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

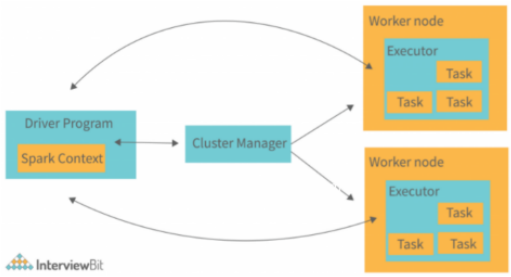
• Bao gồm một bộ sưu tập ngày càng tăng các thuật toán và trình tạo biểu đồ để đơn giản hóa các tác vụ phân tích biểu đồ. • Hỗ trợ một biến thể được tối ưu hóa của Pregel API.

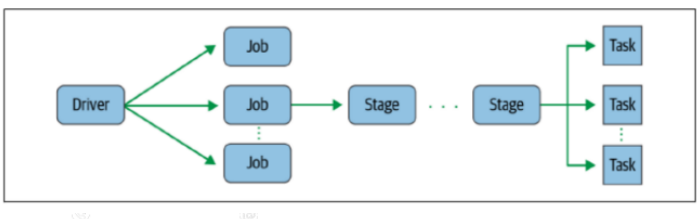
SPARK MLLIB

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.





DRIVER

Bảo trì tất cả thông tin trong suốt vòng đời của một ứng dụng đang chạy trên cụm. •Theo dõi tất cả trạng thái và tác vụ của các executors. • Giao tiếp với quản lý cụm để lấy tài nguyên vật lý và khởi chạy các executor. •Hiển thị như một quy trình trên một máy vật lý.

NGƯỜI THỰC THI

Thực hiện công việc được giao bởi driver

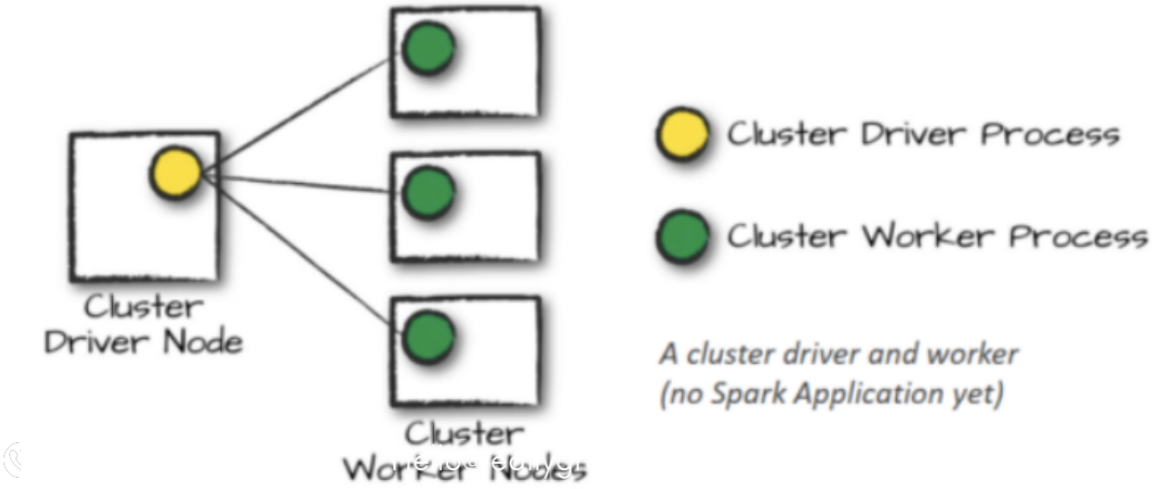
- Thực thi mã được giao cho nó bởi driver • Báo cáo trạng thái tính toán trở lại nút driver
  Sử dụng nhiều luồng để thực thi một số tác vụ

• Mã ứng dụng cần phải được bảo đảm an toàn đối với luồng.
Một executor KHÔNG chạy các tác vụ từ nhiều ứng dụng.

DAG CHO MỘT CHƯƠNG TRÌNH SPARK

Hadoop MapReduce
• DAG là một loạt các tác vụ map và reduce được sử dụng để thực hiện ứng dụng • Một nhà phát triển cần xác định từng tác vụ và nối chúng lại với nhau.
Apache Spark
• Chính engine tạo ra các chuỗi bước phức tạp đó từ logic của ứng dụng • Framework tối ưu hóa công việc, dẫn đến hiệu suất cải thiện.

CLUSTER MANAGER



CHẾ ĐỘ THỰC THI

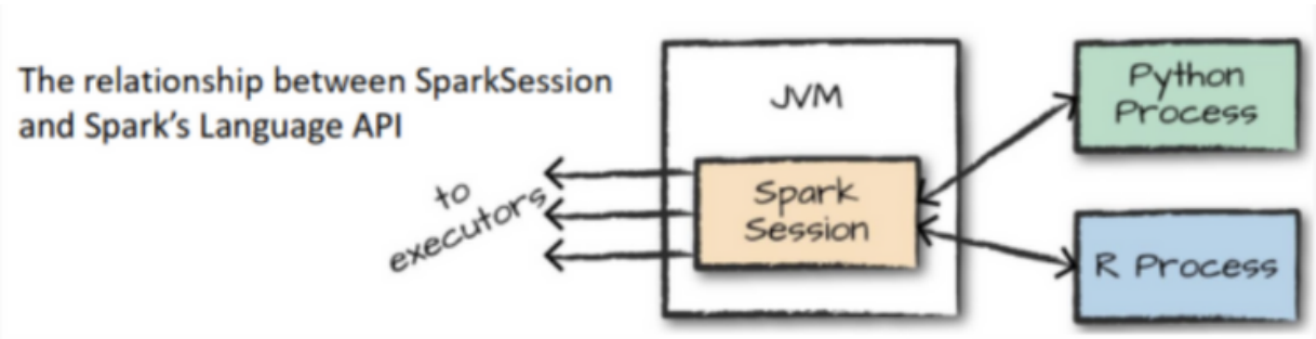
- Cluster mode + Client mode + Local mode

VÒNG ĐỜI CỦA MỘT ỨNG DỤNG SPARK

- Client Request + Launch + Execution + Completion

SPARKSESSION

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.master("local").appName("Word Count")\
        .config("spark.some.config.option", "some-value")\
        .getOrCreate()
```

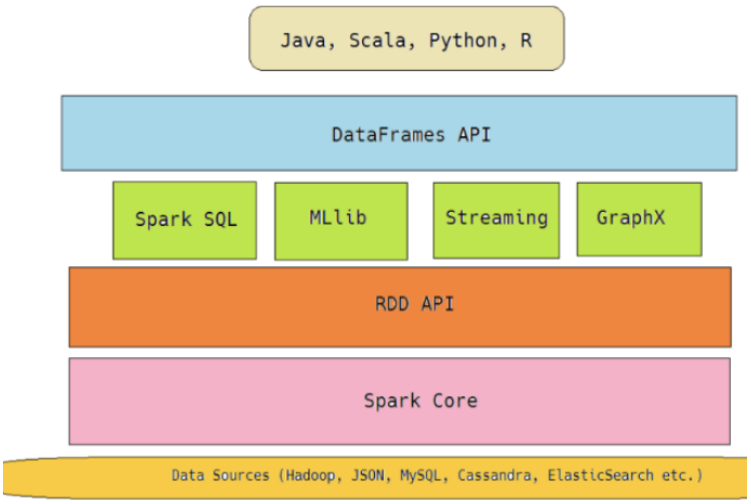


Một đối tượng trong SparkSession kết nối tới cụm
• Tạo RDDs, accumulators, và biến broadcast, chạy mã trên cụm, v.v.
Hoặc, nó nên được tạo ra theo cách tổng quát nhất như sau:

