

***Data Engineer Fresher***

***Final Project***

**Training Assignment**

|  |  |
| --- | --- |
| Document Code |  |
| Version | 1.0 |
| Effective Date | 20/02/2022 |

**Hanoi, 02/2022**

RECORD OF CHANGES

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Effective Date | Change Description | Reason | Reviewer | Approver |
| 1 | 2022-02-06 | Create first version |  | TrangND2 | VinhNV |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Contents

[Final Project 5](#_Toc96284807)

[Problem 1: User Segment 5](#_Toc96284808)

[Objectives: 5](#_Toc96284809)

[Problem Descriptions: 5](#_Toc96284810)

[The marketing team requests Data Engineer to build a User Segment that contains some information about customers. So marketers could divide the users into some groups and send the notification to them. Your mission is to implement this. 5](#_Toc96284811)

[ Download dataset from this link: 5](#_Toc96284812)

[https://kafka.apache.org/quickstart 5](#_Toc96284813)

[ Read requirements in the file User\_Segment.xlsx 5](#_Toc96284814)

[ Choose databases and design the tables 5](#_Toc96284815)

[ Design and implement the Data Pipeline to: 5](#_Toc96284816)

[o Sync data daily from source folder to destination folder. 5](#_Toc96284817)

[o Transform data from destination folder to the final tables 5](#_Toc96284818)

[ Deploy Data Pipeline on the Airflow 5](#_Toc96284819)

[Assumptions: 5](#_Toc96284820)

[Technical Requirements: 5](#_Toc96284821)

[Questions to answer: 5](#_Toc96284822)

[Submit: 5](#_Toc96284823)

[ A graph to show the data pipeline architecture 5](#_Toc96284824)

[ A document to show the idea (optional) 5](#_Toc96284825)

[ A repository containing Spark codes and how to set up the environment for testing 5](#_Toc96284826)

[ A repository containing Airflow codes and how to set up the environment for testing 5](#_Toc96284827)

[Estimated Time to complete: 1 week 5](#_Toc96284828)

|  |  |
| --- | --- |
| D:\FSOFT\SERVER GST\86.WIP\2.Template\6_Logo\Logo FA\Logo FA New-02.png | **CODE : Final Project**  **TYPE : <Type of Final Project>**  **LOC : <Lines of Code>**  **DURATION : 1 Week** |

# Final Project

Problem 1: User Segment

Objectives:

In this final project, you have to work with Spark, Airflow and Databases. You will work in groups, each group consists of 3-4 people.

Problem Descriptions:

The marketing team requests Data Engineer to build a User Segment that contains some information about customers. So marketers could divide the users into some groups and send the notification to them. Your mission is to implement this.

* Read requirements in the file User\_Segment.xlsx
* Choose databases and design the tables
* Design and implement the Data Pipeline to:
  + Sync data daily from source folder to destination folder and save as parquet files.
  + Transform data daily from destination folder to the final tables
* Deploy Data Pipeline on the Airflow

Assumptions:

Technical Requirements:

* Spark
* Airflow
* Databases

Questions to answer:

Submit:

* A graph to show the data pipeline architecture
* A document to show the idea (optional)
* A repository containing Spark codes and how to set up the environment for testing
* A repository containing Airflow codes and how to set up the environment for testing

Estimated Time to complete: 1 week

**-- THE END --**