**MANB 1163**

**Assignment 1: Mini Group Project**

**Data Analytics with Hadoop – Collect & Process Data and Build Applications with Hadoop (mark 25)**

**Lecturer: Saiful Adli Ismail**

This project will combine three component hadoop, apache hive, apache hcatalog and apache pig to form as one solution, i.e.:

* Deployment of Hadoop servers to develop, analyze and operate data directly access from Web servers.
* Present charts for extensive data analysis (if you can develop the graph of visualization that would be great). E.g using Ms Excel 2013, Ms Power BI

Each of these components plays a vital role in data analysis. In fact, data analysis itself largely contributes to a thorough decision means in organization. The internal servers could be hosting database application for data storage.

Your task is to install the **Hortonworks server into your virtualbox.** You are freely to set up your own experiment. Testing your setup and collect and process the data based on respective tutorial at http://hortonworks.com/get-started/develop/. In addition, **provide suggestion for the type of data analytic (diagnostic, prescriptive, predictive and** during data analysis from the cloud or Web. E.g (kaggle.com) data set.

**Prepare your report proposal, which includes your analysis of your data in form of chart.**

**Prepare your 10 minutes presentation (ppt) on 26 NOV 2016.**

Submit both files (report and ppt) to my email at saifuladli@utm.my before xx Oct 2017

Note: Don't procrastinate! An early start will help you do a timely and quality project.

**MANB 1163**

**Individual Assignment 2: Term Paper (mark 25)**

**Lecturer: Saiful Adli Ismail**

Write a conference paper (e.g. survey paper or taxonomy) regarding the topic given below:

1. Data Analytics in Big Data Environment (e.g. methodology, hadoop technology, Healtcare, Energy management, Retail, Telecommunication, etc))
2. Data Analytics in Cloud Computing (e.g. Data at rest, Data in motion, Data flow)
3. Big data Analytics Tools (e.g. tableau, cognosBI, vanilla-BMP, etc)
4. Data Science algorithm in industries (Internet search, Search Recommenders, Digital advertisement

The reference of the conferences and journals paper or citation **must more than 20 citation**.

**Main Components of Paper**

1 Abstract

- Abstract is a brief survey containing

- a. A little introduction to the title

- b. Problem statement

- c. Design study .

- d. Main findings

- e. Conclusion.

All these should be written in the simplest way to describe the overall study. All key variables must be recorded in the abstract but not necessarily described. Write only the design and instruments used in this study with one sample / subject without elaborating procedures review.

2 Introductions

 - grab the ATTENTION of the audience

 - start general ----> narrow to focus

 - present related or relevant background material

3 Main Body

 - methods & result

 - clearly explain experimental procedure

 - present the data that relates to main findings

4 Ending

 - Conclusion, Acknowledgements

 - repeat main finding, relate back to hypothesis

 - future directions / implications of results.

**Prepare your paper with IEEE format which includes your citation and references using citation tool such as zotero, mendeley or endnote.**

**Prepare your 10 minutes presentation (ppt) on 26 NOV 2016.**

Submit both files (paper and ppt) to my email at saifuladli@utm.my before

26 NOV 2016.

Note: IEEE paper format template will be given on the first day of class.

**MANB 1163**

**Group Assignment 3: MapReduce (mark 20)**

Write a Mapreduce program to calculate “word count” in Java using Horton works sandbox. Follow the tutorial from the http://manikumarbachu.blogspot.my/2015/08/how-to-write-first-map-reduce-program.html. Later on download mapreduce source code from http://www.guru99.com/create-your-first-hadoop-program.html and run in into hue account in the hortonwork sandbox. Prepare your report briefly step by step on how to running the mapreduce code in Hadoop environment.

Format of Report [20 mark]

1.Introduction

2.Step to running the source code into Hadoop Sandbox.

3.Explain each line of code

4.Futrue of mapreduce (what current and previous algorithm used in mapreduce?)

5. Conclusion

Demo [10 mark]

Submit both files (report and ppt) to my email at saifuladli@utm.my before

26 NOV 2016.