ELT Data Warehouse

Apache Airflow

Subject: **Data Mining**

Lecturers: **PHAUK Sokkhey** & **CHAN Sophal**

Presenter: **CHOENG Veyseng**

Academic Year: 2021 - 2022

Table of Content

Introduction to Data warehouse and Data Lake

Data warehouse is management system which is specifically designed for creating environment for data analysis and manage data pipelines.

Data Lake is the complementary into the Data warehouse in order to help Data warehouse to scale the storage capacity of raw data with large amount of volume.

ELT versus ETL

Generally known as framework as the data integration methods to transfer data from one device to another from a source of data warehouse.

ETL: Extract Transform, Load

ELT: Extract, Load, Transform

Apache Airflow Introduction

In October 2014, Apache Airflow was first introduced by **Maxime Beauchemin**

In 2016, Apache Airflow is a project joined in **Apache Software Foundation Incubation program**,

It is the proposing solution that is developed by **Airbnb**

Used in more than **200 companies** world wide: Airbnb, PayPal, Intel, Stripe …

Data warehousing is a part of Apache Airflow applications that have the ability to do

Apache
Airflow

What is Apache Airflow?

Defined as a **open source** platform to programmatically author, schedule and monitor workflow as orchestrator

Authoring: written as Directed Acyclic Graph in Python Programming Language

Scheduling: able to specify when work flow should consider as start, end, after interval of the process that should run again

Monitoring: can have the interactive experience with UI tracking of the workflow

It is **scalable**, **dynamic** and attractive **user interface** help Airflow is considered to be part of the data warehousing solution

The functionalities of Airflow could be **extensible** and **customized** based on the plug-in options of the tool itself

Apache Airflow Core Components

- **Web server**: initiated with Flask server (Flask is one of Python web frameworks) using Gunicorn serving the UI
- **Scheduler**: Daemon in char of scheduling workflows
- **Metastore**: database where the meta data is stored
- **Executor**: Class defining how the tasks should be executed
- **Worker**: Process or sub-process executing the task

Apache Airflow

Apache Airflow Implementation Understanding

DAG: Directed Acyclic Graph where no loop is allowed to process and it is written in Python. It is also defined as **data pipeline** which is consist of a group of tasks that could or could not depend on one and another.

Operator: could defined as **task** there are three main types of operators: Sensor, Action and Transfer operator. For example: in order to execute a Python function, we need Python operator or execute a SQL request, we can use PostgresOperator.

Connection: configuration setting process to support DAG and operators connection with the **external components** by identifying the available for airflow to process the pipelines

Apache Airflow Demo

[1] Forex Data Pipeline

[2] Check if the forex rate data is available from the source

[3] Check the file having the currencies to watch

[4] Download forex rate with Python

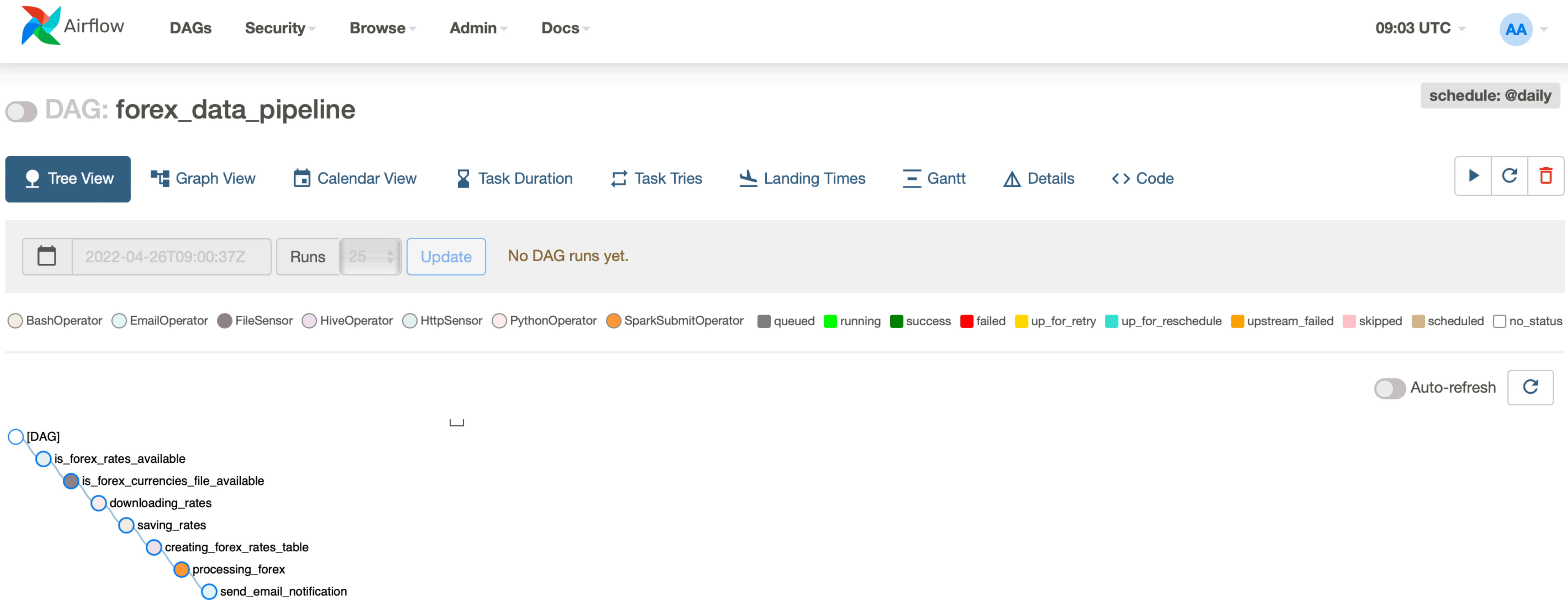
[5] Save the forex rate in HDFS (with Hadoop)

