1. **Core data concepts**
2. **Classify data**

* Structured data:

+ tất cả dữ liệu đều có chung các trường và các thuộc tính.

+ lược đồ cho các thực thể dữ liệu có cấu trúc là tabular

+ được lưu trong cơ sở dữ liệu, nhiều bảng có thể reference nhau bằng cách sử dụng giá trị khóa trong mô hình quan hệ.

* Semi-structured data:

+ cũng là những thông tin có cấu trúc, nhưng cho phép biến thể trong những cấu trúc đó giữa các entity instances. Ví dụ: phần lớn khách hang có một địa chỉ email, nhưng vẫn có người có nhiều địa chỉ email, và có người lại không có địa chỉ email.

+ một format phổ biến cho semi-structured là JSON

* Unstructed data

+ có những dạng dữ liệu không phải là structured cũng như semi-structured. Ví dụ: dữ liệu, hình ảnh audio, video, binary files

* Data stores: có 2 dạng lưu trữ data thường dùng:

+ file storages

+ databases

1. **File storage**

* Là phần quan trọng trong các hệ điều hành.
* Files có thể lưu trữ trong các hệ thống file trên đĩa cứng của máy tính, hojcw trên các thiết bị có thể di chuyển được như USB, nhưng phần lớn các tổ chức lưu trứ những dữ liệu file quan trong trong tập trung được gọi là shared file storage system, và các hệ thống đó được quản lý bởi cloud, tiết kiệm chi phí, bảo mật, và đáng tin cậy cho dữ liệu.
* **Delimited text files:**

+ dữ liệu được lưu trữ trong format thuần túy với các trường được phân cách và các hang có ký tự kết thúc hàng.

+ phần lớn dữ liệu được lưu trữ trong format CSV: mỗi trường được ngăn cách nhau bởi dấu phẩy và các hang được kết thúc bằng return/newline.

+ có tab-separate values (TSV) và space-delimited

+ phù hợp với structured data được truy cập bởi nhiều úng dụng và dịch vụ với format người human-readable.

**A screenshot of a computer

Description automatically generated with medium confidence**

* **JSON:**

+ hierarchy document schema

+ entities có nhiều thuộc tính

+ mỗi thuộc tính có thể là một object{} (hoặc là tập hợp các objects[{}])

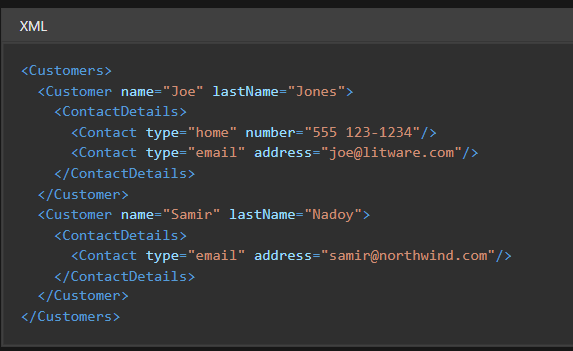
+ phù hợp với cả structed và un-structed.

* **XML:**

**+** humand-readable data format

+ Phần lớn được chuyển sang JSON format để ít phức tạp hơn

+ Sử dụng tag tương tự như html



* **Binary Large Object (BLOB):**

**+** lưu trữ dữ liệu nhị phân,

**+** để theo format dễ đọc thì được map sang các ký tự (ASCII hoặc Unicode)

**+** một số là dịnh dạng tệp, tuy nhiên cụ thể là dữ liệu ko có cấu trúc (images, video, audio, application-specific document)

* **Optimaized file formats:** một số format dữ liệu tối ưu cho storage (nén, chỉ mục,..)

**+** Avro: row-based format: được tạo bởi Apache. Mỗi hàng có header miêu tả structure dữ liệu trong record, header được lưu dưới dạng JSON. Dữ liệu được lưu dạng nhị phân. Sử dụng thông tin trong header để parse dữ liệu nhị phân. Phù hợp cho nén dữ liệu, giảm bộ nhớ lưu trữ và yêu cầu bang thông của mạng

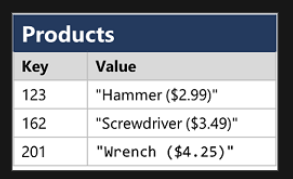
**+ ORC:** tổ chức dưới dạng cột. tối ưu cho các tác vụ đoc và ghi. Chứa các sọc dữ liệu. mỗi stripes chứa dữ liệu của một hoặc 1 tập hợp các cột. mỗi stripes chứa chỉ mục đến hàng trong stripe. Dưới mỗi strips sẽ có những thông itn thống kê (count, sum, mã, min,…)