```
from pyspark import SparkContext
sc = SparkContext.getOrCreate()
```

LOW-LEVEL APIS

Low-level APIs (Các API cấp thấp) là các giao diện lập trình ứng dụng cung cấp một mức độ kiểm soát chi tiết hơn đối với hệ thống hoặc công nghệ cơ bản.
Khi nào nên dùng Low-level APIs • Cần một số chức năng không có sẵn trong các API cấp cao hơn • Ví dụ, kiểm soát chặt chẽ vị trí dữ liệu vật lý trên cụm • Giữ mã nguồn cũ được viết bằng RDDs • Thực hiện một số thao tác tùy chỉnh trên biến chia sẻ.
Tại sao cần hiểu và làm việc với các API cấp thấp? • Tất cả các tải trọng làm việc với Spark đều được biên dịch thành các nguyên tố cơ bản. • Một phép biến đổi DataFrame được chuyển đổi thành một tập hợp các biến đổi RDD.



RDD LÀ GÌ ?

RDD là viết tắt của Resilient Distributed Dataset, là một khái niệm quan trọng trong Apache Spark. RDD là một tập hợp bất biến, được phân tán trên nhiều node trong một cluster và được sử dụng để lưu trữ và xử lý các dữ liệu lớn trong Spark.

RDDS PROPERTIES

Những tính chất này xác định tất cả khả năng lập lịch và thực thi chương trình của Spark

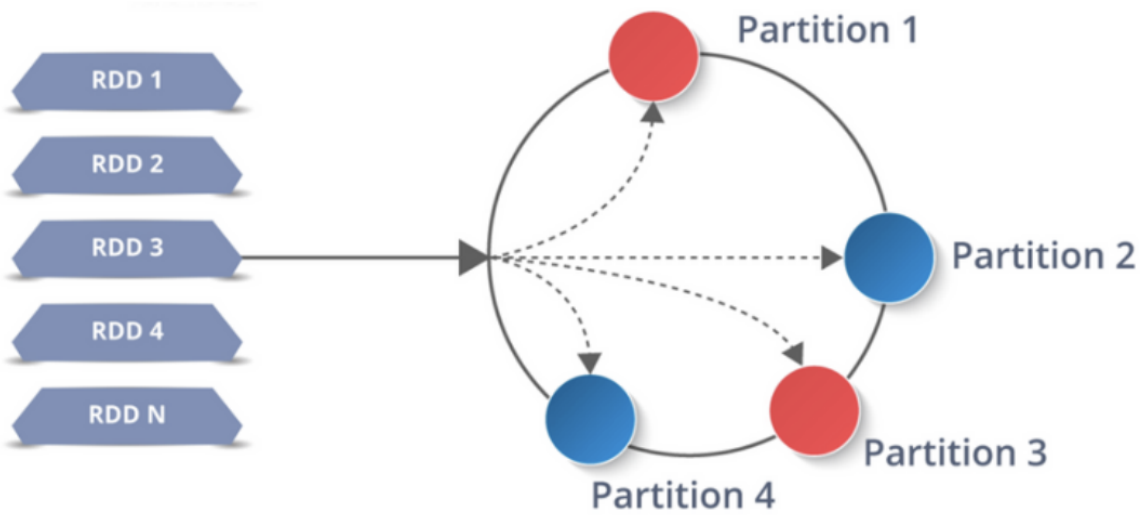
Required	• A list of partitions
	• A function for computing each split
	• A list of dependencies on other **RDDs**
Optional	• A Partitioner for key-value **RDDs**
	• A list of preferred locations on which to compute each split (e.g., block locations of a HDFS file)

TYPES OF RDDS

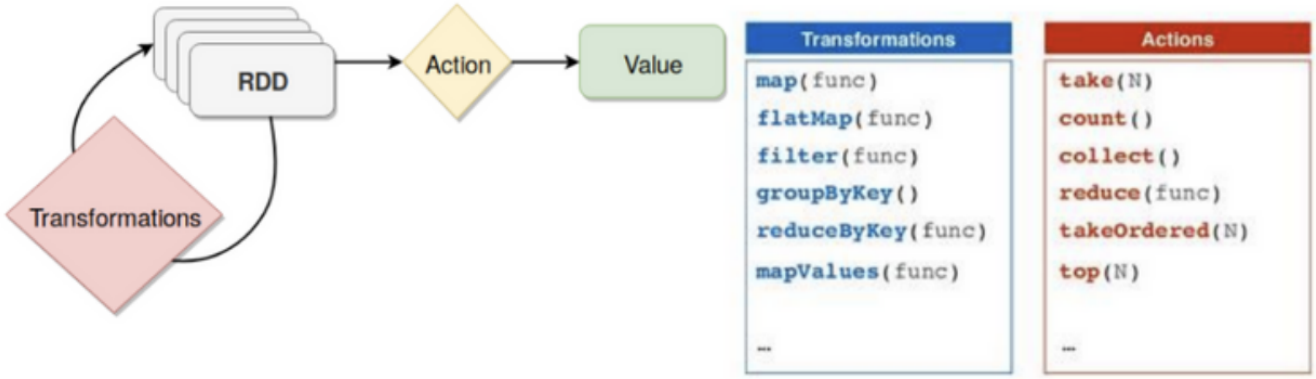
- Generic RDD + Key - value RDD
- RDD kiểu key-value có các thao tác đặc biệt và một khái niệm về phân vùng tùy chỉnh theo key.

LOGICAL PARTITIONS

Mỗi tập dữ liệu trong RDD được chia thành các phân vùng logic, có thể tính toán trên các nút khác nhau của cụm.



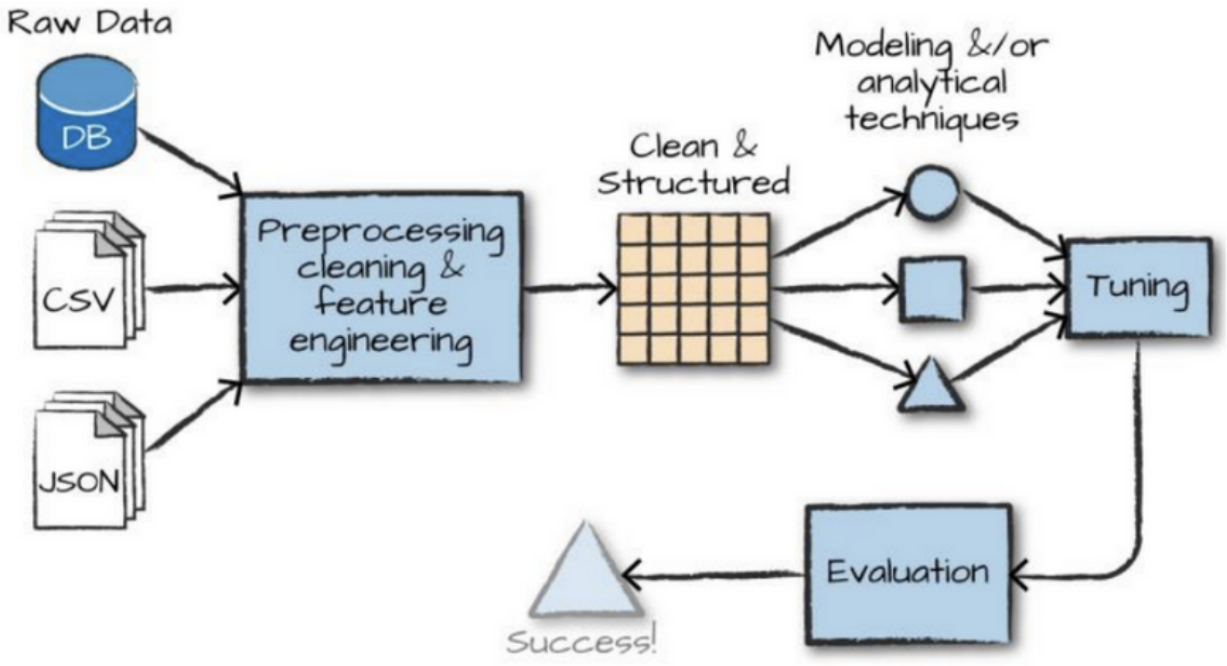
TRANSFORMATIONS AND ACTIONS



• Spark ghi nhớ các phép biến đổi được áp dụng cho một tập dữ liệu cơ bản và tính toán chúng khi một hoạt động yêu cầu trả về một kết quả.

XỬ LÝ DỮ LIỆU

Advanced analytics process



QUÁ TRÌNH PHÂN TÍCH NÂNG CAO

- **Thu thập dữ liệu:** thu thập dữ liệu cần thiết để huấn luyện một thuật toán • Spark có thể xử lý nhiều nguồn dữ liệu và kích thước dữ liệu khác nhau
- **Xử lý dữ liệu:** giải quyết các sự mập mờ và không nhất quán trong dữ liệu • Điền các mục bị thiếu, sửa các giá trị không hợp lệ và giải quyết các xung đột • MLlib cung cấp một loạt các API cho xử lý trước dữ liệu.
- **Kỹ thuật xử lý đặc trưng:** chuyển đổi dữ liệu thành một dạng phù hợp cho các thuật toán ML • Thêm / xóa các thuộc tính, chuẩn hóa / phân hóa giá trị của một thuộc tính, xử lý các biến phân loại, v.v. • Các biến trong MLlib thường là vector của các số thực.

QUÁ TRÌNH PHÂN TÍCH

Huấn luyện mô hình: xây dựng một mô hình để dự đoán đầu ra chính xác, cho một số đầu vào • Các mô hình có thể được sử dụng để thu thập thông tin hoặc dự đoán tương lai
Điều chỉnh và đánh giá mô hình: thử nghiệm các siêu tham số khác nhau và so sánh các biến thể mô hình mà không bị quá khớp • Tập huấn luyện so với Tập xác thực so với Tập kiểm tra
Tận dụng mô hình và/hoặc thông tin • Chúng ta kết thúc với một mô hình hoạt động tốt sau khi chạy mô hình qua quá trình huấn luyện, cũng như điều chỉnh và đánh giá mô hình.

PHÂN TÍCH DỮ LIỆU VỚI SPARK

MLlib cung cấp các giao diện để xử lý trước dữ liệu, huấn luyện và điều chỉnh các mô hình quy mô lớn và sử dụng chúng trong sản xuất • org.apache.spark.ml cho việc thao tác với DataFrame • org.apache.spark.mllib cho các API RDD cấp thấp của Spark
MLlib chủ yếu được thiết kế cho các vấn đề về khả năng mở rộng. • Các công cụ trên một máy tính đơn lẻ thường là bổ sung cho MLlib, ví dụ như scikit-learn, TensorFlow, R, v.v.

ƯU ĐIỂM CỦA MLLIB

• Xử lý dữ liệu lớn • Đa dạng thuật toán • Tích hợp với Apache Spark • API dễ sử dụng

NHƯỢC ĐIỂM CỦA MLLIB

• Hạn chế về deep learning • Hạn chế về tính linh hoạt • Không hỗ trợ tính toán GPU • Có thể cần phải tùy chỉnh thêm

ƯU ĐIỂM SPARK

- Dễ sử dụng + Được hỗ trợ bởi nhiều thư viện cấp cao + Xử lý dữ liệu lớn nhanh hơn đáng kể so với các công nghệ xử lý dữ liệu truyền thống như MapReduce. + Có khả năng tích hợp với hệ thống Hadoop và các công nghệ lưu trữ dữ liệu khác
- Hỗ trợ nhiều ngôn ngữ lập trình như Java, Scala, Python và R + Cung cấp nhiều API và thư viện hỗ trợ các tác vụ xử lý dữ liệu phức tạp + Lazy evaluation giúp tối ưu hiệu suất tính toán và tối thiểu hoá việc truyền dữ liệu trên mạng.

NHƯỢC ĐIỂM

- yêu cầu nhiều tài nguyên hơn so với các công nghệ xử lý dữ liệu truyền thống + Khó khăn trong việc xử lý dữ liệu không đồng nhất hoặc dữ liệu có cấu trúc phức tạp. + Có thể gây ra sự cố trong quá trình triển khai và vận hành trên các cụm máy tính lớn.

SO SÁNH SPARK VÀ HADOOP

- Hiệu suất + Khả năng tích hợp (HDFS, Cassandra và HBase) + Ngôn ngữ lập trình + Tính linh hoạt

---

Power BI

What

Power BI là một công cụ Business Intelligence (BI) cho phép người dùng kết nối và tổ chức dữ liệu từ nhiều nguồn khác nhau, biến chúng thành báo cáo và đồ thị trực quan, dễ hiểu để tiến hành phân tích và đưa ra quyết định.
