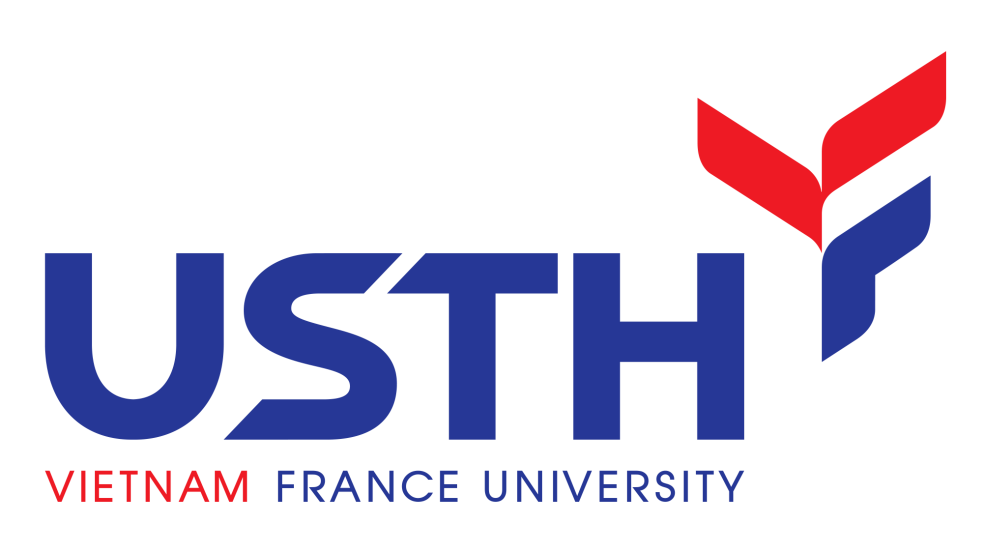
# DECLARATION

UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI

****

**MASTER THESIS**

By

**DANG THAI SON (2440045)**

*Information and Communication Technology (ICT)*

Title

**ETL Microservice for Data Ingestion into a Data Lake**

------------------------------------------------------------------------------------------------------------------

Supervisor: **Assoc. Prof. Trần Giang Sơn**

*ICT Laboratory - USTH*

**Hanoi, August 2025**

I, hereby, Dang Thai Son, declare that all the work and result in this thesis are entirely my own and are not plagiarized from any source. This thesis was written based on my research which was carried out at the Information and Communication Technology laboratory at University of Science and Technology of Hanoi, under the guidance of Assoc. Prof. Tran Giang Son.

Any scientific result, method, comment, and statistics inherited during the research from other authors has been cited thoroughly. In case there is any plagiarism in my thesis, I understand that this work will not be evaluated and I will take full responsibility for penalties from the thesis defense committee and my university.

# ACKNOWLEDGEMENTS

I would like to express my gratitude to all those who offered me the support to be capable of carrying out my project and then complete this thesis.

First and foremost, I cannot begin to express my thanks to Assoc. Prof. Tran Giang Son, my supervisor, for his invaluable guidance and relentless support throughout my internship. With his expert advice along with immerse patience, and considerable encouragement, I could have a thorough look of the project plan and then define proper methods to execute it. Had it not been for him, I would not have been able to have a grasp of the topic. Working with Assoc. Prof. Tran Giang Son was a tremendous opportunity for me to strengthen not only my knowledge but also my professional experience, which shall carry me a long way in the future.

In addition, I would like to extend my sincere gratitude to all the members of my thesis committee for their time and effort on reviewing my work. Their insightful comments and constructive criticism shall greatly improve the quality of this project.

Last but not least, I very much appreciate all staff and lecturers in the Information and Communication Technology Department at University of Science and Technology of Hanoi for foundational courses in the master program which serves as a decisive part in this research. Besides, I would like to thank Mr. Le Nhu Chu Hiep specifically, for his active cooperation and additional critical information about the ULake during the internship. The completion of my dissertation would not have been possible without these enormous helps.

# TABLE OF CONTENTS

[DECLARATION 1](#_Toc206629169)

[ACKNOWLEDGEMENTS 3](#_Toc206629170)

[TABLE OF CONTENTS 4](#_Toc206629171)

[LIST OF ABBREVIATIONS 6](#_Toc206629172)

[LIST OF TABLES 7](#_Toc206629173)

[LIST OF FIGURES 8](#_Toc206629174)

[ABSTRACT 9](#_Toc206629175)

[INTRODUCTION 10](#_Toc206629176)

[I/ Overview (Include related work = literature review) 10](#_Toc206629177)

[1. Data lake 10](#_Toc206629178)

[2. ETL 10](#_Toc206629179)

[3. Purpose: ULake hiện tại thiếu ETL cho landsat =>> Objective 10](#_Toc206629180)

[II/ Objectives 10](#_Toc206629181)

[MATERIALS AND METHODS 11](#_Toc206629182)

[I/ Tools 11](#_Toc206629183)

[1. Quarkus framework 11](#_Toc206629184)

[2. Postgresql 11](#_Toc206629185)

[a) Use-case diagram 11](#_Toc206629186)

[b) Sequence diagrams 11](#_Toc206629187)

[3. gdal\_translate 11](#_Toc206629188)

[II/ Dataset 11](#_Toc206629189)

[1. Mô tả dữ liệu vệ tinh 11](#_Toc206629190)

[2. Mô tả nguồn dữ liệu mình dùng 12](#_Toc206629191)

[a) Copernicus là gì? Bao gồm cả Landsat/Sentinel 12](#_Toc206629192)

[b) Tại sao cần: mục đích cụ thể của các vệ tinh từ Copernicus 12](#_Toc206629193)

[c) Nếu ko có thì sao? 12](#_Toc206629194)

[III/ Methods 12](#_Toc206629195)

[1. Microservice 12](#_Toc206629196)

[2. Data Lake 12](#_Toc206629197)

[3. ETL microservice flow 12](#_Toc206629198)

[RESULT AND DISCUSSIONS 14](#_Toc206629199)

[1. Result 14](#_Toc206629200)

[2. Discussion 14](#_Toc206629201)

[CONCLUSION 15](#_Toc206629202)

[REFERENCES 16](#_Toc206629203)

# LIST OF ABBREVIATIONS

# LIST OF TABLES

# LIST OF FIGURES

No table of figures entries found.

