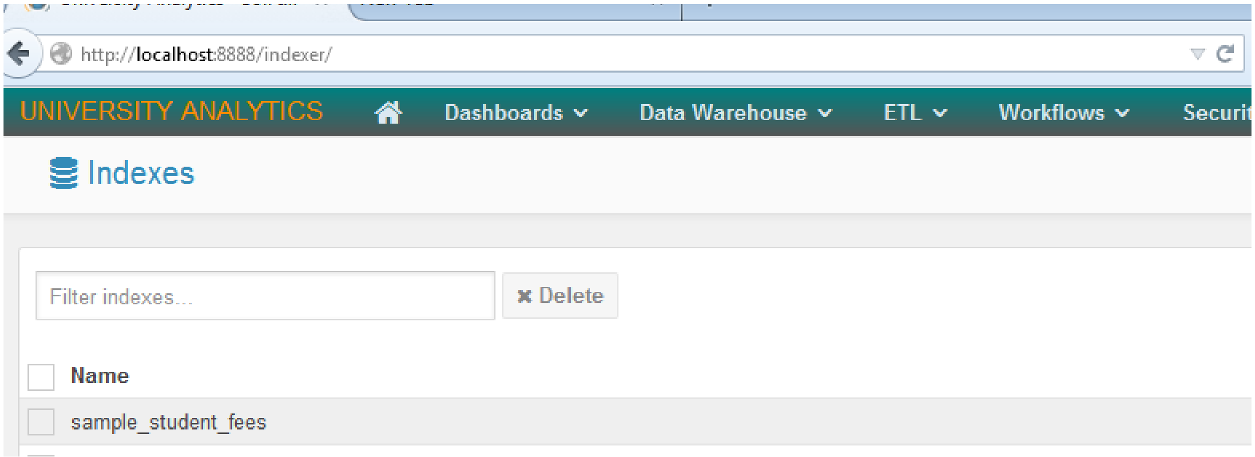


Indexes > fields

Name	Type	Required	Indexed	Stored	Unique Key	Default Field	
department	string	☑	☑	☑	□	□	−
gender	string	☑	☑	☑	□	□	−
fees	int	☑	☑	☑	□	□	−
course	string	☑	☑	☑	□	□	−
studID	int	☑	✓	☑	✓	✓	

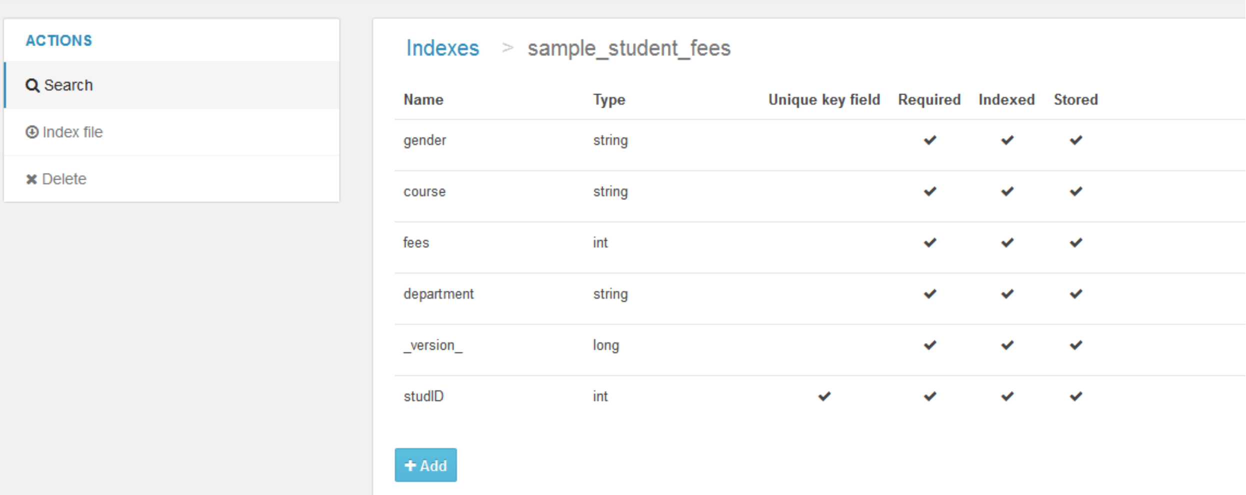
Figure 41: Dashboard Generation Part 2

Once the index is created, then it will be selected for search operation.



