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 circular blue and white circle with text

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

Phần 1: CHUẨN BỊ MÔI TRƯỜNG 1

*1.1.* *Repository* 1

*1.2.* *Cài đặt Airflow* 1

*1.3.* *Cài đặt Power BI* 1

*1.4.* *Cài đặt PgAdmin4* 1

*1.5.* *Cài đặt VSCode* 1

Phần 2: THỐNG KÊ – DỰ ĐOÁN GIÁ VÀNG 2

*2.1.* *Giới thiệu* 2

*2.2.* *Mô tả repository* 2

*2.3.* *Lược đồ hoạt động* 3

*2.4.* *Lược đồ ER ở database* 4

*2.5.* *Model view ở PowerBI* 4

*2.6.* *Dashboard* 4

*2.7.* *Code xử lý ở Airflow* 4

*2.8.* *Code xử lý ở Database* 5

*2.9.* *Code xử lý ở PowerBI* 5

Phần 3: THAM KHẢO 6

# PHÂN CÔNG VIỆC

**TRẦN THỊ THUẬN – 23C23002**

* Làm dashboard cho Seminar + Project
* Cấu hình airflow, viết DAGs files

**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**

# 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, theo loại vàng

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
* Notebook folder: lưu trữ file sử dụng cho 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

## *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*

## *Code xử lý ở Airflow*

* Link: <https://github.com/ThuanTran262/Big_data_project/tree/main/pipeline>
  + Files: model\_training\_pipeline.py, model\_prediction\_pipeline.py
  + model\_training\_pipeline.py:
    - load data từ bảng fact\_gold\_data
    - xây dựng mô hình
    - lưu mô hình vào database
    - lưu mô hình mẫu vào repository
  + model\_prediction\_pipeline.py:
    - load mô mình mẫu từ repository
    - load data từ bảng fact\_gold\_data
    - dự đoán dữ liệu
    - lưu mô hình sau khi dự đoán vào database
* Link: <https://github.com/ThuanTran262/Big_data_project/tree/main/dags>
  + Files: model\_training\_dag.py
* Link: <https://github.com/ThuanTran262/Big_data_project/tree/main/model>
  + Files: my\_model.h5
  + my\_model.h5 là file mô hình mẫu sau khi xây dựng mô hình

## *Code xử lý ở Database*

* Link: <https://github.com/ThuanTran262/Big_data_project/tree/main/pipeline>
  + Files: storing\_pipeline.py
  + Lưu data vào các bảng:
    - **dim\_date**: lưu trữ thứ tự tuần, tháng, quý năm sau khi phân tách ngày của dữ liệu thành tuần, tháng, quý, năm
    - **dim\_symbol**: lưu trữ thông tin loại vàng theo ký hiệu
    - **fact\_gold\_data**: lưu trữ dữ liệu lấy từ website finance.yahoo.com mỗi ngày
    - **real\_time\_fact\_gold\_data**: lưu trữ dữ liệu lấy từ website finance.yahoo.com sau mỗi phút
    - **real\_time\_dim\_date**: lưu trữ thứ tự tuần, tháng, quý năm của dữ liệu thực lấy được sau mỗi phút và sau khi phân tách ngày của dữ liệu thành tuần, tháng, quý, năm
* Link: <https://github.com/ThuanTran262/Big_data_project/tree/main/utils>
  + File: crawling\_utils.py
  + Lấy dữ liệu từ website finance.yahoo.com sau mỗi ngày và sau mỗi phút

## *Code xử lý ở PowerBI*

# 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>