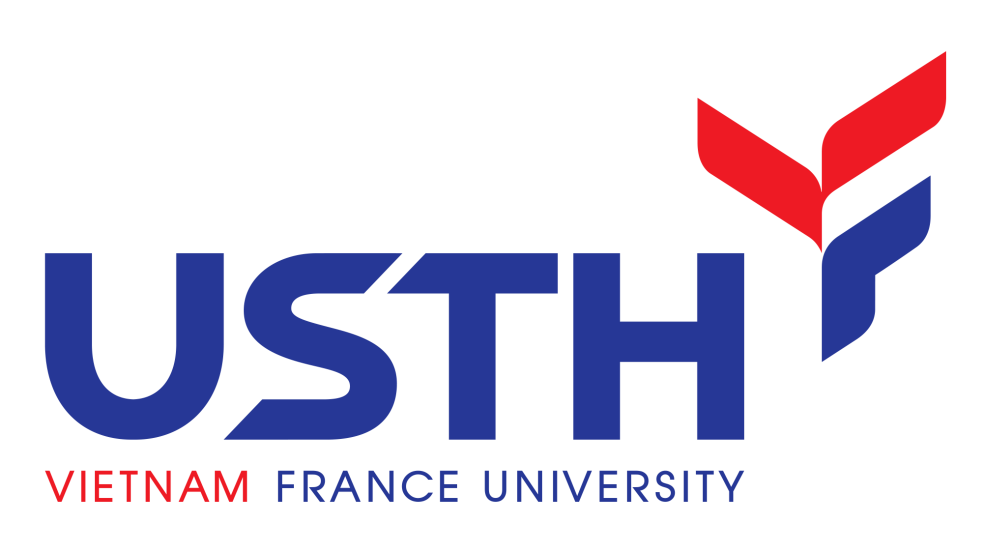
# DECLARATION

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 chances 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 their overwhelming support throughout 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](#_Toc206973276)

[ACKNOWLEDGEMENTS 3](#_Toc206973277)

[TABLE OF CONTENTS 4](#_Toc206973278)

[LIST OF ABBREVIATIONS 5](#_Toc206973279)

[LIST OF TABLES 6](#_Toc206973280)

[LIST OF FIGURES 7](#_Toc206973281)

[ABSTRACT 8](#_Toc206973282)

[INTRODUCTION 10](#_Toc206973283)

[I/ Overview & Purpose 10](#_Toc206973284)

[II/ Objectives 11](#_Toc206973285)

[MATERIALS AND METHODS 12](#_Toc206973286)

[I/ Tools 12](#_Toc206973287)

[1. Quarkus framework 12](#_Toc206973288)

[2. Postgresql (persistence layer) 12](#_Toc206973289)

[a) Use-case diagram (ETL pipeline, microservice architecture, data flow) 12](#_Toc206973290)

[b) Sequence diagrams 12](#_Toc206973291)

[3. gdal\_translate 12](#_Toc206973292)

[II/ Dataset 12](#_Toc206973293)

[1. Mô tả dữ liệu vệ tinh 12](#_Toc206973294)

[2. Mô tả nguồn dữ liệu mình dùng 13](#_Toc206973295)

[a) Copernicus là gì? Bao gồm cả Landsat/Sentinel 13](#_Toc206973296)

[b) Tại sao cần: mục đích cụ thể của các vệ tinh từ Copernicus 13](#_Toc206973297)

[c) Nếu ko có thì sao? 13](#_Toc206973298)

[d) Phương thức truy cập: Catalog/Process APIs 13](#_Toc206973299)

[III/ Methods 13](#_Toc206973300)

[1. Microservice 13](#_Toc206973301)

[2. Data Lake 13](#_Toc206973302)

[3. ETL microservice flow 13](#_Toc206973303)

[RESULT AND DISCUSSIONS 15](#_Toc206973304)

[1. Result 15](#_Toc206973305)

[2. Discussion 15](#_Toc206973306)

[CONCLUSION 16](#_Toc206973307)

[REFERENCES 17](#_Toc206973308)

# LIST OF ABBREVIATIONS

IDC International Data Corporation

AI Artificial Intelligence

ETL Extract, Transform, Load

GA4 Google Analytics 4

ICT Information and Communication Technology

CDSE Copernicus Data Space Ecosystem

RDBMS Relational Database Management System

SQL Structured Query Language

GDAL Geospatial Data Abstraction Library

TIFF Tagged Image File Format

NASA

USGS

ESA

ALOS World 3D

JAXA

MODIS

# LIST OF TABLES

# LIST OF FIGURES

No table of figures entries found.