□	Name
□	sample_student_fees

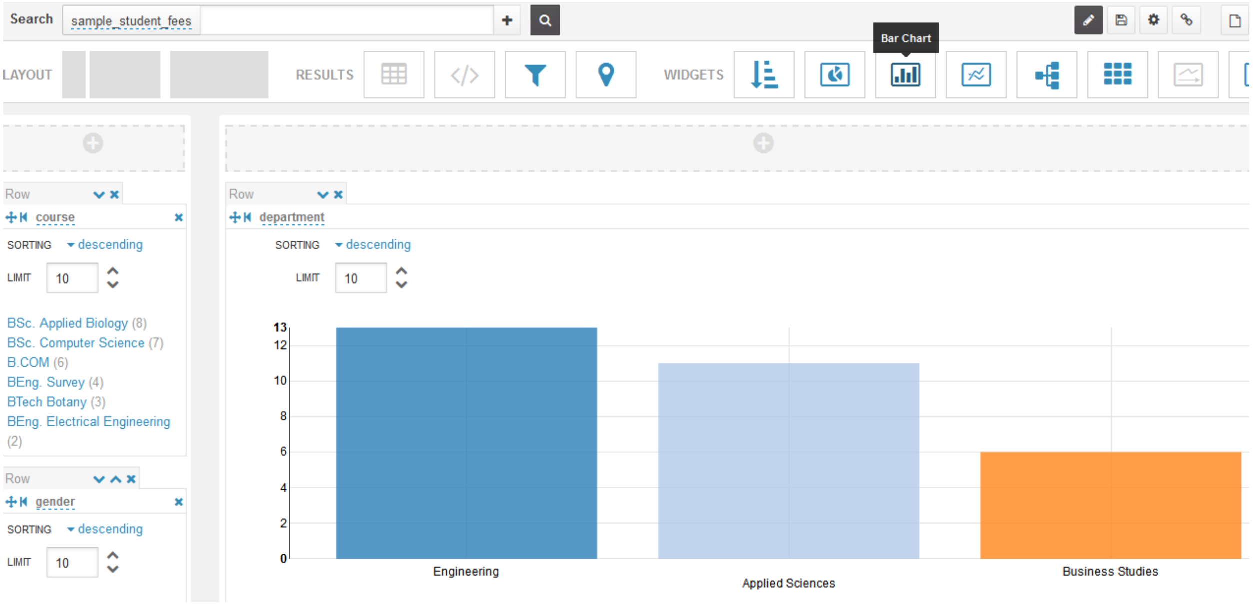
Figure 42: Dashboard Generation Part 3



Name	Type	Unique key field	Required	Indexed	Stored
gender	string		✓	✓	✓
course	string		✓	✓	✓
fees	int		✓	✓	✓
department	string		✓	✓	✓
version	long		✓	✓	✓
studID	int	✓	✓	✓	✓

Figure 43: Dashboard Generation Part 4

By selecting appropriate widgets, one can sophisticated dashboards for users.



Search: sample_student_fees

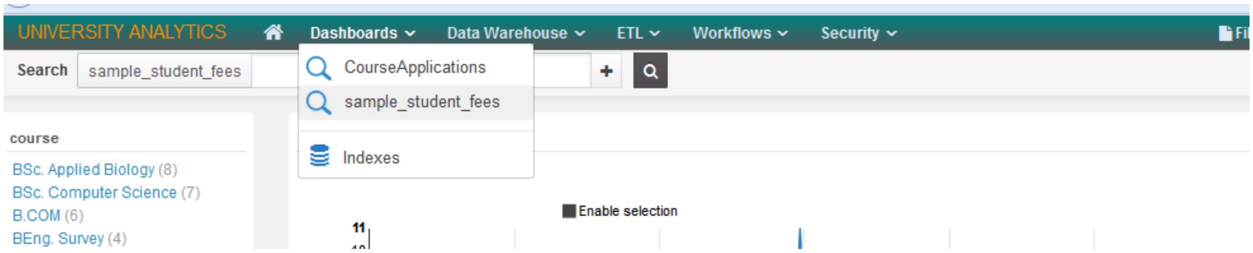
course: SORTING descending, LIMIT 10

gender: SORTING descending, LIMIT 10

department: SORTING descending, LIMIT 10

Figure 44: Dashboard Generation Part 5

Once you save the dashboard, it will be added under the ‘Dashboards’ link.