Các sản phẩm gồm: + Power BI Desktop (Free) + + Power BI Pro + + Power BI Premium + + Power BI Mobile + + Power BI Embedded + + Power BI Report Server

Các tính năng của Power BI

- **Kết nối và tích hợp dữ liệu**: Power BI cho phép kết nối và tích hợp dữ liệu từ nhiều nguồn bao gồm cả cơ sở dữ liệu định hướng và không định hướng, dữ liệu trong excel, dữ liệu trong máy tính, trên điện toán đám mây và nhiều nguồn khác.
- **Biểu đồ và đồ thị:** Power BI cho phép người dùng tạo ra các biểu đồ và đồ thị trực quan để phân tích dữ liệu, bao gồm các biểu đồ cột, biểu đồ vòng, biểu đồ thác nước và nhiều loại biểu đồ khác.
- **Chia sẻ và xuất báo cáo**: Power BI cho phép người dùng chia sẻ và xuất báo cáo một cách nhanh chóng và dễ dàng, bao gồm chia sẻ báo cáo trong tổ chức, chia sẻ báo cáo cho khách hàng và xuất báo cáo sang nhiều định dạng khác như PDF, Excel, PowerPoint, …
- **Tích hợp với các ứng dụng khác**: Power BI tích hợp với nhiều dịch vụ Microsoft khác nhau như Microsoft Excel, SharePoint, Teams và Dynamics 365. Điều này giúp người dùng có thể chia sẻ dữ liệu và báo cáo của mình với các dịch vụ khác trong tổ chức một cách dễ dàng và nhanh chóng.

Ưu điểm

- Đa nền tảng + + Dễ sử dụng + + Tích hợp dữ liệu từ nhiều nguồn + + Khả năng chia sẻ + + Phân tích dữ liệu nâng cao + + Tính linh hoạt + + Trực quan hóa với khả năng tương tác + + Tính bảo mật cao

Khuyết điểm

- Hạn chế về khả năng xử lý với dữ liệu lớn + + Giao diện người dùng chật chội + + Xử lý mối quan hệ bảng chưa tốt + + Khó để hiểu và làm chủ các công cụ phức tạp của Power BI

Ứng dụng của power BI

- Phân tích kinh doanh + Theo dõi hiệu suất + Phân tích khách hàng + Quản lý dự án + Ngoài ra power BI còn có một số ứng dụng khác như: - Tổ chức và chia sẻ thông tin: Power BI có thể được sử dụng để tổ chức và chia sẻ thông tin với đồng nghiệp hoặc khách hàng của bạn. - Theo dõi các chỉ số chính: Power BI cung cấp các bảng điều khiển và báo cáo để theo dõi các chỉ số chính của doanh nghiệp, từ đó giúp bạn tăng cường quản lý và đưa ra quyết định.

So sánh

Power BI vs Tableau

Phương diện so sánh	Power BI	Tableau
Truy cập dữ liệu	Không thể kết nối với cơ sở dữ liệu Hadoop(cho phép trích xuất dữ liệu từ phân tích Azure, Salesforce và Googles) Người dùng bị hạn chế quyền truy cập vào các cơ sở dữ liệu và máy chủ khác	Cho phép người dùng truy cập dữ liệu trên đám mây và kết nối với cơ sở dữ liệu Hadoop cung cấp cho người dùng quyền truy cập không giới hạn.
Hình ảnh trực quan	Cung cấp nhiều điểm dữ liệu để tạo ra hình ảnh trực quan.	Giao diện người dùng, dễ sử dụng, không cần sử dụng bất kỳ ngôn ngữ mã hóa nào
Thiết lập	Cloud Desktop Windows Mobile Android	Tableau Destop, Tableau Online hoặc Tableau Server.
Triển khai	Triển khai theo mô hình SaaS	Triển khai Cloud & On-Premise.
Machine Learning	Được tích hợp với Microsoft Azure	Khả năng học máy của Python được tích hợp sẵn trong Tableau
Hiệu suất	Có thể xử lý một khối lượng dữ liệu giới hạn	Có thể xử lý một khối lượng dữ liệu khổng lồ với hiệu suất tốt hơn hẳn.
Giá	Power BI Desktop được hỗ trợ miễn phí. Các gói dịch vụ rẻ hơn Tableau.	Giá thành cho các sản phẩm và dịch vụ của Tableau cao hơn Power BI.

Power BI vs QlikView

Phương diện so sánh	Power BI	QlikView
Truy cập dữ liệu	Kết nối với tất cả các công nghệ của Microsoft và các nền tảng bên ngoài như MySQL, Google Analytics, Oracle, Facebook,	Kết nối với các dịch vụ đám mây như Azure, Twitter, MS office, Hadoop, SAP,....
Phân tích và báo cáo	Có nhiều tính năng để xây dựng báo cáo theo đa dạng yêu cầu và khả năng mở rộng báo cáo dễ dàng như: truy vấn (drill-down), lọc chéo (cross-filtering), lưu trữ chế độ nhìn dữ liệu (bookmark), ...	Yếu hơn về các tính năng phân tích và báo cáo so với Power BI.
Người dùng	Nhiều đối tượng (dễ sử dụng)	Được sử dụng hầu hết bởi những người có kiến thức chuyên môn về lập trình dữ liệu vì vậy đại đa số thành viên trong cộng đồng của họ là những người làm chuyên môn và viết lập trình
Thiết lập	Cloud Desktop Windows Mobile Android	Cloud Desktop Windows,macOS On-premise Windows and Linux Mobile - Android and iPhone
Tốc độ	Có thể hoạt động với tốc độ tốt mà không gặp nhiều khó khăn.	Tốc độ QlikView phụ thuộc vào tốc độ RAM của thiết bị.
Giá	Rẻ	Giá thành cao hơn Power BI.

Tableau

What

- Tableau is a software tool for business intelligence and data visualization created by Tableau Software. + Allows users to connect to various data sources, create interactive and dynamic visualizations and dashboards, and share insights with others. + It is known for its flexibility, scalability, and ease of use, making it a popular choice for businesses of all sizes.

Why

Easy to use User-friendly interface and drag-and-drop functionality, making it easy for users to create visualizations and dashboards.
Advanced analytics features Provides advanced features for data preparation, data analysis, and predictive analytics
