final exam

Airflow

What data pipline

- Data pipeline là một khái niệm trong lĩnh vực xử lý dữ liệu (data processing), đề cập đến quá trình chuyển đổi dữ liệu từ nguồn đến đích thông qua một chuỗi các bước xử lý. + Một Data pipeline bao gồm các công đoạn như trích xuất (extraction), tiền xử lý (preprocessing), biến đổi (transformation), tích hợp (integration), lưu trữ (storage) và phân phối (delivery) dữ liệu...

What is airflow

- Airflow là một công cụ lập lịch trình cho luồng công việc cũng như hỗ trợ quản lý, theo dõi từng phần trong quy trình giúp bạn sửa lỗi, bảo trì code thuận tiện và dễ dàng. + Cung cấp giao diện dễ sử dụng để quản lý và theo dõi quy trình làm việc. Airflow cung cấp các bảng điều khiển, báo cáo và trực quan hóa để theo dõi tiến trình công việc và xem các lỗi hoặc cảnh báo + Có thể tùy chỉnh Airflow bằng cách viết mã Python để định nghĩa các công việc và luồng công việc phức tạp hơn.

Why is airflow

- Lập lịch và quản lý quy trình: Airflow cho phép bạn xác định và lập lịch chạy các task dựa trên các phụ thuộc, thời gian và điều kiện khác nhau. + Theo dõi và ghi lại quy trình: Airflow giúp bạn theo dõi và ghi lại thông tin về quy trình làm việc. + Điều phối và xử lý lỗi: Airflow có khả năng điều phối công việc giữa các máy chủ và tự động xử lý các lỗi phát sinh trong quy trình. + Mở rộng và tích hợp: Airflow cho phép bạn mở rộng và tích hợp với các công cụ và dịch vụ khác trong hệ thống của bạn.

What is DAG?

- DAG(Directed Acyclic Graph) là một thuật ngữ để chỉ một biểu đồ hướng và không có chu trình + Trong Airflow, DAG được sử dụng để biểu diễn một luồng task hoặc quy trình làm việc, theo một chiều nhất định

DAG runs

- DAG Runs là các phiên bản cụ thể của DAG được thực thi. + Bao gồm một bản sao của DAG và các thông tin liên quan đến quá trình thực thi của nó. Nó bao gồm các thông tin như thời gian bắt đầu và kết thúc, trạng thái của DAG Run (running, success, failed, skipped, v.v.), + Cho phép theo dõi và quản lý quá trình thực thi của DAG + Airflow giữ lại các DAG Runs trong cơ sở dữ liệu metadata

TASK

- Task là một thành phần cơ bản của một DAG + Task đại diện cho một công việc cụ thể cần thực hiện trong quy trình làm việc + Thuật ngữ upstream và downstream được sử dụng để chỉ các mối quan hệ phụ thuộc giữa các task

TASK INSTANCE

- Task Instance là các phiên bản cụ thể của một task trong một DAG Run. Mỗi khi một task trong DAG được thực thi, một Task Instance mới được tạo ra. + Task Instance bao gồm thông tin về trạng thái của task, gian bắt đầu và kết thúc, trạng thái tiến trình, ... + Cho phép theo dõi và quản lý từng bước trong quá trình thực thi của một task + Airflow lưu trữ thông tin về các Task Instances trong cơ sở dữ liệu metadata + Các trạng thái của task cho 1 flow lý tưởng: none→scheduled→queued→running→finally→success

TASKFLOW API

- TaskFlow API là một API cao cấp hơn Task Instance + Cung cấp một cách tiếp cận mới để định nghĩa và quản lý các quy trình công việc phức tạp hơn. + TaskFlow API cho phép bạn xác định các luồng công việc và quyết định điều kiện để thực hiện các task trong Airflow.

Why

TaskFlow API trong Apache Airflow khắc phục một số nhược điểm của Task Instance: + Đơn giản hóa việc định nghĩa quy trình công việc phức tạp + Xử lý lỗi linh hoạt và thực thi tái thử + Quản lý luồng điều kiện phức tạp + Dễ dàng kiểm soát và tái sử dụng

What is operator?

- Operator là một lớp hoặc đối tượng Python đại diện cho một công việc cụ thể trong một DAG + Đóng vai trò thực hiện các công việc cụ thể, như thực thi chương trình, truy vấn cơ sở dữ liệu, gửi email, tải dữ liệu và nhiều tác vụ khác. + Giúp xác định các bước và hành động cụ thể mà một task phải thực hiện

Popular operatorsk

- BashOperator: chạy các command line + PythonOperator: chạy các python function + Python Virtualenv Operator: chạy các python function nhưng sử dụng virtualenv + Ngoài ra còn có nhiều Operator khác như: Email Operator, Dummy Operator, SSH Operator nhưng chỉ tập trung 3 nhóm trên.