Search: sample_student_fees

Figure 45: Dashboard Generation Part 6

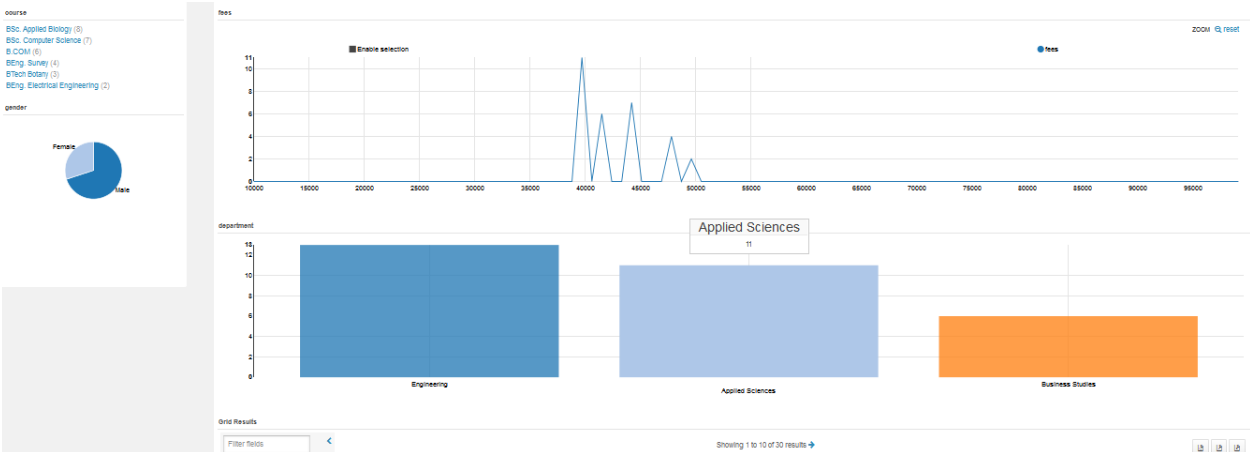


Figure 46: Dashboard Generation Part 7

4.5 Benchmarking with other State-of-Art Open Source BI Systems

The seven (7) most common open source BI platforms are Actuate, JasperSoft, OpenBI, Palo, Pentaho, SpagoBI and Vanilla (Bernardino & Figueiredo, 2014). In this section a comparison is carried to benchmark the system with some of the state-of-art BI platforms.

Bernardino & Tereso (2012) used the following features to compare open source and commercial BI tools, and hence what was used in the benchmarking.

Table 4: Business Intelligence System Indicators/ Features

Business Intelligence Indicators/ Features
▪ Reports
▪ Dashboards
▪ OLAP – Online Analytical Processing
▪ ETL – Extraction, transformation and loading
▪ Data Mining
▪ KPIs – Key Performance Indicators
▪ GEO/ GIS – Geo Information System
▪ Ad-Hoc Queries
▪ Linux
▪ Windows
▪ Unix
▪ Mobile

Table 5: Comparison with Other State-of-Art Systems

Reports	**Developed System**	**Actuate**	**JasperSoft**	**OpenBI**	**Pentaho**	**SpagoBI**	**Vanilla**	**Palo**
Dashboards	yes	yes	Yes	yes	yes	yes	yes	yes
OLAP	yes	yes	Yes	yes	yes	yes	yes	yes
ETL	yes	yes	Yes	yes	yes	yes	yes	yes
Data Mining	yes	yes	No	yes	yes	yes	yes	yes
KPIs	no	yes	Yes	yes	yes	yes	yes	yes
GEO/GIS	yes	yes	Yes	yes	yes	yes	yes	yes
Ad-hoc Queries	yes	yes	Yes	yes	yes	yes	yes	yes
Unix	yes	yes	No	yes	yes	yes	yes	yes
Linux	yes	yes	Yes	yes	yes	yes	yes	yes
Windows	yes	yes	Yes	yes	yes	yes	yes	yes
Mobile	yes	yes	Yes	yes	yes	yes	yes	yes