**+ Parquet:** một dạng lưu dữ liệu dạng cột khác. Dữ liệu mỗi cột được lưu cùng nhau cho mỗi row group. Bao gồm metadata. Đặc biệt được sử dụng trong việc lưu trữ và xử lý các kiêu dữ liệu lồng nhau. Hỗ trợ tốt lược đồ nén và mã hóa

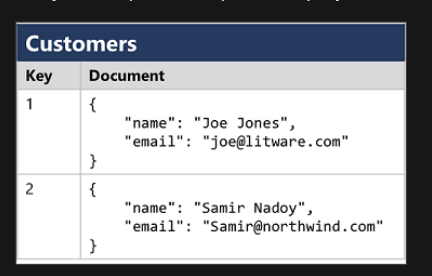
1. **Databases**

* **Relational databases**
* **Non-relational databases (nosql database): có 4 dạn non-relational databases:**

**+** key-valuew database: mỗi record có unique key, và các giá trị có format tùy ý



**+** Document database: một dạng đặc biết của key-value, giá trị được lưu dưới dạng JSON (tối ưu hóa cho parse và query)



+ column family database

A screenshot of a computer

Description automatically generated with medium confidence

+ Grap database (vd: neo4j)

A picture containing font, screenshot, line, circle

Description automatically generated

1. **Transactional data processing**

* Đóng gói các sự kiện cần được theo dõi (ví dụ: chuyển tiền ngân hàng)
* Small, discrete, và là một đơn vị công việc (all or nothing)
* Hệ thống giao tác thường high-volumn, xử lý hàng triệu giao tác trong một ngày
* Dữ liệu phải được truy xuất nhanh
* Online Transactional Processing (OLTP): giải pháp dựa trên hệ cơ sở dữ liệu, tối ưu cho đọc và ghi, để hỗ trợ tải lượng giao tác (created, retrieved, updated, deleted). Đảm bảo tính nhất quán trong dữ liệu được lưu trữ trong csdl. OLTP buộc các giao tác ACID:

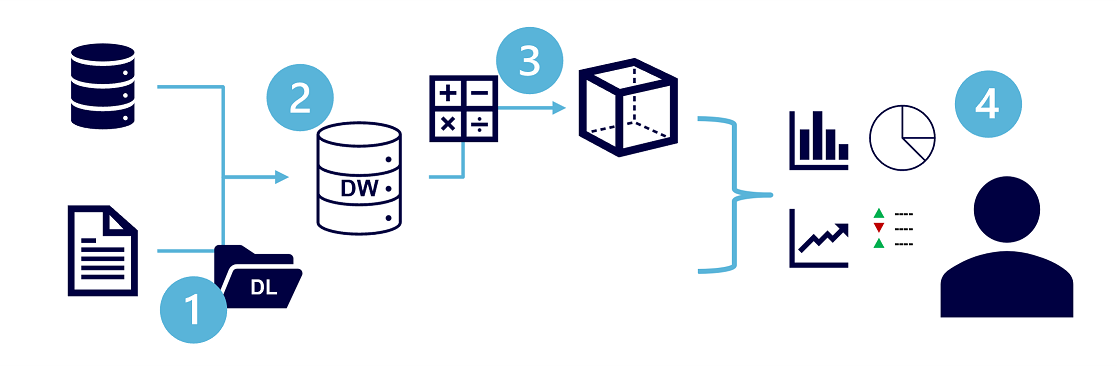
+ Atomic: single unit, all or nothing

+ Consistency: chỉ được lấy dữ liệu trong csdl từ trạng thái valid này sang trạng thái valid khác

+ Isolation: các giao tác đồng thời không bị chồng đè lên nhau và kết quả trong một csdl nhất quán

+ Durability: khi một giao tác đã commit thì dữ liệu được ghi xuống csdl, và ko bị thay đổi hay abort

1. **Analytical data processing**



1. File dữ liệu được lưu trong data lake để phân tích
2. Extract, transform, load (ETL) process copies dât từ files và OLTP database vao trong data warehouse, được tối ưu hóa để đọc. Lược đồ data warehouse dựa trên bảng thự tế chứa numeric values cần được phân tích (vd: sales amounts) với những bảng liên quan dimension chứa những thực hể dựa trên đó có thể phân tích (customer, product)
3. Dữ liệu trong data warehouse có thể được tính gộp và được nạp vào mô hình online analytical processing (OLAP), hoặc cube. Từ bảng thực tế có thể tính intersection của dimensions từ dimension table (vd: tính doanh thu dựa theo ngày, khách hàng, và sản phẩm)
4. Dữ liệu trong datalake, datawarehouse, và trong analytic model có thể được truy vấn để tạo báo cáo, hiển thị hóa, và dashboard.

