ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

Trường Đại học Khoa học Tự nhiên

Đề tài

**THỐNG KÊ - DỰ ĐOÁN**

**GIÁ VÀNG**

**A blue circle with icons and a person standing in front of a diagram

Description automatically generated**

Ngành học: **LÝ THUYẾT XÁC SUẤT THỐNG KÊ**

Môn học: **KỸ THUẬT XỬ LÝ DỮ LIỆU LỚN**

Giảng viên: PGS. TS. Nguyễn Thanh Bình

Học viên: Trần Thị Thuận – 23C23002

Nguyễn Bích Trâm – 23C23010

Đỗ Thị Thanh Thảo – 23C23009

Nguyễn Trường Giang – 23C23006

Năm học: 2024

MỤC LỤC

PHÂN CÔNG VIỆC 1

Phần 1: CHUẨN BỊ MÔI TRƯỜNG 2

*1.1.* *Repository* 2

*1.2.* *Cài đặt Airflow* 2

*1.3.* *Cài đặt Power BI* 2

*1.4.* *Cài đặt PgAdmin4* 2

*1.5.* *Cài đặt VSCode* 2

Phần 2: THỐNG KÊ – DỰ ĐOÁN GIÁ VÀNG 3

*2.1.* *Giới thiệu* 3

*2.2.* *Mô tả repository* 3

*2.3.* *Lược đồ hoạt động* 4

*2.4.* *Pipelines* 5

*2.5.* *Lược đồ ER ở database* 6

*2.6.* *Model view ở PowerBI* 6

*2.7.* *Dashboard* 7

*2.8.* *Mô tả code xử lý ở Airflow* 9

*2.9.* *Mô tả code xử lý ở Database* 9

Phần 3: THAM KHẢO 10

# PHÂN CÔNG VIỆC

**TRẦN THỊ THUẬN – 23C23002**

* Thiết kế và xây dựng dashboard cho Seminar + Project
* Thiết kế lược đồ luồng dữ liệu
* Cấu hình Airflow và viết DAGs để tự động hóa luồng dữ liệu

**NGUYỄN BÍCH TRÂM – 23C23010**

* Báo cáo Seminar
* Làm file power point cho Seminar + Project
* Review + update dashboard cho Semiar + Project

**ĐỖ THỊ THANH THẢO – 23C23009**

* Crawl data
* Xử lý data lưu xuống database
* Làm file doc cho Project

**NGUYỄN TRƯỜNG GIANG – 23C23006**

* Xây dựng mô hình dự đoán giá vàng

# CHUẨN BỊ MÔI TRƯỜNG

## *Repository*

* <https://github.com/ThuanTran262/Big_data_project>

## *Cài đặt Airflow*

* <https://www.freecodecamp.org/news/install-apache-airflow-on-windows-without-docker/>

## *Cài đặt Power BI*

* <https://www.youtube.com/watch?v=JTs2EBGmTBE>

## *Cài đặt PgAdmin4*

* <https://www.youtube.com/watch?v=v1d2Fa9FPOQ>

## *Cài đặt VSCode*

* <https://www.youtube.com/watch?app=desktop&v=cu_ykIfBprI>

# THỐNG KÊ – DỰ ĐOÁN GIÁ VÀNG



## *Giới thiệu*

* Dữ liệu giá vàng lấy từ website <https://finance.yahoo.com/quote/GC=F/>. Từ đó, xây dựng Dashboard theo dõi:
  + Giá vàng lúc mở cửa, đóng cửa, giá cao nhất, giá thấp nhất, giá hiệu chỉnh, khối lượng giao dịch theo ngày, tuần, tháng, quý
  + Lịch sử giá vàng theo khoảng thời gian tùy chọn.

Và dự đoán giá vàng tăng hay giảm vào ngày mở cửa kế tiếp

## *Mô tả repository*

A screenshot of a computer

Description automatically generated

* ERD: lưu trữ lược đồ ER
* Chart folder: lưu trữ lược đồ
* Dags: lưu trữ những file dags sử dụng cho cài đặt lịch lấy dữ liệu, xây dựng mô hình, dự đoán giá vàng
* Dashboard: lưu trữ những dashboard của airflow, powerBI
* Doc folder:
  + Project folder: lưu trữ file tài liệu .doc, .ppt cho đề tài
  + Seminar folder: lưu trữ file ppt cho seminar
* Model folder: lưu trữ model đã được huấn luyện
* Notebook folder: lưu trữ notebook sử dụng cho phân tích và lựa chọn model dự đoán giá vàng
* Params folder: lưu trữ các file hằng số
* Pipeline folder: lưu trữ các pipeline files dùng để crawl data, xử lý dữ liệu, xây dựng mô hình, dự đoán giá vàng
* Utils: lưu trữ các file thư viện dùng chung

## *Lược đồ hoạt động*

A screenshot of a computer

Description automatically generated

## *Pipelines*

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## *Lược đồ ER ở database*

A screenshot of a computer

Description automatically generated

## *Model view ở PowerBI*

A screenshot of a computer

Description automatically generated

## *Dashboard*

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated

## *Mô tả code xử lý Airflow*

* Link: <https://github.com/ThuanTran262/Big_data_project/tree/main/pipeline>
  + Files: storing\_data\_pipeline.py, model\_training\_pipeline.py, model\_prediction\_pipeline.py
  + **storing\_data\_pipeline.py**: là 1 module chứa function giúp crawl data từ web và lưu xuống database
    - Đọc data từ web
    - Transform data, tách thành các bảng dim và fact
    - Lưu data xuống database
  + **model\_training\_pipeline.py**: là 1 module chứa function giúp huấn luyện mô hình
    - load data từ bảng fact\_gold\_data
    - xây dựng mô hình
    - lưu kết quả đánh giá mô hình (chỉ số RMSE) vào database
    - lưu mô hình đã được huấn luyện vào repository (folder model)
  + **model\_prediction\_pipeline.py:** là 1 module chứa function giúp dự đoán
    - load mô mình mẫu từ repository
    - load data từ bảng fact\_gold\_data
    - dự đoán dữ liệu
    - lưu kết quả dự đoán vào database
* Link: <https://github.com/ThuanTran262/Big_data_project/tree/main/dags>
  + Files: crawling\_data\_dag.py, model\_training\_dag.py, model\_prediction\_dag.py
  + **crawling\_data\_dag.py:** tạo task thêm dữ liệu vào database (tần suất daily)
  + **model\_training\_dag.py**: tạo task để xây dựng mô hình (tần suất monthly)
  + **model\_prediction\_dag.py:** tạo task để dự đoán giá vàng (tần suất daily)
* Link: <https://github.com/ThuanTran262/Big_data_project/tree/main/model>
  + Files: my\_model.keras
  + **my\_model.keras:** lưu model sau khi đã được huấn luyện
* Link: <https://github.com/ThuanTran262/Big_data_project/tree/main/utils>
  + File: crawling\_utils.py
  + **crawling\_utils.py:** lấy dữ liệu từ website finance.yahoo.com sau mỗi ngày

# THAM KHẢO

* Slide bài giảng của Thầy Nguyễn Thanh Bình
* <https://www.astronomer.io/docs/learn/vscode-local-dev/>
* <https://www.youtube.com/watch?v=IpvwxE9zSQQ>
* <https://www.youtube.com/watch?v=rvfHeAPdTL0>
* <https://github.com/microsoft/winget-cli/issues/3303>
* <https://www.youtube.com/watch?v=ES4LfzbSxMA>