Apart from the above features, this system was customized to be used in a university setup with major data warehouse schema already done. It was intended to be used in by university with less modifications or configurations.

This system took into consideration the nature of data sizes and with Hadoop HDFS, it is believed that the system if implemented on physical server clusters may perform better without any hardware issues. Running on MapReduce ensures fault-free operations.

4.6 Hadoop Framework Opportunities in BI

Hadoop framework has shown great opportunity in the implementation of Business Intelligence systems. This is because of the power to store and manipulate large datasets on commodity servers. Hadoop is mainly composed of two components; they are Hadoop File System (HDFS) and MapReduce. HDFS enables the storage of larger data sets than normal file systems can handle. This is achieved by breaking down the large data into smaller units on a set of nodes. MapReduce allows the manipulation of data on each node and consequently the overall job. The programming model of MapReduce has opened up opportunities for other Hadoop projects which makes easy development of Hadoop systems. They provide tools that can be plugged into various processes required in the implementation of Hadoop BI systems.

4.7 Big Data Analytics in Universities

Since the advent of big data in the other sectors of the economy, universities in Kenya have not been left out. Increased use of information systems, social media and other unstructured data in the universities, has resulted into the phenomenon of ‘Big Data’. The term ‘Big Data’ is described by the 4Vs; Volume, Velocity, Variety and Veracity; meaning data that is huge, growing fast, in all sorts of formats, and which is uncertain. Analytics around this kind of data is seen to pick momentum. Possible analytics around big data in universities include inter-university competitive intelligence, student retention, student sponsorship ranking, donor funding and university sponsorship.

4.8 Data Analysis using R Software

R statistical software was integrated to Hadoop to achieve the transfer of data stored in the Hadoop cluster to R environment for analysis. Once data is in the R working space, several statistical analyses can be carried out. With the various integration options available as discussed above, R computational capabilities are brought right into Hadoop.

Although Hadoop has analytics capabilities, R being a mature open source statistical software, has natural ways of data analysis. R has inbuilt algorithms for data mining and prediction.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

This project study managed to implement a Business Intelligence system for universities in Kenya through the use of open source tools mainly Hadoop and R statistical software. The major components of a BI system are; data sources, extraction transformation and loading (ETL), data warehouse, data visualization and reporting, and analytics. It has been shown that Hadoop provides a great opportunity to developers who aspire to implement solutions dealing with large amounts of data typically termed as “Big Data”. Universities should take a lead to incorporate this low-cost technology into their information systems so as to able to take advantage of Business Intelligence. R is also powerful and can work well with Hadoop to analyze data for further insights. Universities have continued to automate most of their operations leading to increased amounts of data. Again with the advent of social media, a lot of useful data from these sources would be used to improve educational standards of universities in Kenya. They can only get the most of these systems when they understand data being generated. This will lead to evident-based decision-making.

Data is an asset to organizations, universities included. It is becoming a challenge to most universities, although, with the right tools and technologies it is a great opportunity. Developers can take advantage of these open source tools especially Hadoop framework and implement sophisticated solutions that can improve organizations’ performance.

However, Hadoop technology is still developing and more research is required to make it mature for development of enterprise systems. The challenge faced in this project was that Hadoop framework has a long learning curve and most developers in organizations especially in Universities may find it a bit challenging to incorporate it in their projects.

5.2 Recommendations

This project would further be developed by incorporating real-time Business Intelligence. This implies that, Hadoop should be connected to the information systems and also social media to stream live content and update dashboards in near real-time. Other Hadoop projects like Apache Flume are heading in this direction.

