# IMPLEMENTATION OF RDBMS VIA SQOOP IN A SMALL HADOOP CLUSTER

CHIA LI YEN

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

**BORANG PENGESAHAN STATUS TESIS***

JUDUL: **IMPLEMENTATION OF RDBMS VIA SQOOP IN A SMALL HADOOP CLUSTER**

SESI PENGAJIAN: **2012/2013**

Saya **CHIA LI YEN**

mengaku membenarkan tesis (PSM) ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. ** Sila tandakan (/)

| | | |
|---|---|---|
| _____ | SULIT | (Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972) |
| _____ | TERHAD | (Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan) |
| ___/___ | TIDAK TERHAD | |

| | |
|---|---|
| _____ | _____ |
| (TANDATANGAN PENULIS) | (TANDATANGAN PENYELIA) |
| Alamat tetap: A-1-08 P/Puri Suakasih Psn Suakasih BTHO 43200 Bt 9, Cheras, Selangor. | PM Dr. Azah Kamilah binti Draman @ Muda |
| Tarikh: | Tarikh: |

CATATAN: * Tesis dimaksudkan sebagai Laporan Akhir Projek Sarjana Muda (PSM)

** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

# IMPLEMENTATION OF RDBMS VIA SQOOP IN A SMALL HADOOP CLUSTER

CHIA LI YEN

This report is submitted in partial fulfillment of the requirements for the
Bachelor of Computer Science (Database Management)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA
2015

**DECLARATION**

I hereby declare that this project report entitled

**IMPLEMENTATION OF RDBMS VIA SQOOP IN A SMALL HADOOP CLUSTER**

is written by me and is my own effort and that no part has been plagiarized without citations

STUDENT : _____ Date: _____
(CHIA LI YEN)

SUPERVISOR : _____ Date: _____
(PM DR. AZAH KAMILAH BINTI DRAMAN @ MUDA)

# DEDICATION

I dedicate this to my parents, lecturers, and friends. I earnestly feel that without their inspiration, able guidance and dedication, I would not be able to pass through the tiring process of this work.

## ACKNOWLEDGEMENT

I would like to express the deepest appreciation to my supervisor, Associate Prof Dr. Azah Kamilah binti Draman @ Muda, who introduced me to this new technology and gave me the chance to do the project. Without her guidance and support, this dissertation would not have been possible.

I would like to thank a lecturer, Associate Prof Dr. Choo Yun Huoy, who gave me some guidance throughout the project. Without her guidance, I could not complete the project so successful.

In addition, a thank you to Dr. Lim Kim Chuan, who borrow me his lab and the hardware that fulfill the system and hardware requirement of this project. Without his hardware, the project could not end successfully.

Last but not least, I would like to thank my beloved family and my friends who always support me. Thank you very much.

## ABSTRACT

Hadoop is an open source project for distributed storage and processing of large sets of data on commodity hardware. Hadoop works well with structured as well as unstructured data. Basically, Hadoop is not a database, it is a distributed file system (HDFS) to let user store large amount of data on a cloud of machines and handling data redundancy. On top of it, Hadoop provides an API for processing the stored data, which is known as Map-Reduce. The basic idea is, since the data is stored in many nodes, so better process the data in a distributed way where each node can process the data stored on it rather than spend a lot of time moving it over the network. Sqoop (SQL-to-Hadoop) is used to extract data from non-Hadoop data stores, transform the data into a form usable by Hadoop, and then load the data into HDFS. This process is called ETL, for Extract, Transform, and Load. For those existing big company that want to use Hadoop for data storage, Sqoop will be used to maintain the old existing data and bring those old data into Hadoop. Sqoop also can export the data from Hadoop to non-Hadoop data stores. Therefore, it provides bi-directional data transfer between Hadoop and non-Hadoop data stores. This project is a research on implementation of RDBMS via Sqoop in a small Hadoop cluster. The existing old data in RDBMS will be imported into Hadoop cluster by using Sqoop component of Hadoop and the new data will be inserted into HDFS of Hadoop. After the import process is carried out, an application is created and designed to show how does the old data from those standalone databases can integrate well with each other and combine with the new data that will be inserted via interface.