# ABSTRACT

**Keywords**: *PHP, Magento, E-commerce, Module, Order process, Knockout.js*.

# INTRODUCTION

## I/ Overview (Include related work = literature review)

### 1. Data lake

- Là gì

- Tại sao lại cần

- Nếu ko có thì nnao

- Các data lake phổ biến

- Các nguồn dữ liệu lớn phổ biến (sentinel, landsat, dữ liệu vệ tinh của Nhật …)

### 2. ETL

- Là gì

- Tại sao cần

- Nếu ko có thì nnao

- Một số ví dụ về ETL phổ biển

### 3. Purpose: ULake hiện tại thiếu ETL cho landsat =>> Objective

## II/ Objectives

- Xây dựng 1 microservice cho data ingestion ingest dùng ETL process =>> Sẽ giúp ích cho các ứng dụng phân tích và giúp quá trình thu thập dữ liệu cho các machine learning model thuận tiện hơn

# MATERIALS AND METHODS

## I/ Tools

### 1. Quarkus framework

- Là gì

- Why: phục vụ tốt khi làm việc với dữ liệu và tương thích với data lake hiện tại

- Cụ thể ở project: include extension, code style and structure

### 2. Postgresql

- Là gì

- Why: giúp việc lưu trữ dễ dàng

- How: Nếu ko có thì so sánh với usual mysql

#### a) Use-case diagram

#### b) Sequence diagrams

### 3. gdal\_translate

- Là gì

- Tại sao cần: cho bước Transform

- Nếu ko có thì nnao

## II/ Dataset

### 1. Mô tả dữ liệu vệ tinh

- Landsat/Sentinel/Ảnh vệ tinh nói chung là gì?

- Tại sao cần?

- Nếu ko có thì nnao?

### 2. Mô tả nguồn dữ liệu mình dùng

#### a) Copernicus là gì? Bao gồm cả Landsat/Sentinel

#### b) Tại sao cần: mục đích cụ thể của các vệ tinh từ Copernicus

- Trong project dùng landsat hay sentinel, vệ tinh gì, theo dõi dữ liệu của vùng nào, với mục đích gì

#### c) Nếu ko có thì sao?

## III/ Methods

### 1. Microservice

- Là gì (cụ thể ở đây là package usth.m1)

- Tai sao lai can microservice

- Nếu ko có thì phải dùng cgi thay thế? Bất tiện hơn nnao

### 2. Data Lake

- Loại data lake mình dùng là gì: Ulake

- Mô tả figure của data lake (mail thầy gửi)

- Chỉ ra microservice của mình nằm ở chỗ nào

### 3. ETL microservice flow

- Mô tả luồng chung: pull => save locally => upload  User requirements và use case diagrams (UML)???

- Extract (pull process)

- Setup token cho mỗi lần request

- Dữ liệu extract: ảnh true-color, ảnh full band, metadata

- Khi lưu local: ảnh theo folder, metadata vào postgresql + time info + directory info

- Sequence diagram?

- Transform (preprocess: downsize image)

- Problem: raw images can be very large => need a downsize technique to be able to upload

- Nếu ko có thì quy trình sẽ phức tạp hơn: lưu ảnh ở cloud và chỉ save link ở lake

- Sequence diagram?

- Load (upload to lake)

- Retrieve token cho mỗi lần call

- Dùng file/folder microservice (nhắc về bất lợi Ulake nên ko dùng đc Object)

- Thông tin lưu: ảnh kèm 1 số thông tin cơ bản (nhắc về bất lợi Ulake nên ko upload đc metadata)

- Sequence diagram?

# RESULT AND DISCUSSIONS

### 1. Result

- Hoạt động nnao?

- Extract đc loại dữ liệu gì

- Dữ liệu preprocess thành công ko? Kết quả đầu ra preprocess nnao?

- Upload dữ liệu vào datalake đc ko? Du lieu gi?

### 2. Discussion

Magento is a complex framework due to its dense design patterns and programming techniques. If considering only the functions of this project, it may seem simple and minor. However, to finish the module without being familiar with Magento is impossible. Actually, according to the timeline, two thirds of the internship is spent to understand just the most fundamental parts of the platform. Therefore, as a whole, this project is not straightforward at all. Even when I was able to implement all use cases, there are still some bugs that I could not manage to fix including failure to deal with complex products and to show correct price form in the search area. I did not have enough time to remove those flaws, but I was suggested to improve the controller for the former and to develop a new data binding for the latter in the future.

# CONCLUSION

- Overview: Paraphrase lại intro

- Future works

- Tạo log để theo dõi quá trình ETL

- Tạo cronjob chạy hàng ngày cho ETL để theo dõi 1 vài khu vực nhất định phục vụ nghiên cứu

- Thêm các bước preprocess để đảm bảo dữ liệu sạch: deduplication, avoid các ảnh trống (thường các ảnh này sẽ chỉ toàn màu đen. Dẫn chứng 1 ảnh).

# REFERENCES

- Tầm 15 mục, chủ yếu sẽ nằm ở intro