Developers can take advantage of social media unstructured content and aggregate to structured data of information systems so as to improve analytics in the universities. For example, a university can analyze its admission pattern at the same time harvest social media for students' interests and capabilities to plan well for future student intakes. The university can identify in advance best students with good academic backgrounds.

REFERENCES

Anandarajan, M. A., & Srinivasan, C. A. (2004). *Business Intelligence techniques – a perspective from accounting and finance.* Berlin: Springer.

Apache Hadoop. (2015). *Apache Hadoop*. Retrieved 01 20, 2015, from Hadoop: http://hadoop.apache.org/

Baars, H., & Kemper, H. G. (2008). Management support with structured and unstructured data – an integrated Business Intelligence framework. *Information Systems Management , 25* (2), 132-148.

Baepler, P., & Murdoch, C. J. (2010). Academic Analytics and Data Mining in Higher Education. *International Journal for the Scholarship of Teaching and Learning , 4* (2).

Belaire, K., Matthews, E., McInnis, A., Scime, L., & Weisensee, D. (2005). *Core Team Members*. Retrieved 01 19, 2015, from McMaster University: http://www.mcmaster.ca/bi/coreteam.htm

Bernardino, J., & Tereso, M. (2013). Business Intelligence Tools. *Computational Intelligence and Decision Making* , 267-276.

Bernardino, L. J., & Figueiredo, A. (2014). A comparative analysis of open source business intelligence platforms. *Information Systems and Design of Communication* (pp. 86-92). ACM.

Borthakur, D. (2008). *HADOOP APACHE PROJECT*. Retrieved 01 15, 2015, from HADOOP APACHE PROJECT: http://hadoop. apache. org/common/docs/current/hdfs design. pdf

Burton, B. (2009). *Toolkit: Maturity Checklist for Business Intelligence and Performance Management.* Gartner Research.

Chaudhary, S. (2004). Management factors for strategic BI success. (M.S., Ed.) *Raisinghani* .

Che-Lun, H., & Guan-Jie, H. (2014). Local Alignment Tool Based on Hadoop Framework and GPU Architecture. *BioMed Research International* .

Claessens, H. (2008). *Business Intelligence Implementation at Maastricht University*. Retrieved 01 18, 2015, from Maastricht University: www.herug2008.uct.ac.za/18%20Claessens.pdf

Clavier, P. R., Lotriet, H., & Loggerenberger, J. (2012). Business Intelligence Challenges in the Context of Goods –and Service -Domain Logic, 45th Hawaii International Conference. *System Science* (pp. 4138 -4147). IEEE Computer Society.

Cody, W. F., Kreulen, J. T., Krishna, V., & Spanler. (2002). The integration of business intelligence and knowledge management. *IBM Systems Journal , 41* (4), 697 – 713.

Condie, T., Conway, N., Alvaro, P., & Hellerstein, J. (2010). MapReduce Online. *NSDI , 10* (4), 20.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Harvard Business Press.

Dawson, L., & Van Belle, J.-p. (2013). Critical success factors for business intelligence in the South African financial services sector. *SA Journal of Information Management* , 15(1) 12-pages.

Dean, J., & Sanjay, G. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM , 51* (1), 107-113.

Dimokas, N., Mittas, N., Nanopoulos, A., & Angelis, L. (2008). A prototype system for educational data warehousing and mining. *2008. PCI'08. Panhellenic Conference* (pp. 199-203). Informatics.

Duan, Y., Cao, G., One, V., & Woolley, M. (2013). *Intelligent student engagement management: applying business intelligence in higher education*. Retrieved 1 20, 2015, from http://uobrep.openrepository.com/uobrep/handle/10547/308828

Emurugat, A. (2008). *Data warehousing: a case of a Public University in Kenya*. Retrieved 11 2, 2014, from eRepository, University of Nairobi: http://erepository.uonbi.ac.ke/handle/11295/9168?show=full

Frolick, M. N., & Ariyachandra, T. R. (2006). Business performance management: one truth. *Information Systems Management , 23* (1), 41.

Gangadharan, G. R., & Swami, S. N. (2004). Business intelligence systems: design and implementation strategies. *Information Technology Interfaces, 2004. 26th International Conference on* (pp. 139-144). IEEE.