# ABSTRACT

**This should be written last.**

1. WHY: Context and Problem (1-2 sentences)
   1. Context: The importance of data lakes
   2. Specific problem the project solves: The challenge of ingesting complex, authenticated data
   3. Keywords: Data Lake, data ingestion, analytics, machine learning, diverse data sources, authentication, pre-processing.
2. WHAT: Aim and Objectives (1 sentence)
   1. What is set out to achieve?
   2. Example: "This thesis presents the design and development of a robust ETL microservice for..."
3. HOW: Methodology and Implementation (2-3 sentences)
   1. The technology stack (Reactive microservice, Quarkus)
   2. The data source and APIs (Copernicus Data Space Ecosystem, Sentinel Hub Catalog/Process APIs)
   3. The persistence layer (PostgreSQL for metadata)
   4. The overall process (Extract STAC metadata, download imagery, persist records)
4. SO WHAT: Results and Significance (1-2 sentences)
   1. What was the final outcome? What was successfully built?
   2. Significance of the work. Why does it matter? What contribution does it make?
   3. Keywords: Functional pipeline, improved data availability, data quality, robust solution, scalable architecture, blueprint for modern data services.

Data lakes are central repositories for storing vast amounts of raw data for analytics and machine learning applications. However, a significant challenge lies in the automated and reliable ingestion of data from diverse and complex sources, especially those requiring authentication and extensive pre-processing. This thesis addresses this challenge by presenting the design, development, and implementation of a robust ETL microservice for the automated ingestion of satellite imagery from the Copernicus Data Space Ecosystem into a data lake. The solution is architected as a reactive microservice using the Quarkus framework, leveraging the Sentinel Hub Catalog and Process APIs to extract STAC metadata and corresponding satellite imagery. The extracted metadata is transformed and loaded into a PostgreSQL database, creating a searchable catalog linked to the locally stored image files. The primary outcome is a fully functional, non-blocking ETL pipeline capable of handling API authentication, processing geospatial queries, and atomically persisting both the metadata and the image data. This work provides a practical and scalable blueprint for building modern data ingestion services, improving data availability and quality within the data lake ecosystem.

* Word Count: Aim for 150-300 words.
* .

**Keywords**: *..*.

# INTRODUCTION

## I/ Overview & Purpose

In today’s modern world, data grows exponentially and continuously in numerous domains. According to a report by International Data Corporation (IDC) and Seagate, the global datasphere which is the total amount of data created and duplicated all over the world is estimated to reach 163 zettabytes (1.63 × 1014 gigabytes) by 2025 [1]. This statistic is approximately ten times bigger than that of a decade ago. This data deluge results in a new terminology called “big data” which generally involves enormous, fast-moving, and diverse datasets. Thanks to it, various information-based applications has been developed to serve critical needs such as analytical models in agriculture, or forecast frameworks in weather and climate fields [2]. Some popular big data systems supporting scientific research include Copernicus Global Land Services, Sentinel-2, Landsat Program, ALOS World 3D, and ERA5 that provide valuable environmental information. However, coming with major advantages, big data has raised 3 V challenges (Volume. Variety, Velocity) to traditional data systems, such as relational databases, which struggle with issues like scalability, handling unstructured formats, and real-time processing. Specifically, traditional systems may face performance bottlenecks and storage constraints when processing petabytes of data and managing the velocity of streaming inputs [3]. These limitations have urged for the development of new solutions, one of which is data lake.

In order to conveniently store, manage, and extract value from big data where sources are massive and complex, a new type of data storage architecture which is data lake has emerged. A data lake is defined as a highly scalable centralized repository that allows storing structured, semi-structured, and unstructured data in its raw format without the need to define a schema at the time of ingestion. In other words, this schema-on-read mechanism provide a flexible way to store and access heterogenous and high-volume data coming from sources like satellites, sensors, or user logs, without forcing early transformation [4]. Without such scalable storage architecture, traditional systems would face challenges when working with modern data streams, which would drastically reduce the ability to implement cutting-edge AI models or to fulfil analytical work. To seal that gap, several large-scale data lakes have been developed and widely adopted for either scientific or industrial use such as Amazon S3-based data lake, Google Cloud Storage, and Microsoft Azure data lake.

In the context of data lake, ETL process (Extract, Transform, Load) is widely considered the backbone of the entire architecture since it performs the core operations needed to deliver up-to-date and consistent data for analysis and decision-making. As the name suggested, ETL is a sequential three-phase data integration procedure. Firstly, diverse raw data from multiple sources, for example structured databases, GA4 statistics, or admin transaction logs, is extracted. Then using methods like format conversion, deduplication, and resizing, it is transformed into a clean and consistent format before being loaded into a destination system, here is the data lake, for analytical platform or model training purposes [5]. Its importance lies in the fact that it not only automates data movement from source systems to target repositories but also ensures that data is clean, and compliant with business rules before it is stored. Without a properly implemented ETL process, organizations could have problems with data inconsistencies, schema mismatches, and loss of traceability, especially when working with high-volume or heterogeneous sources [6]. Take major data lake platforms, Amazon S3, Google Cloud, and Microsoft Azure, as examples, they all use ETL pipelines which are AWS Glue, Google Cloud Dataflow, and Azure Data Factory respectively for loading clean and structured data. Overall, ETL is indispensable for ensuring raw and unstructured data becomes reliable and usable for downstream applications.