Wide range of data sources Supports a wide range of data sources which makes it easy to connect and integrate data from different sources.
Real-time collaboration Supports real-time collaboration and allows users to share their visualizations and insights with others.
Powerful visualization capabilities Provides a wide range of chart types and customization options.
Scalability Tableau is designed to handle large and complex data sets, making it a scalable solution for businesses of all sizes.

SOME APPLICATIONS OF TABLEAU

Business intelligence Create interactive dashboards and reports that provide real-time insights into key performance indicators (KPIs) and other business metrics
Finance Analyze financial data, track revenue and expenses, and create financial reports and forecasts.
Marketing Analyze customer behavior, track marketing campaigns, and create visualizations to communicate marketing insights.
Healthcare Analyze patient data, track healthcare outcomes, and create visualizations to communicate healthcare insights.
Education Analyze student data, track student performance, and create visualizations to communicate educational insights.
Government Analyze public data, track government performance, and create visualizations to communicate government insights.

SOME PRODUCT OFFERINGS

Tableau Desktop This is the primary authoring and publishing tool for creating interactive data visualizations, dashboards, and reports.
Tableau Server This is a web-based platform that allows users to publish and share interactive dashboards, reports, and visualizations.
Tableau Online This is a cloud-based version of Tableau Server that provides the same functionality as Tableau Server but hosted in the cloud.
Tableau Mobile This is a mobile app that allows users to access and interact with Tableau dashboards and reports on their mobile devices.
Tableau Public This is a free version of Tableau that allows users to create and share public data visualizations on the web.
Tableau Reader This is a free desktop application that allows users to view and interact with Tableau workbooks created by others.

DATA TYPES IN TABLEAU

- Text (string) + Numerical + + Date + Boolean + + Geographic + + Cluster + Date & Time

TABLEAU TURTORIAL

TOTAL CASES BY COUNTRIES

- Columns: SUM(Total cases) + Rows: Country + Marks: Color: SUM(Total cases), Text: SUM(Total cases)

TOTAL DEATHS BY COUNTRIES

- Columns: SUM(Total deaths) + Rows: Country + Marks: Color: SUM(Total cases), Text: SUM(Total cases)

DEATH RATE BY COUNTRIES

- Columns: AGG(Death rate) + Rows: Country + Marks: Color: AGG(Death rate), Text: AGG(Death rate)

CASE MOVEMENT

- Columns: DAY(Date reported) + Rows: SUM(Total cases), SUM(New cases) + Marks: Color: Measure Names

DEATH MOVEMENT

- Columns: DAY(Date reported) + Rows: SUM(Total deaths), SUM(New deaths) + Marks: Color: Measure Names

SUMMARY

- Columns: Measure Names + Measure Values: SUM(New cases), SUM(New deaths), SUM(Total cases), SUM(Total deaths) + Marks: Text: Measure Names, Text: Measure Values + Filter: YEAR(Date reported), Country

MAP

- Columns: Longitude (generated) + Rows: Latitude (generated) + Marks: Size: SUM(Total cases), Color: AGG(Death rate), Detail: Country

GENERAL REVIEW

Tableau is a visual data analysis tool that helps users easily, quickly explore and understand their data. It can connect to various data sources, from Excel spreadsheets to large databases, allowing users to create visually appealing reports, charts, and dashboards. Tableau also supports sharing and collaboration with stakeholders through Tableau Server or Tableau Online. Tableau is a suitable solution for businesses, organizations, and individuals looking to maximize the value of their data.

ADVANTAGES

Easy to learn and use Tableau has a user-friendly interface, allowing users to easily create charts by dragging and dropping data fields.
Flexible and diverse Tableau has many chart types such as column, line, circle, map, tree, and grid. Combine multiple chart types to create dynamic and intuitive dashboards.
Powerful and efficient Tableau processes large and complex data quickly and accurately. Tableau shares reports and dashboards with many other users

DISADVANTAGES

High cost Tableau has a high price so with other tools it is necessary to pay for Tableau Desktop or Tableau Server. This can limit accessibility for individuals or organizations on a tight budget.
Limitations of charts Tableau has no radar, sankey, or waterfall charts. Users may need to seek alternative solutions or utilize open-source codes to generate these specific chart types.
Hard to integrate with other tools Tableau is difficult to integrate with other tools. It may have feature limitations and scalability when compared to other tools

---

Data visualization tools

D3.js

WHAT