Gartner Research. (2009, june 12). *Gartner Says Worldwide Business Intelligence, Analytics and Performance Management Grew 22 Percent in 2008*. Retrieved 01 20, 2015, from Gartner Research: http://www.gartner.com/newsroom/id/1017812

Golfareli, M. (2009). *Open source BI platforms: a functional and architectural comparison.* Berlin Heidelberg: Springer.

Golfarelli, M., & Rizzi, S. (2009). *Data warehouse design: Modern principles and methodologies.* McGraw-Hill.

Golfarelli, M., & Rizzi, S. (2009). Expressing OLAP preferences. *In Scientific and Statistical Database Management* , 83-91.

Gray, P. (2003). Business intelligence: A new name or the future of DSS. *DSS in the uncertainty of the Internet age* .

Herring, J. P. (1992). The role of intelligence in formulating strategy. *Journal of Business Strategy* , 54-60.

Kimball, R., & Ross, M. (2011). *The data warehouse toolkit: the complete guide to dimensional modeling.* John Wiley & Sons.

Madureira, L. (2012). *Computational Intelligence and Decision Making: Trends and Applications.*

Malik, S. (2005). *Enterprise dashboards design and best practices* (1 ed.). New Jersey: Wiley.

Mallach, E. G. (2000). *Decision support and data warehouse systems.* McGraw-Hill Higher Education.

McMaster University. (2005). *Business Intelligence Project*. Retrieved 01 20, 2015, from http://www.mcmaster.ca/bi

Moss, L. T., & Atre, S. (2003). *Business intelligence roadmap: the complete project lifecycle for decision-support applications.* Addison-Wesley Professional.

Muhammad, G., Ibrahim, J., Bhatti, Z., & Waqas, A. (2014). Muhammad, G., Ibrahim, J., Bhatti, Z., & Waqas, A. (2014). Business Intelligence as a Knowledge Management Tool in Providing Financial Consultancy Services. *American Journal of Information Systems* , 2(2) 26-32.

Nikolaos, D., Nikolaos, M., & Alexandros, N. (2008). A prototype system for educational data warehousing and mining. *Panhellenic Conference* (pp. 199-203). Samos: IEEE Computer Society.

Oancea, B., & Dragoescu, R. M. (2014). *Integrating R and Hadoop for Big Data Analysis.* arXiv preprint arXiv:1407.4908.

Olszak, C. M., & Ziemba, E. (2007). Approach to building and implementing business intelligence systems. *Interdisciplinary Journal of Information Knowledge and Management* , 2, 134-148.

Olszak, C. M., & Ziemba, E. (2003). Business Intelligence as a Key to Management of an Enterprise. *Informing Science Institute, Informing Science+ Information Technology Education* .

Piedade, M. B., & Santos, M. Y. (2010). Business intelligence in higher education: enhancing the teaching-learning process with a SRM system. *2010 5th Iberian Conference* (p. .). Santiago de Compostela: Information Systems and Technologies (CISTI).

Piedade, M. B., & Yasmina, M. S. (2009). Semana da Escola de Engenharia October 24-27, 2011 BUSINESS INTELLIGENCE IN HIGHER EDUCATION Promoting the students success with a SRM system.

Popovic, A., Coelho, P. S., & Jaklic, J. (2009). The Impact of Business Intelligence System Maturity on Information Quality. *Information Research* , paper 417.

Power, D. J. (2007). *A brief history of decision support systems*. Retrieved 2015, from DSSResources.COM: http://dssresources.com/history/dsshistoryv28.html

Radcliffe, P. M. (2010). *Business Intelligence at the University of Minnesota*. Retrieved 01 20, 2015, from http://university-relations.umn.edu/assets/pdf/ur_article_288616.pdf

Ranjan, J. (2009). Business intelligence: concepts, components, techniques and benefits. *ournal of Theoretical and Applied Information Technology* , 60-70.

Shim, J. P., Warketin, M., Courtney, J. F., Power, D. J., Sharda, R., & Carlsson, C. (2002). Past, present, and future of decision support technology. Decision support systems. *Decision support systems , 33* (2), 111-126.