In recent years, the acceleration of climate change and environmental degradation has boosted the global demand for reliable Earth observation and forecasting systems. To understand and respond to these needs, data scientists and analysts must work with high-resolution satellite imagery and geospatial datasets which are enormous, heterogenous, and frequently update. These datasets which are considered as a big data source then need a scalable infrastructure to be ingested, stored, and processed effectively. As previously discussed, data lakes and ETL pipelines are fundamental for such workflow. At the ICT Lab, a centralized data lake infrastructure called ULake has been established, supporting ingestion pipelines for Kaggle and GitHub. However, there is currently no dedicated ETL process for ingesting satellite data into the lake, which limits the lab’s ability to fully integrate remote sensing data into its analytical workflows. Therefore, the goal of this project is to fulfil the need by developing an ETL microservice for data ingestion for the ULake that automates the end-to-end process of retrieving, transforming, and storing satellite data.

## II/ Objectives

The objective of this project is to design and implement a microservice-based ETL pipeline used for the ingestion of satellite imagery into the ULake data lake. The microservice supports interactions with satellite data providers, here specifically is the Copernicus Data Space Ecosystem (CDSE), to extract relevant imagery and metadata. It will then perform downsizing for full-band-color images as a preprocessing step before loading the data into appropriate storage locations within ULake. By doing so, this project will not only fill a current technical gap in the laboratory’s data infrastructure, but also enhance the efficiency and accessibility of downstream applications such as climate modeling and machine learning pipelines. On top of that, the ETL microservice will contribute to a streamlined, reusable, and scalable data ingestions job for Earth observation datasets in research and practical applications.

# MATERIALS AND METHODS

## I/ Tools

### 1. Quarkus framework

The microservice developed in this project was built using Quarkus, a modern full-stack Java framework optimized for building cloud-native, container-first, and reactive applications. Quarkus, which supports both imperative and reactive programming models, provides high performance and low memory footprint, making it well-suited for event-driven architectures and data-intensive workflows [7]. Therefore, one main reason why Quarkus was selected is its efficiency in handling data streams and non-blocking I/O operations, which is especially suitable when working with large volumes of satellite imagery and metadata from external APIs. Additionally, this choice ensured compatibility with the existing ULake ecosystem where other microservices were developed using Quarkus, which allows consistent in deployment and configuration across the data lake.

Several Quarkus extensions utilised to support the system’s functionality include: quarkus-oidc-client for secure OAuth2-based communication with external APIs, quarkus-hibernate-reactive-panache to simplify entity-repository mappings using reactive Hibernate, and quarkus-rest-jackson for handling RESTful endpoints and JSON serialization. The ETL pipeline was made following a modular structure comprising entity, model, proxy, repository, resource, and service packages with reactive programming principles which leverages Mutiny to ensure non-blocking behaviour across service layers.

### 2. PostgreSQL

In this project, PostgreSQL was used as a local relational database to store structured metadata extracted during the first phase of the ETL process. PostgreSQL is an open-source relational database management system (RDBMS) known for its stability, strong community support, and adherence to SQL standards [8]. It can handle structured tabular data which is an appropriate choice for storing metadata fields such as image identifiers, capture timestamps, source URLs, coordinates, and file paths related to downloaded satellite imagery. The primary reason for choosing PostgreSQL was to save metadata locally during the development and testing of the ETL microservice, which enabled intermediate storage between extraction and transformation phases, and supported data traceability and debugging. Without such tool, metadata shall need to be stored directly in flat files which are inconsistent, hard to query, and prone to loss if the system crashes.

### 3. GDAL translate utility

To support the transformation phase of the ETL process, this project took advantage of the gdal\_translate utility, which is a part of the Geospatial Data Abstraction Library (GDAL), a widely used open-source toolkit for reading, writing, and converting raster geospatial data formats. GDAL supports over 150 raster formats and capabilities for re-projection, resampling, and data compression. Specifically, gdal\_translate in employed in this project to downsize large multi-band satellite images by reducing their spatial resolution while preserving all spectral bands. This step is critical to ensure the data lake input file size remain below the upload threshold, allowing successful ingestion into the ULake. Without this tool or any similar one in the preprocessing step, raw satellite imagery, especially full-band TIFF would exceed storage or transfer limitations set by the data lake’s ingestion interface, which may result in failure or performance degradation, and then block the data ingestion workflow.