- D3.js là một thư viện JavaScript được sử dụng để tạo trực quan hóa tương tác trong trình duyệt. + D3.js có thể trực quan hóa dữ liệu, tương tác biểu đồ dưới dạng Scalable Vector Graphics (SVG), HTML5, and Cascading Style Sheets (CSS). + D3.js được biết đến rộng rãi từ năm 2011 khi phiên bản 2.0.0 của nó được phát hành vào tháng 8 – 2011.

Tính năng nổi bật của D3.js

- Rất linh hoạt + + Dễ dàng sử dụng + + Hỗ trợ tập dữ liệu lớn + + Lập trình bằng phương thức khai báo + + Tính tái sử dụng + + Có đa dạng các hàm tạo đường cong + Kết hợp dữ liệu với một phần tử hoặc nhóm các phần tử trên trang HTML. + Thao tác với DOM dễ dàng

Ưu điểm

- D3.js là một thư viện JavaScript. Vì vậy, nó có thể được sử dụng với bất kỳ Framework JavaScript nào mà bạn chọn như Angular, React hoặc Vue.js. + D3 tập trung vào dữ liệu, vì vậy nó là công cụ thích hợp và chuyên biệt nhất để trực quan hóa dữ liệu. + D3 là mã nguồn mở. Vì vậy, bạn có thể làm việc với mã nguồn và thêm các tính năng của riêng bạn. + Nó hoạt động với các tiêu chuẩn web như HTML, CSS và SVG nên bạn không cần bất kỳ công nghệ hoặc plugin nào khác ngoài trình duyệt để sử dụng D3. + D3 không cung cấp bất kỳ tính năng cụ thể nào, vì vậy, nó cho phép bạn kiểm soát hoàn toàn hình ảnh của mình để tùy chỉnh nó theo cách bạn muốn. Điều này mang lại cho nó một lợi thế so với các công cụ phổ biến khác như Tableau hoặc QlikView. + Vì D3 có kích thước khá nhẹ và hoạt động trực tiếp với các tiêu chuẩn web nên nó cực kỳ nhanh và hoạt động tốt với các bộ dữ liệu lớn

Nhược điểm

- Cần rất nhiều thời gian để hiểu rõ và thông thạo được D3.js + Các trình duyệt cũ không được hỗ trợ nhiều + D3.js có thể chậm hơn các thư viện khác khi hiển thị các bộ dữ liệu lớn, đặc biệt là khi sử dụng trực quan hóa phức tạp. + D3.js không cung cấp các loại biểu đồ được xây dựng sẵn, vì vậy người dùng cần xây dựng biểu đồ từ đầu.

Highcharts

- Highcharts là một thư viện JavaScript mã nguồn mở được sử dụng để tạo các biểu đồ tương tác và trực quan. + Được phát triển bởi công ty Highsoft từ năm 2009, Highcharts đã trở thành một lựa chọn hàng đầu cho các nhà phát triển và nhà thiết kế trong việc hiển thị dữ liệu phức tạp một cách trực quan và dễ hiểu.

Tổng quan về Highcharts

- Đa dạng các loại biểu đồ + Có thể tương tác với người dùng + Tương thích cao với JS và CSS +

Ưu điểm

- Dễ sử dụng: Highcharts cung cấp cú pháp rõ ràng và dễ hiểu. + Đa dạng và linh hoạt: Hỗ trợ nhiều loại đồ thị cũng như cho phép tùy chỉnh các yếu tố của đồ thị dễ dàng. + Tương tác: Cho phép người dùng tương tác với các đồ thị thông qua các sự kiện như di chuột, nhấp chuột và chạm. + Tích hợp dễ dàng: Có thể tích hợp dễ dàng vào các ứng dụng web hiện có. + Hỗ trợ đa nền tảng: Tương thích với hầu hết các trình duyệt web phổ biến. + Tính di động: Highcharts hỗ trợ tốt cho các thiết bị di động. + Cộng đồng phát triển mạnh mẽ: Highcharts có một cộng đồng phát triển lớn và nhiều tài liệu hướng dẫn. + Hỗ trợ kỹ thuật: Có nhiều kênh liên lạc và hỗ trợ kỹ thuật như tài liệu, diễn đàn và trang web hỗ trợ trực tuyến.

Nhược điểm

- Giới hạn miễn phí: Để sử dụng Highcharts trong các dự án thương mại hoặc dự án có yêu cầu đặc biệt, bạn cần phải mua giấy phép. + Tính tùy chỉnh hạn chế: Trong một số trường hợp, Highcharts có thể hạn chế tính tùy chỉnh cao đối với một số yêu cầu đặc biệt của người dùng. + Yêu cầu kiến thức kỹ thuật: Để sử dụng Highcharts hiệu quả, người dùng cần có kiến thức về HTML, CSS và JavaScript. + Hiệu suất: Trong một số trường hợp, Highcharts có thể trở nên chậm hoặc gặp khó khăn khi xử lý và hiển thị các tập dữ liệu lớn. + Phụ thuộc vào JavaScript: Nếu người dùng không hỗ

trợ hoặc vô hiệu hóa JavaScript trên trình duyệt, đồ thị sẽ không hoạt động. + Khả năng tương thích trình duyệt: Highcharts vẫn có thể gặp một số vấn đề tương thích đối với một số phiên bản trình duyệt cũ hoặc trình duyệt không phổ biến.

echarts

WHAT

- Ban đầu được tạo tạo Baidu, sau đó được quản lý bởi Apache Software Foundation + Là thư viện mã nguồn mở hỗ trợ việc trực quan hóa trên website + Hỗ trợ nhiều loại đồ thị khác nhau

Chart configuration

- title: cấu hình tiêu đề của biểu đồ + xAxis: cấu hình trục x của biểu đồ + yAxis: cấu hình trục y của biểu đồ + series: bao gồm dữ liệu và kiểu hiển thị của dữ liệu đó + legend: chú thích cho dữ liệu + grid: cấu hình khung cho biểu đồ, dùng khi hiển thị nhiều biểu đồ + tooltip: hiển thị thông tin đồ thị khi di chuyển chuột + toolbox: hiển thị các công cụ cho biểu đồ: lưu biểu đồ, zoom, reset,... + backgroundColor: hiển thị màu nền + Một số cấu hình cho animation

Ưu điểm

- Rất nhiều feature: ECharts cung cấp một bộ tính năng và tùy chọn toàn diện, cho phép người dùng tạo nhiều loại biểu đồ và hình ảnh trực quan. + Tương thích nhiều trình duyệt: tương thích với tất cả các trình duyệt web chính, đảm bảo rằng các biểu đồ của bạn sẽ trông và hoạt động giống nhau trên tất cả các thiết bị. + Mã nguồn mở: ECharts là một thư viện mã nguồn mở, nghĩa là nó miễn phí sử dụng và có thể được tùy chỉnh để đáp ứng các nhu cầu cụ thể của bạn. + Document tuyệt vời: rõ ràng, cung cấp đủ thông tin, ví dụ cụ thể cho từng tính năng, loại biểu đồ + Cộng đồng lớn, tích cực: ECharts có một cộng đồng các nhà phát triển lớn và tích cực, điều đó có nghĩa là các lỗi nhanh chóng được sửa và các tính năng mới thường xuyên được bổ sung. + Được cập nhật thường xuyên