Stackowiak, R., Rayman, J., & Greenwald, R. (2007). *Oracle Data Warehousing and Business Intelligence Solutions.* Indianapolis: Wiley Publishing, Inc.

Surajit, C., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. *Communications of the ACM , 44* (8).

Tabman, E., & Chi, R. T. (1995). Distributed intelligent executive information systems. *Decision Support Systems , 14* (2), 117-130.

Tereso, M., & Bernardino, J. (2011). Open source business intelligence tools for SMEs. *011 6th Iberian Conference* (pp. 1-4). Information Systems and Technologies (CISTI).

The R Project for Statistical Computing. (2015). *R*. Retrieved 01 20, 2015, from The R Project for Statistical Computing: http://www.r-project.org

Webopedia. (2015). *Open Source Tools*. Retrieved 01 20, 2015, from Webopedia: http://www.webopedia.com/TERM/O/open_source_tools.html

Wise, L. (2012). *Using Open Source Platforms for Business Intelligence: Avoid Pitfalls and Maximize ROI.* Newnes.

Xuejian, Y., & Li, X. (2011). A multidimensional data analysis system based on MDA for educational data warehousing. *2011 6th International Conference* (pp. 88-94). Computer Science & Education (ICCSE).

APPENDICES

Appendix 1: Interview Guide

OPEN SOURCE IMPLEMENTATION OF BUSINESS INTELLIGENCE SYSTEMS FOR UNIVERSITIES IN KENYA: A CASE OF THE TECHNICAL UINVERSITY OF KENYA

Interview Guide

This interview aimed to assess the Business Information Requirements of the organization, Data Warehouse and Business Intelligence Awareness of the Interviewee, the current Business Intelligence Maturity Level of the organization, how the organization can implement a Business Intelligence Solution, benefits of a Business Intelligence tool to the organization, challenges that the organization would face in implementing Business Intelligence Solution, and general comments concerning Business Intelligence usage.

Name of Institution:

Office/ Division:

Title:

Date:

Q1. What kind of processes do you carry out in your office?

Q2. How do you handle reporting of information to the management? What tools or technologies do you involve to achieve this? How long does it take to prepare a required report?

Q3. In your own experience, what things inform decision-making both in your office and the top management?

Q4. What kind of challenges to you encounter in preparing reports?

Q5. Do you have a central data warehouse where staff members can access the same information?

Q6. What benefits would you experience if you have a data warehouse in your organization?

Q7. What are the hindrances to having a central ‘one-source-of-truth’ data warehouse in your organization?

Q8. What kinds of reports have you being asked that were unavailable or you were unable or to provide? Provide a reason for this.

Q9. What kind of business answers would you require from a data warehouse?

Q10. Would you recommend data driven decision-making?

Appendix 2: Request Letter for Institutional Data

ISHMAEL OBONYO

DIRECTORATE OF ICT SERVICES

THE TECHNICAL UNIVERSITY OF KENYA

P.O. BOX 52428 – 00200 NAIROBI.

DATE: 3RD DECEMBER 2014

THE DEPUTY VICE CHANCELLOR, ACADEMIC RESEARCH AND STUDENTS

THE TECHNICAL UNIVERSITY OF KENYA

P.O. BOX 52428 – 00200 NAIROBI

Dear Sir,

RE: REQUEST FOR STUDENTS' AND APPLICANTS' DATA

I am a student pursuing postgraduate studies in the University of Nairobi and carrying out a research study titled "Open Source Implementation of Business Intelligence Systems in Kenyan Universities: A Case of the Technical University of Kenya". This project is aimed at investigating open source tools that can be employed by universities to develop their own BI systems without incurring costs associated with propriety BI tools. This is to request your office to allow me access and use data pertaining to students and applicants so as to enable me complete my project for the partial fulfillment of Masters of Science in Computational Intelligence, University of Nairobi.

Thanking you in advance.

Yours faithfully,

ISHMAEL N. OBONYO

TECHNOLOGIST, DIRECTORATE OF ICT SERVICES

STAFF NUMBER: NT0552