# ABSTRAK

Hadoop merupakan satu projek *open source* berfungsi untuk penyimpanan data secara agihan dan pemprosesan set data yang besar dalam perkakasan komoditi. Hadoop boleh berfungsi dengan data yang berstruktur dan data yang tidak berstruktur. Pada asasnya, Hadoop bukan satu pangkalan data, ia merupakan sistem fail yang diagihkan kepada pengguna untuk menyimpan saiz data yang besar pada *cloud machine* dan pengendalian data yang berlebihan. Selain daripada itu, Hadoop juga telah menyediakan *API* untuk memproses data yang telah disimpan, iaitu *MapReduce*. Idea asasnya lebih baik memproses data yang telah disimpan dalam sistem fail daripada menggunakan masa yang banyak untuk berfungsi. Sqoop *(SQL-to-Hadoop)* digunakan untuk mengambil data dari pangkalan data yang bukan Hadoop, dan menukarkan data ke dalam bentuk yang boleh digunakan oleh Hadoop dan kemudian memuatkan data tesebut ke dalam HDFS. Proses ini dipanggil ETL: *Extract, Transform and Load*. Bagi sesetengah syarikat besar yang ingin menggunakan Hadoop untuk penyimpanan data, Sqoop boleh digunakan untuk mengekalkan data lama yang sedia ada dan membawa data-data tersebut ke dalam Hadoop. Sqoop juga boleh mengeksport data dari Hadoop ke dalam pangkalan data yang bukan Hadoop. Oleh itu, Sqoop membolehkan pemindahan data dalam dua arah antara Hadoop dan RDBMS. Projek ini mengkaji prelaksanaan RDBMS melalui Sqoop dalam kelompok Hadoop yang kecil. Data yang sedia ada dalam RDBMS akan diimport ke dalam Hadoop dengan menggunakan komponen Sqoop Hadoop dan data baru akan dimasukkan ke dalam HDFS Hadoop. Selepas proses import dijalankan, aplikasi akan dibangunkan untuk menunjukkan bagaimana data lama dari pangkalan data yang *standalone* boleh mengintegrasikan dengan baik antara satu sama lain dan menggabungkan dengan data baru yang akan dimasukkan melalui aplikasi.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATION

| ABBREVIATION | DESCRIPTION |
|---|---|
| HDFS | Hadoop Distributed File System |
| YARN | Yet Another Resource Negotiator |
| CDH | Cloudera Hadoop Distribution |
| HDP | Hortonworks Data Platform |
| NFS | Network File System |
| ACID | Atomicity, Consistency, Isolation and Durability |
| SSH | Secure Shell |
| RSA | Rivest, Shamir and Adelman |
| SELINUX | Security-Enhanced Linux |
| NTP | Network Time Protocol |
| THP | Transparent Huge Pages |
| VM | Virtual Machine |

# CHAPTER I

# INTRODUCTION

## 1.1 Project Background

Hadoop is an open source project that offers another approach to store and process data. For those hoping to tackle the capability of big data, Hadoop is the platform of choice. Hadoop has two core components out of so many components, that is, file store (HDFS) to store large amount of file data on a cloud of machines, handling data redundancy and programming framework (MapReduce), an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster.

This project implemented by using five nodes (m1.hdp2, m2.hdp2, w1.hdp2, w2.hdp2 and w3.hdp2). m1.hdp2 acts as master or NameNode in HDFS, m2.hdp2 acts as Secondary NameNode or knows as checkpoint node, while w1.hdp2, w2.hdp2 and w3.hdp2 are act as slaves or DataNode in HDFS. The job of NameNode are managing namespace in HDFS and store MetaData (about the data being stored in DataNodes). Secondary NameNode purpose is to have a checkpoint in HDFS. DataNodes stores the actual Data and send the information of data stored to the NameNode.

At the same time, m1.hdp2 acts as JobTracker in MapReduce, while w1.hdp2, w2.hdp2 and w3.hdp2 are act as TaskTracker. The job of JobTracker is divide job and assigns it to TaskTracker, while TaskTracker run the tasks and report the status of task to JobTracker.