[6] Create Hive table to store HDFS

[7] Process Rate with Spark

[8] Send Email Notification

Apache Airflow Demo



DAG: forex_data_pipeline

schedule: @daily

2022-04-26T09:00:37Z Runs 25

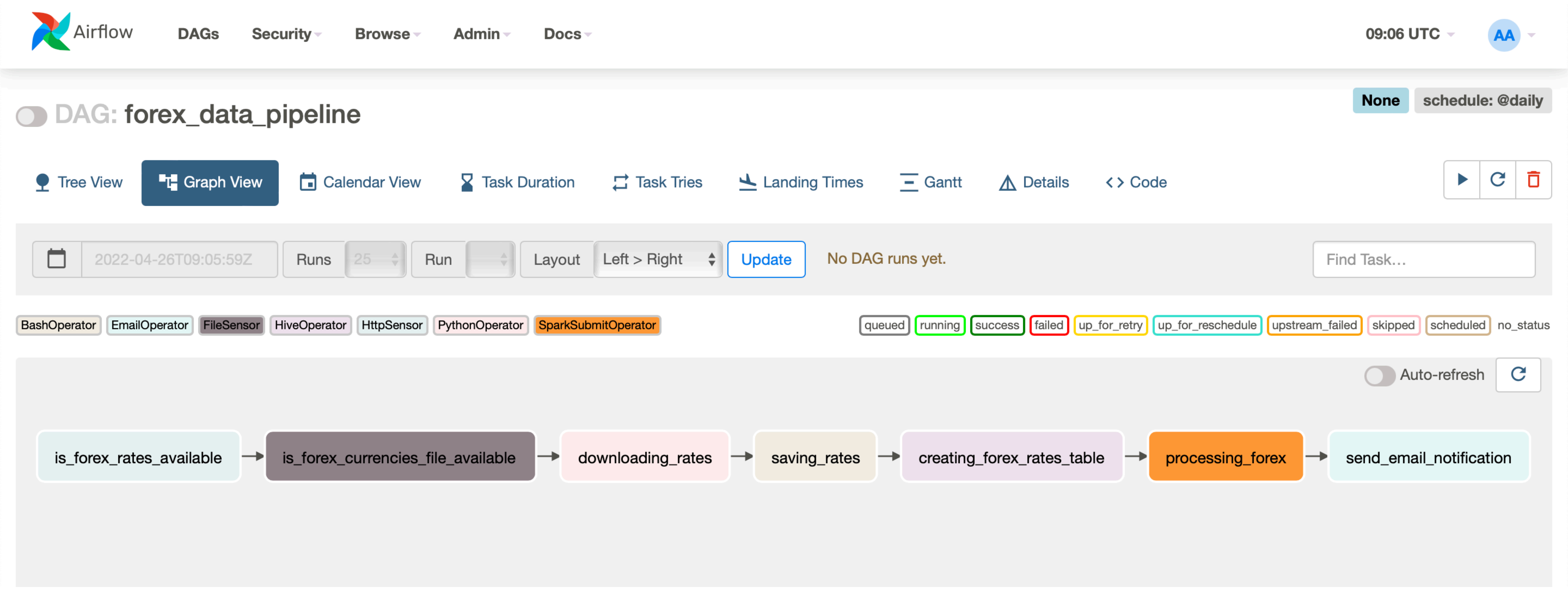
No DAG runs yet.

○ BashOperator ○ EmailOperator ● FileSensor ○ HiveOperator ○ HttpSensor ○ PythonOperator ● SparkSubmitOperator ■ queued ■ running ■ success ■ failed ■ up_for_retry ■ up_for_reschedule ■ upstream_failed ■ skipped ■ scheduled □ no_status

```
○ [DAG]
  ○ is_forex_rates_available
    ● is_forex_currencies_file_available
      ○ downloading_rates
        ○ saving_rates
          ○ creating_forex_rates_table
            ● processing_forex
              ○ send_email_notification
```

Apache Airflow

Apache Airflow Demo



DAG: forex_data_pipeline

None schedule: @daily

2022-04-26T09:05:59Z Runs 25 Run Layout Left > Right Update

No DAG runs yet.

References

[1] https://www.snowflake.com/trending/data-lake-vs-data-warehouse
[2] https://www.udemy.com/course/the-ultimate-hands-on-course-to-master-apache-airflow/
[3] https://airflow.apache.org
[4] https://docs.docker.com/
[5] https://github.com/tuanavu/airflow-tutorial
[6] https://www.applydatascience.com