BINUS University

Academic Career: ***Undergraduate / ~~Master~~ / ~~Doctoral~~ *)***			**Class Program:** ***~~International~~/Regular/~~Smart Program~~/~~Global Class~~*)***
☐ **Mid Exam** ☑ **Final Exam** ☐ **Short Term Exam** ☐ **Others Exam : _____**			**Term : Odd/~~Even~~/~~Short~~ *)**
☑ **Kemanggisan** ☑ **Alam Sutera** ☐ **Bekasi** ☐ **Senayan** ☐ **Bandung** ☐ **Malang**			**Academic Year :** **2020 / 2021**
Faculty / Dept.	:	School of Computer Science	Deadline: Day / Date : Tuesday / Feb 16th, 2021 Time : 13.00
Code - Course	:	COMP6579 - Big Data Processing	Class : All classes
Lecturer	:	Team	Exam Type : Online
*) *Strikethrough the unnecessary items*			
The penalty for CHEATING is DROP OUT!!!			

I. CASE (100%)

1. **[15%]** YARN (Yet Another Resource Negotiator) is a very important component in the Big Data ecosystem.
   a. Each application under YARN will be handled by the Application Master. Explain what is the main function of the Application Master, how the mechanism will run in the event of Application Master failure?
   b. What does Negotiator mean in YARN? When the negotiations in YARN take place?

2. **[10%]** Explain how Zookeper can prevent a "split brain" condition where there are two Resource Managers active at the same time.

3. **[15%]** Hadoop provides two scripting languages i.e., PIG and HiveQL to help Big Data application developers to develop a program that use the Map Reduce programming model. What do you know about the two scripting languages? What are the advantages and disadvantages of PIG and HiveQL?

4. **[15%]** One of the NoSQL database available in the Big Data ecosystem is the Columnar Database (column-based database). What do you know about the Columnar Database? In which database operations (insert, update, or delete) a columnar database is superior to a row-based database, and in which database operations is a columnar database no better than a row-based database? Explain why this is so.

5. **[20%]** Give an example of Big Data implementation where Stream Processing is required in the Data Ingestion or Data Analytics step. What components of the Big Data ecosystem can be implemented in that case. Explain why Stream Processing capabilities are needed in this case, what is the impact if batch processing is used?

Verified by,

Kristien Margi Suryaningrum (D6414) and sent to Program on Jan 18, 2021

6. **[25%]** Create a Python script that uses Apache Spark Machine Learning Library (Spark ML lib) to classify Iris data (https://archive.ics.uci.edu/ml/datasets/iris). Use the Naïve Bayes algorithm with 70% training data and the rest for testing. Show the accuracy of the results of the classification process.

-- Good Luck --

Verified by,

Kristien Margi Suryaningrum (D6414) and sent to Program on Jan 18, 2021