WORKFLOW

- Workflow là một tập hợp các task được tổ chức và quản lý để thực hiện một quy trình tự động. + Bao gồm định nghĩa và xác định thứ tự của các task, các phụ thuộc giữa chúng và lịch chạy để đảm bảo các task được thực hiện theo đúng thứ tự và thời gian. + Giúp tổ chức, lập lịch và thực thi các quy trình làm việc tự động phức tạp. + Workflow được biểu diễn dưới dạng DAG(dạng biểu đồ hướng và không có chu trình) và chứa các Task riêng lẻ.

AIRFLOW INSTALLATION

Một Airflow installation thông thường bao gồm các thành phần sau: ● Scheduler ● Executor ● Web server ● Dags folder ● Metadata database

WHAT IS CATCHUP, BACKFILL?

- Catchup, backfill là thuật ngữ được sử dụng để liên quan đến việc xử lý và thực thi các DAGs trong quá khứ. + Catchup là quá trình chạy các task trong DAG từ thời điểm bắt đầu của DAG đến thời điểm hiện tại. + Backfill là quá trình chạy lại các task của một DAG trong quá khứ từ một ngày bắt đầu cụ thể.

Cron expression ?

- Cron Expression là một phần của cú pháp được sử dụng để định lịch thực hiện các DAGs hoặc các task trong DAGs dựa trên một lịch trình thời gian định kỳ. + Sử dụng cú pháp đặc biệt để xác định các mốc thời gian cụ thể mà các DAGs hoặc task sẽ được thực hiện.

Advans

- Tự động hoá và quản lý công việc phức tạp + Nhiều tích hợp và khả năng mở rộng. + Giao diện người dùng trực quan, dễ theo dõi + Mã nguồn mở, cộng đồng lớn.

Disadvantages

- Phức tạp trong việc cấu hình ban đầu. + Không phù hợp cho các công việc cần tính real+time. + Cần phải bảo trì thường xuyên và tốn dung lượng để lưu trữ log metadata. + Tính bảo mật không cao.

Spark

What

Apache Spark là một công nghệ tính toán một cách thống nhất và một bộ thư viện cho xử lý dữ liệu song song trên các cụm máy tính.
2022 MapReduce @ Google -> 2004 MapReduce paper -> 2006 Hadoop @ Yahoo -> 2008 Hadoop Submit -> 2010 Spark paper -> 2014 Apache Spark top+level

Tính năng

- Tốc độ + Tính toán thời gian thực + Tích hợp Hadoop + Đa ngôn ngữ + Lazy evaluation + Học máy + Nhiều định dạng

Why and purpose

Data scientists

- Thống kê (statistics, machine learning, SQL) + Chuyển đổi dữ liệu thành định dạng có thể sử dụng được + Phân tích mô hình hóa dữ liệu
  Data engineer
- Phát triển hệ thống hoặc ứng dụng xử lý dữ liệu + Kiểm tra và điều chỉnh các ứng dụng + Lập trình với spark's API

Spark core

Bộ máy thực thi phân tán: đây là bộ máy cơ sở cho xử lý dữ liệu phân tán quy mô lớn trên nhiều máy tính song song. • Hỗ trợ các API Java, Scala và Python • Các thư viện bổ sung được xây dựng trên đó cho phép xử lý các tải công việc đa dạng. • Quản lý bộ nhớ và khôi phục lỗi, lên lịch, phân phối và giám sát công việc trên một cụm, và tương tác với các hệ thống lưu trữ. Spark tập trung vào thực hiện tính toán trên dữ liệu, bất kể nó được lưu trữ ở đâu.

Spark streaming

Cho phép xử lý luồng dữ liệu trực tiếp với khả năng xử lý dữ liệu lớn và khả năng chịu lỗi.

Spark SQL

- Tích hợp xử lý quan hệ với lập trình chức năng • Hỗ trợ truy vấn dữ liệu qua SQL hoặc Hive Query Language

SPARK GRAPHX

• Bao gồm một bộ sưu tập ngày càng tăng các thuật toán và trình tạo biểu đồ để đơn giản hóa các tác vụ phân tích biểu đồ. • Hỗ trợ một biến thể được tối ưu hóa của Pregel API.

SPARK MLLIB

MLlib là một thư viện gồm các thuật toán hỗ trợ Machine Learning với quy mô Big data. Thư viện này có hiệu suất hoạt động nhanh và khả năng thực hiện nhiều công việc hơn.

MÔ HÌNH THỰC THI SPARK

Một ứng dụng Spark có một tiến trình driver duy nhất và một tập các tiến trình executor được phân tán trên các máy chủ trong một cụm.