Nhược điểm

- Không phù hợp những người mới bắt đầu không quen với trực quan hóa dữ liệu và JavaScript + Có quá nhiều tùy chọn, có thể gây khó khăn khi mới làm quen với Echarts + Gặp một số vấn đề khi xử lý các tập dữ liệu lớn hoặc hình ảnh phức tạp, đặc biệt trên các thiết bị cũ + Được thiết kế chủ yếu cho ứng dụng web và có thể không phù hợp để sử dụng trong các loại ứng dụng hoặc môi trường khác

Leaflet

WHAT

- Leaflet là một thư viện mã nguồn mở JavaScript được sử dụng để hiển thị bản đồ tương tác trên các trang web. + Leaflet được phát triển bởi Vladimir Agafonkin và được cập nhật và duy trì bởi một cộng đồng lớn các nhà phát triển trên toàn thế giới. + Leaflet cũng có khả năng tích hợp với nhiều dịch vụ bản đồ trực tuyến như OpenStreetMap, Mapbox, Esri, Google Maps và bất kỳ dịch vụ nào hỗ trợ Web Map

Các Thành Phần Chính

Title layer Leaflet cho phép hiển thị các lớp dữ liệu trên bản đồ, gọi là Tile layer. Các Tile layer có thể được lấy từ nhiều nguồn khác nhau như OpenStreetMap, Mapbox, hoặc Google Maps
Maker Đây là thành phần được sử dụng để thêm các đối tượng như địa điểm, địa chỉ, hoặc điểm quan trọng khác trên bản đồ.
Popup Leaflet cho phép thêm Popup, là một cửa sổ nhỏ hiển thị thông tin chi tiết về các đối tượng trên bản đồ khi người dùng nhấp vào.
Layer control Leaflet cung cấp Layer Control, là một công cụ để người dùng chuyển đổi giữa các lớp dữ liệu trên bản đồ: + **baseMaps**: Một đối tượng chứa các lớp dữ liệu được sử dụng như lớp cơ sở của bản đồ. Ví dụ: các lớp dữ liệu bản đồ địa hình, bản đồ vệ tinh, bản đồ địa lý. + **overlayMaps:** Một đối tượng chứa các lớp dữ liệu thêm vào bản đồ. Ví dụ: lớp dữ liệu điểm địa danh, lớp dữ liệu đường phố, lớp dữ liệu vùng địa lý.
GeoJSON Leaflet hỗ trợ định dạng dữ liệu địa lý GeoJSON, cho phép dễ dàng hiển thị các đối tượng địa lý trên bản đồ.
Event handling Leaflet cho phép xử lý các sự kiện như nhấp chuột, kéo thả, hoặc lướt trên bản đồ để tương tác với các đối tượng trên bản đồ.
Plugins Leaflet có nhiều plugin hỗ trợ cho các tính năng bổ sung, ví dụ như hiển thị đường đi, hiệu ứng động, và tương tác với cơ sở dữ liệu. + Leaflet.markercluster + Leaflet.draw + Leaflet.ajax + Leaflet.heat

Ưu Điểm

Nhẹ và nhanh: Leaflet là một thư viện JavaScript nhỏ gọn, được thiết kế để có thể tải nhanh và sử dụng hiệu quả trên các thiết bị di động và máy tính bảng.
Dễ sử dụng: Leaflet được thiết kế để dễ sử dụng cho cả những người mới bắt đầu và những người có kinh nghiệm với các thư viện bản đồ. Nó cung cấp một API đơn giản và dễ hiểu để thao tác các đối tượng địa lý trên bản đồ.
Đa nền tảng: Leaflet hỗ trợ nhiều nền tảng và trình duyệt khác nhau, bao gồm cả thiết bị di động và máy tính bảng. Điều này đảm bảo rằng ứng dụng bản đồ sử dụng Leaflet có thể hoạt động trên đa nền tảng và đa trình duyệt.
Hỗ trợ các dịch vụ bản đồ phổ biến: Leaflet hỗ trợ nhiều dịch vụ bản đồ phổ biến như OpenStreetMap, Google Maps và Bing Maps. Nó cũng cho phép người dùng tạo các bản đồ tùy chỉnh với các dữ liệu địa lý của riêng họ.
Cộng đồng sử dụng lớn: Leaflet là một thư viện mã nguồn mở, có một cộng đồng sử dụng lớn và tích cực phát triển và cập nhật các plugin, bản vá lỗi và tài liệu để giúp người dùng sử dụng thư viện một cách dễ dàng và hiệu quả hơn.
Tính mở rộng và tích hợp dễ dàng: Leaflet cho phép tích hợp và mở rộng các plugin và thư viện khác để cung cấp các tính năng bổ sung cho ứng dụng bản đồ. Nó cũng hỗ trợ nhiều định dạng dữ liệu địa lý khác nhau để tải và hiển thị trên bản đồ.

Nhược Điểm

Giới hạn khả năng xử lý dữ liệu lớn: Leaflet được thiết kế để hiển thị các bản đồ tương tác đơn giản và không phù hợp với các ứng dụng yêu cầu xử lý dữ liệu lớn. Khi số lượng đối tượng trên bản đồ tăng, nó có thể dẫn đến tốc độ chậm và hiệu suất giảm.
Thiếu tính năng mở rộng về 3D: Leaflet chủ yếu hỗ trợ bản đồ 2D và không có tính năng tích hợp 3D tích cực. Có một số plugin có thể giúp tích hợp 3D vào Leaflet, nhưng chúng thường không đáp ứng được yêu cầu cao về hiệu suất và tính năng.
Không có tính năng định vị GPS tích hợp: Leaflet không tích hợp sẵn tính năng định vị GPS, điều này có nghĩa là người dùng cần phải sử dụng các thư viện và API bên ngoài để tích hợp tính năng này vào ứng dụng.
Hạn chế về tính năng và tính năng tùy chỉnh: Leaflet cung cấp các tính năng cơ bản để hiển thị bản đồ và đối tượng địa lý trên bản đồ, nhưng nó thiếu một số tính năng và tính năng tùy chỉnh so với một số thư viện bản đồ khác.
Tài liệu không đầy đủ: Mặc dù Leaflet có tài liệu chi tiết và hướng dẫn sử dụng, nhưng nó không đầy đủ và chi tiết bằng một số thư viện khác. Điều này có thể làm cho việc học và sử dụng Leaflet có thể khó khăn hơn đối với những người mới bắt đầu.