Sqoop is a tool used for data transfer from relational database into HDFS. There are three types of different relational databases are using in this project, which are, ORACLE, MySQL and SQL Server. The data will be imported from the selected relational database into Hive, which act as data warehouse in Hadoop. User can retrieve the data from Hive by using HiveQL language, which is a SQL like statement. Hive will convert the HiveQL language into a MapReduce job.

## 1.2 Problem Statement(s)

The problems are identified and described as below:

*i.* Hadoop is an open source project, so there are many vendors have developed their own distributions by adding some new functionality. Consequently, user does not know which Hadoop distribution to choose.

*ii.* Hadoop is a new platform to store and process data. But, how does it work? Are there any differences with Relational Database Management System? If there are existing data in old relational database, can Hadoop build on top of it?

*iii.* How can an application work with Hadoop?

## 1.3 Objective

The purposes to develop the project are listed as below:

Objective 1: To compare Hadoop Framework Distribution among different open source distributor (Cloudera, Hortonwork, MapR and etc.).

Objective 2: To design a database using the proposed Hadoop Framework.

Objective 3: To implement the proposed design in a prototype.

## 1.4 Project Significance

Users can see the comparison of Hadoop Framework Distribution among different open source distribution. The result of the comparison can help user to understand more about the Hadoop that developed by different open source. Users also can see how Hadoop works with its components. Besides, users also can know how the data can be imported into Hadoop from RDBMS by using Sqoop. From a prototype that is developed, users can see how well does the Hadoop work with an application.

## 1.5 Expected Output

The expected outputs from the project are listed as below:

Output 1: Can get a comparative analysis on different Hadoop Framework Distribution.

Output 2: Database design using the proposed Hadoop Framework is created.

Output 3: Simple prototype or interface to implement the database is developed.

## 1.6 Conclusion

This chapter has briefly introduced what is the project all about and how can Hadoop works with RDBMS. The objectives have clearly stated the purpose of the project to carry out and how to solve the problems found. The next chapter will discuss about the literature review of the selected services that will contribute to the project.

# CHAPTER II

# LITERATURE REVIEW

## 2.1 Introduction

In this chapter, some services will be discussed about their function and how the services contribute in a cluster. The services that will be described in this chapter are those selected to be used in this project, which is based on the need of the project. From this chapter, user can know how important for those services to be installed in the cluster.

The selected services are HDFS, MapReduce, YARN, Hive, Pig, HBase, Sqoop and ZooKeeper. An overview of a service will be described for each service. But, the main components of a Hadoop cluster are HDFS and MapReduce. Therefore, some explanation on the way how do they work in the cluster will be discussed in this chapter in more detail.

## 2.2 Apache Hadoop

As the World Wide Web grew, search engines and indexes were made to help people find relevant information. During the early years, search results were returned by humans. But as the number of web pages grew from dozens to millions, computerization was needed. Web crawlers were made, many as university-led research projects and search engine startups took off (Yahoo, etc).

One such project was Nutch – an open-source web search engine – and the brainchild of Dough Cutting and Mike Cafarella. Their goal was to invent a way to return web search results faster by distributing data and calculations across different computers so multiple tasks could be accomplished simultaneously. Also during this time, another search engine project called Google was in progress. It was based on the same concept – storing and processing data in a distributed, automated way so that more relevant web search results could be returned faster.

In 2006, Cutting joined Yahoo and took with him the Nutch project as well as ideas based on Google's early work with automating distributed data storage and processing. The Nutch project was divided. The web crawler portion remained as Nutch. The distributed computing and processing portion became Hadoop (named after Cutting's son's toy elephant). In 2008, Yahoo released Hadoop as an open-source project, and, today Hadoop's framework and family of technologies are managed and maintained by the non-profit Apache Software Foundation (ASF), a global community of software developers and contributors.

In a "normal" relational database, data is found and analyzed using queries, based on Structured Query Language (SQL). Non-relational databases use queries too,