## II/ Dataset

### 1. Satellite imagery

Satellite imagery refers to images of Earth or other planets captured by imaging satellites operated by governments or commercial providers. These images are typically obtained using onboard sensors that record electromagnetic radiation reflected or emitted from the Earth's surface. Depending on sensor specifications, satellite images can include multiple spectral bands, enabling advanced analysis of land cover, vegetation health, water bodies, and atmospheric conditions [8]. This type of data plays a vital role in a wide range of scientific and environmental applications, including climate monitoring, agricultural forecasting, and disaster management.

In this project, satellite imagery serves as the primary raw data source extracted and processed through the ETL pipeline. These images are essential for enabling downstream geospatial analysis and model training tasks. Without access to satellite data, people would face serious limitations in observing environmental patterns, particularly in regions lacking ground sensors. The ability to capture large-scale and repeatable observations makes satellite imagery irreplaceable for Earth observation and global-scale analysis.

Some of the most widely used open-access satellite data programs include the Landsat program operated by NASA and USGS which has provided continuous Earth observation since the 1970s, and the Sentinel constellation operated by ESA under the Copernicus program which offers high-resolution multispectral data. Other valuable sources include the ALOS World 3D dataset by JAXA and MODIS imagery by NASA.

### 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?

#### d) Phương thức truy cập: Catalog/Process APIs

## III/ Methods

### 1. Microservice

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

- Tai sao lại cần microservice

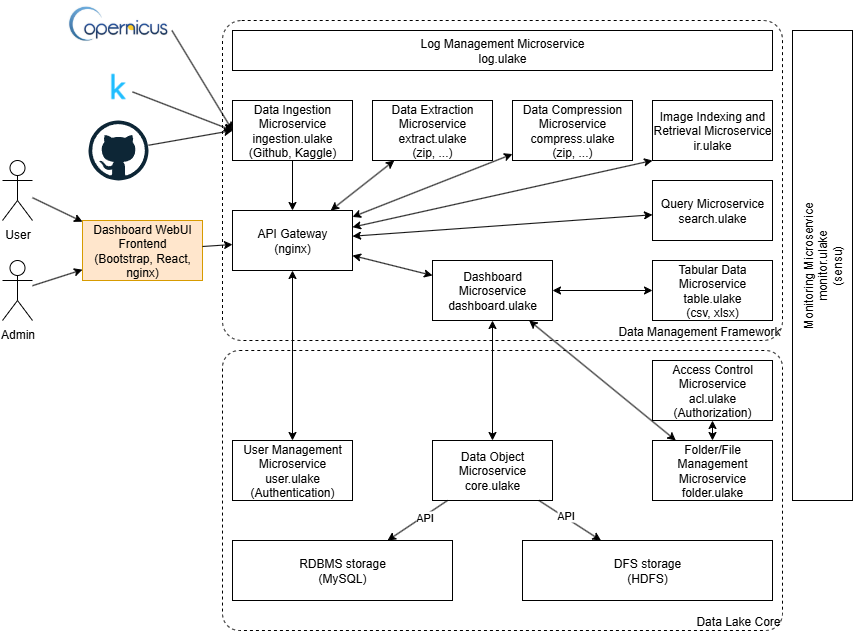
- Nếu ko có thì phải dùng cgi thay thế? Bất tiện hơn nnao

### 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

#### a) Use-case diagram (ETL pipeline, microservice architecture, data flow)

#### b) Sequence diagrams

- 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, (STAC) 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

[1] <https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf>

[2] <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00659-3>

[3] <https://www.wecmelive.com/open-access/big-data-vs-traditional-data-data-warehousing-ai-and-beyond.pdf>

[4] <https://www.sciencedirect.com/science/article/pii/S0306437924001182>

[5] <https://journalofcloudcomputing.springeropen.com/articles/10.1186/s13677-023-00571-y>

[6] <https://www.researchgate.net/publication/340938742_ETL_in_Near-real-time_Environment_A_Review_of_Challenges_and_Possible_Solutions>

[7] <http://103.203.175.90:81/fdScript/RootOfEBooks/E%20Book%20collection%20-%202025%20-%20C/CSE%20%20IT%20AIDS%20ML/Reactive_Systems_in_Java_Resilient,_Event_Driven_Architecture_with.pdf>

[8] <https://scispace.com/pdf/the-postgres-next-generation-database-management-system-506hq9wdwc.pdf>

[9] <https://api.pageplace.de/preview/DT0400.9781351659116_A43723976/preview-9781351659116_A43723976.pdf>

- Tầm 15 mục, chủ yếu sẽ nằm ở intro