Ứng Dụng

Ứng dụng định vị GPS: Leaflet có thể được sử dụng để tạo các ứng dụng định vị GPS như các ứng dụng điều hướng đường phố, ứng dụng theo dõi vị trí, ứng dụng đi du lịch, v.v. Leaflet có thể kết hợp với các dịch vụ bản đồ như OpenStreetMap, Mapbox và Google Maps để tạo ra các ứng dụng định vị GPS mạnh mẽ.
Ứng dụng quản lý tài nguyên và môi trường: Leaflet có thể được sử dụng để hiển thị các bản đồ tài nguyên và môi trường, bao gồm các khu vực bảo tồn thiên nhiên, địa điểm khai thác mỏ và các bản đồ chủ đề liên quan đến môi trường. Các đối tượng địa lý có thể được hiển thị với các biểu tượng và màu sắc khác nhau để giúp người dùng phân biệt giữa chúng.
Ứng dụng theo dõi và giám sát thời tiết: Leaflet có thể được sử dụng để hiển thị các bản đồ thời tiết tương tác, bao gồm nhiệt độ, độ ẩm, tốc độ gió và các thông số khác. Người dùng có thể chọn các khu vực cụ thể trên bản đồ để xem thông tin thời tiết và dự báo cho các khu vực đó.

Ứng dụng định vị tài sản và theo dõi hàng hóa: Leaflet có thể được sử dụng để theo dõi vị trí của tài sản và hàng hóa, bao gồm các phương tiện vận chuyển và đường ống dẫn. Các đối tượng địa lý có thể được đánh dấu trên bản đồ và người dùng có thể theo dõi chúng trong thời gian thực.
Ứng dụng quản lý dữ liệu địa lý: Leaflet có thể được sử dụng để hiển thị và quản lý dữ liệu địa lý, bao gồm các lớp dữ liệu và thông tin địa lý liên quan. Người dùng có thể chọn các lớp dữ liệu cụ thể để hiển thị trên bản đồ và tùy chỉnh chúng theo cách của riêng mình.

Datapine

WHAT

Datapine là một công cụ Business Intelligence phân tích và trực quan hoá dữ liệu trực tuyến. Nền tảng phân tích dữ liệu của datapine được cung cấp và quản lý trên một hạ tầng đám mây thay vì cài đặt trên máy cục bộ. Nó kết nối và tổng hợp dữ liệu từ nhiều nguồn khác nhau, sau đó phân tích và hiển thị kết quả dưới dạng bảng, biểu đồ và báo cáo để hỗ trợ việc ra quyết định kinh doanh. Nó có thể lấy data từ nhiều nguồn: Các hệ CSDL như MySQL, MongoDB, SQL Server, Oracle,… Các tệp file như: csv, excel,… Các MXH: Facebook, Twitter, Instagram,… Các dịch vụ Marketing và quảng cáo: Google Ads, Google Analytics,...

Lợi ích

Datapine đã giúp nhiều doanh nghiệp và tổ chức trên toàn thế giới trong việc quản lý dữ liệu và phân tích dữ liệu.
Một số lợi ích tiêu biểu của datapine:

- **Dễ sử dụng** + **Tính tương thích và tích hợp** + **Tốc độ phân tích** + **Bảo mật**

Những thách thức

Giới hạn về việc tuỳ chỉnh - Datapine khá giới hạn trong việc tuỳ chỉnh các biểu đồ nên đôi khi chúng ta khó có thể tạo ra một biểu đồ đúng ý mình hoàn toàn. - Giải pháp: Cung cấp thêm những chức năng tùy chỉnh.
Yêu cầu về kết nối Internet - Cần kết nối mạng để sử dụng. Tốc độ mạng sẽ ảnh hưởng rất lớn đến hiệu suất và tốc độ truy cập vào datapine. - Giải pháp: sử dụng mạng có tốc độ cao để sử dụng ổn định.
Tải lượng người dùng đồng thời + Nhiều người dùng truy cập và làm việc trên datapine thì có thể ảnh hưởng đến hiệu suất và tốc độ xử lý, gây ức chế cho người dùng. + Giải pháp: Nâng cấp server để cải thiện tải lượng
Tốc độ và chi phí + Datapine tốn khá nhiều thời gian để load và xử lý những tập dữ liệu lớn + Tốn phí để sử dụng, chỉ được sử dụng 14 ngày free

Tiêu chí	Data pine	Tableu	PowerBI
Giao diện người dùng	Thân thiện, dễ sử dụng, tùy chỉnh	Thân thiện, trực quan nhưng có một số yêu cầu kỹ thuật để sử dụng	Thân thiện, trực quan, nhưng có một số tính năng phức tạp hơn Datapine
Truy cập dữ liệu	Đều cho phép người dùng kết nối với nhiều nguồn dữ liệu khác nhau	Same	Same
Phân tích dữ liệu	Tính năng phân tích dữ liệu cao cấp: phân tích dữ liệu tương quan, thời gian thực, định lượng và định tính	Tableau và Power BI có các tính năng phân tích dữ liệu mạnh mẽ hơn với các công cụ tạo đồ thị, tính toán trên nhiều bảng và phân tích dữ liệu địa lý	Same
Hiệu suất và tốc độ	Đều có tốc độ xử lý nhanh chóng, cho phép người dùng truy cập và phân tích dữ liệu nhanh chóng, hiệu quả	Same	Same
Giá cả	Cung cấp các gói dịch vụ khác nhau	Giá khá cao, phù hợp với các doanh nghiệp lớn	Same

Tổng kết

- Phù hợp với các doanh nghiệp nhỏ chưa có đủ kinh phí
- Những người mới bắt đầu và có đam mê trong việc phân tích dữ liệu nhưng chưa có nhiều kĩ thuật chuyên môn

---

MongoDB and MapReduce

MongoDB

ĐỊNH NGHĨA

- MongoDB là một hệ quản trị cơ sở dữ liệu phi cấu trúc hướng tài liệu (document+oriented) được phát triển bởi công ty MongoDB Inc. MongoDB là một cơ sở dữ liệu NoSQL + MongoDB là một lựa chọn phổ biến cho việc lưu trữ lượng lớn dữ liệu cần truy cập và cập nhật dễ dàng. + MongoDB là phần mềm miễn phí và mã nguồn mở, và có sẵn trên nhiều nền tảng. Nó cũng có sẵn dưới dạng dịch vụ dựa trên đám mây.

CẤU TRÚC CỦA MongoDB

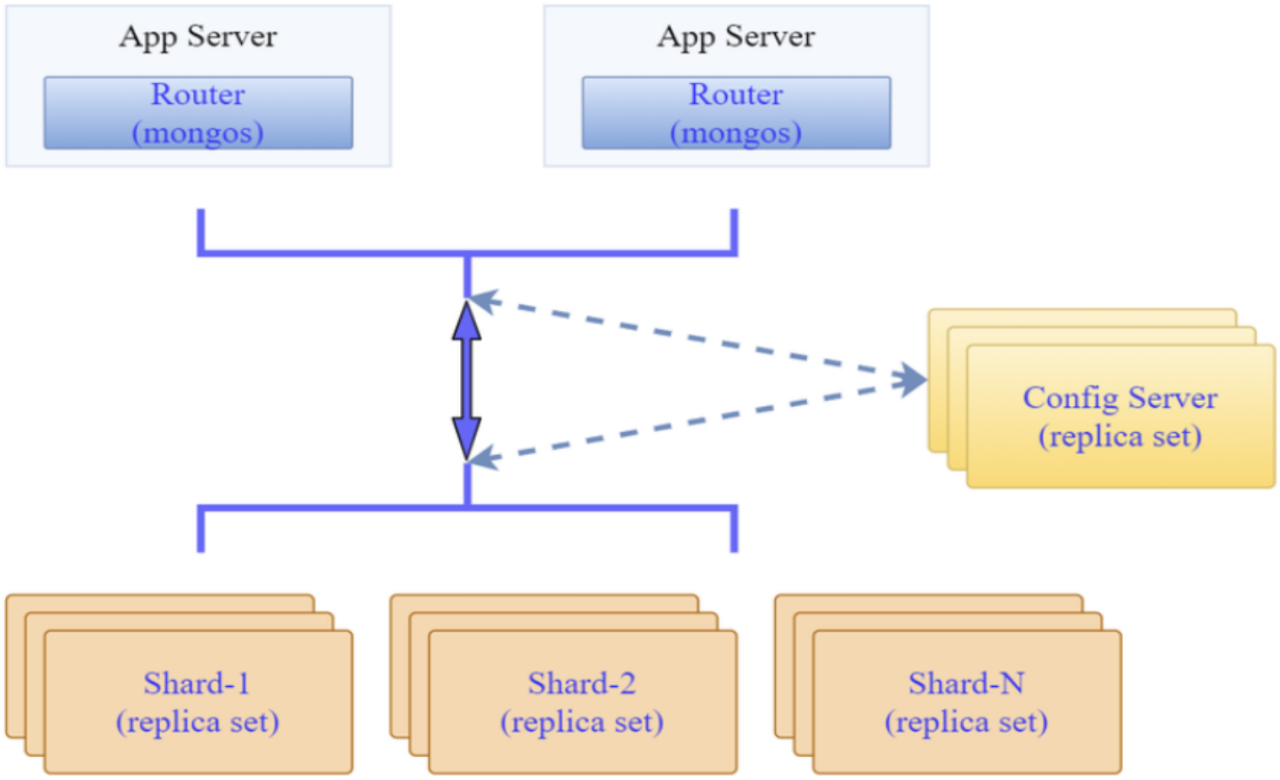
- **Database:** Nó có thể được coi là vùng chứa vật lý cho dữ liệu. Mỗi database có tập file riêng trên file system. Mỗi database có thể chứa nhiều tập hợp (collections), và mỗi tập hợp chứa nhiều documents. + **Collection:** Một nhóm các database document có thể được gọi là một collection. RDBMS tương đương với collection là một table. + **Document:** Một tập hợp các cặp key – value có thể được chỉ định là một document. Các document được liên kết với các dynamic schema.

CÁCH HOẠT ĐỘNG CỦA MongoDB

- MongoDB hoạt động theo cơ chế replica set và sharding để đảm bảo tính sẵn sàng và khả năng mở rộng của hệ thống. + Replica set là cơ chế sao lưu dữ liệu bằng cách sao chép các tài liệu từ một node (nút) chính đến các node sao lưu (secondary nodes). + MongoDB sử dụng cơ chế sharding để phân tán dữ liệu trên nhiều node để mở rộng khả năng lưu trữ và xử lý của hệ thống. Các tài liệu được phân tán vào các shard (node lưu trữ dữ liệu) theo một key được xác định trước. Các yêu cầu truy vấn được chuyển đến shard tương ứng với key đó để tối ưu hiệu suất truy vấn.

- Và một mongodb sử dụng sharding sẽ được gọi là một "sharded cluster". Một MongoDB sharded cluster bao gồm các thành phần chính sau: + **shard:** Mỗi shard chứa một phần của dữ liệu đã được shard. Mỗi shard này lại có thể được triển khai dưới dạng một replicaset để tăng tính dự phòng cho dữ liệu của nó quản lý. + **mongos:** Hoạt động như một query router, là phần giao diện với các client với sharded cluster. Client sẽ chỉ cần biết kết nối tới mongos, phần còn lại là kết nối tới shard nào, replicas nào sẽ do mongos điều phối và trong suốt với client + **config servers**: Chứa thông tin

metadata và các tham số cấu hình cho cluster. Ví dụ thông tin cấu hình các shard, các replicaset.. được lưu ở config server này. Và config server cũng có thể triển khai dưới dạng replicaset.



Hình ảnh tổng quan MongoDb khi hoạt động theo cơ chế replica set và sharding

ƯU ĐIỂM

- **Dữ Liệu Linh Hoạt:** Dữ liệu trong mongoDB được lưu trữ dưới dạng JSON, không bị bó buộc về số lượng field, kiểu dữ liệu …, có thể thoải mái insert dữ liệu mình muốn + **Hiệu Suất Cao:** MongoDB được thiết kế cho hiệu suất cao, và nó có thể mở rộng để xử lý lượng dữ liệu lớn + **Khả Năng Mở Rộng:** MongoDB có khả năng mở rộng bằng cách thêm một hoặc nhiều node vào cluster + **Dễ Sử Dụng:** MongoDB tương đối dễ sử dụng, ngay cả đối với các lập trình viên không quen thuộc với các cơ sở dữ liệu NoSQL + **Mã Nguồn Mở**: MongoDB là phần mềm mã nguồn mở và có cộng đồng phát triển rất lớn

NHƯỢC ĐIỂM

- **Không Ràng Buộc Dữ Liệu:** MongoDB không có các tính chất ràng buộc như trong RDBMS nên dễ bị làm sai dữ liệu + **Không hỗ trợ Transaction:** MongoDB không hỗ trợ transaction vì vậy có thể gặp khó khi xây dựng các hệ thống cần sử dụng transaction (như ngân hàng) + **Không Hỗ Trợ Join:** Không hỗ trợ join giống như RDBMS nên khi viết function join trong code ta phải làm bằng tay khiến cho tốc độ truy vấn bị giảm. + **Sử Dụng Nhiều Bộ Nhớ**: do dữ liệu lưu dưới dạng key+value, các collection chỉ khác về value do đó key sẽ bị lặp lại. + **Bị Giới Hạn Kích Thước Bản Ghi**: mỗi document không được có kích thước quá 16Mb và level các document con trong 1 document không được quá 100

SO SÁNH CÁC DATABASE SQL VỚI MongoDB NoSQL

	SQL	NoSQL
Khả năng mở rộng	SQL databases có thể mở rộng theo chiều dọc	NoSQL databases có thể mở rộng theo chiều ngang
Được sử dụng tốt nhất cho	RDBMS database là tùy chọn thích hợp để giải quyết các vấn đề về ACID. Atomicity – Tính toàn vẹn Consistency - Tính nhất quán Isolation - Tính độc lập Durability - Tính bền vững	NoSQL được sử dụng tốt nhất để giải quyết các vấn đề về tính khả dụng của dữ liệu

	SQL	NoSQL
Hiệu suất	Thường có hiệu suất tốt khi truy vấn dữ liệu trong các bảng với quan hệ phức tạp.	Thường có hiệu suất tốt khi truy vấn dữ liệu không cần quan hệ với nhau và truy vấn dữ liệu theo tài liệu.
Độ tin cậy và độ bảo mật	Có các tính năng bảo mật và kiểm soát truy cập tốt hơn, đồng thời có các tính năng sao lưu và phục hồi dữ liệu nhanh chóng.	Các hệ thống NoSQL thường được thiết kế để có thể xử lý các cấp độ lỗi khác nhau. Tuy nhiên, do tính năng mở rộng ngang của NoSQL, các hệ thống này thường khó bảo mật hơn các hệ thống SQL.
Kết luận	Dự án đã có yêu cầu dữ liệu rõ ràng xác định quan hệ logic có thể được xác định trước.	Phù hợp với những dự án yêu cầu dữ liệu không liên quan, khó xác định, đơn giản mềm dẻo khi đang phát triển

	SQL	**NoSQL**
Thiết kế cho	RDBMS truyền thống sử dụng cú pháp và truy vấn SQL để phân tích và lấy dữ liệu để có thêm thông tin chi tiết.	Hệ thống cơ sở dữ liệu NoSQL bao gồm nhiều loại công nghệ cơ sở dữ liệu khác nhau. Các cơ sở dữ liệu này được phát triển để đáp ứng nhu cầu trình bày cho sự phát triển của ứng dụng hiện đại.
Loại	SQL databases là cơ sở dữ liệu dựa trên bảng	NoSQL databases có thể dựa trên tài liệu, cặp key-value, cơ sở dữ liệu biểu đồ

	SQL	NoSQL
Cấu trúc dữ liệu	Cấu trúc dữ liệu phải được xác định trước và có thể thay đổi ít. Việc thêm hoặc sửa các cột phải được thực hiện bằng các câu lệnh ALTER TABLE và có thể ảnh hưởng đến hiệu suất truy vấn	Cấu trúc dữ liệu không cần được xác định trước và có thể thay đổi dễ dàng.