* Datalakes thường là trong kịch bản xử lý phân tích dữ liệu quy mô lớn, nơi mà khối lượng lớn dữ liệu lưu trữ dựa trên file được thu tập và phân tích
* Data warehouse lữu dữ liệu trong các lược đồ quan hệ để tối ưu hóa cho việc đọc, và truy vấn để hỗ trợ báo váo và hiển thị hóa dữ liệu. datawarehouse có thể yêu cầu giảm chuẩn hóa về data trong OLTP (duplication để làm cho việc truy vấn trở nên nhanh chóng hơn)
* Mô hình OLAP là một nơi lưu trữ kiểu dữ liệu gộp, tối ưu hóa cho việc phân tích. Có thể nhìn dữ liệu gộp tại các cấp dộ khách nhau (vd: doanh số theo region, city, individual address)
* Data scientists có thể làm việc tực tiếp trên file dữ liệu trong data lake
* Data analysts truy vấn bảng trực tiếp từ data warehouse để sinh ra những báo cáo phức tạp và mô phỏng hóa.
* Người sử dụng kinh doanh có thể sử dụng dữ liệu pre-aggregated trong analytic model trong dạng của báo cáo và dashboard

1. **Job roles in the world of data**

* **Database Administrator:**

**+** thiết kế, hiện thực, duy trì, và các khía cạnh vận hành trên csdl (on-premise, cloud).

**+** làm việc với các bên liên quan để hiện thực policies, tools và những quy trình để backup và phục hồi khi có thiên tai hoặc do lỗi từ con người.

**+** quản lý tính an toàn của dữ liệu, và cấp đặc quyền thóng qua dữ liệu, cấp hoặc thu hồi quyền truy cập của người dùng

* **Data Engineer:**

+ phối hợp với các bên liên quan để thiết kế và hiện thực liên hệ dữ liệu. bao gồm: pipelines, các hoạt động làm sạch và chuyển đổi dữ liệu, và dữ liệu lưu trữ cho các hoạt động phân tích

+ sử dụng các công nghệ platform: relational, non-relational db, file stores, và data streams

+ bảo đảm tính bảo mật của dữ liệu và duy trì trong cloud, mở rộng từ on-premise đến cloud.

* **Data Analyst:**

+ khám phá dữ liệu để tìm ra xu hướng và mối liên hệ, thiết kế và xây dựng mô hình phân tích, cho phép phân tích bậc cao thông qua báo cáo và hiển thị hóa dữ liệu

1. **Azure SQL: tên cho các giải pháp csdl quan hệ dựa trên MS sql server bao gồm:**

* Azure SQL DB: PaaS database
* Azure SQL Managed Instance: tự động bảo trì, cho phép cấu hình hiều hơn so vơi SQL DB nhưng đòi hỏi nhiệm vụ administrative nhiều hơn
* Azure SQL VM: máy ảo có thiết lập SQL server, cho phép cấu hình toàn bộ.

1. **Azure Database cho nhiều csdl quan hệ open-source:** MySQL, MariaDB, PostgreSQL
2. **Azure Cosmos DB:**

* Hệ thống Csdl global-scale non-relational (NoSQL) hỗ trợ nhiều APIs, cho phé lữu trữ và bảo trì dữ liệu dưới dạng JSON, key-value paires, column families, và graphs

1. **Azure Storage:** dịch vụ core cho phép lưu tữ dữ liệu trong:

* **Blob containers –** có thể mở rộng, lưu trữ với chí phí hiệu qua cho file nhị phân
* **File shares –** network file shares(vd: mạng trong công ty)
* **Tables –** lưu trữ key-value cho phép đọc ghi nhanh

1. **Azure Data Factory:**

* cho phép định nghĩa và lập lịch data pipeline để chuyển đổi và chuyển dạng dữ liệu.
* có thể tích hợp piplines với những dihcj vụ khác của Azure,
* cho phép nhập dữ liệu từ cloud, xử lý dữ liệu sử dụng điện toán dựa trên cloud
* được data engineer sử dụng để xây dụng giải pháp extract, transform, và load (ETL)

1. **Azure Synapse Analytics:** Giải pháp phân tích dữ liệu toàn diện và thống nhất, cung cấp một giao diện dịch vụ cho nhiều khả năng phân tích, bao gồm:

* Pipelines – cùng công nghệ với Azure Data Factory.
* SQL – csdl SQL có khả năng mở rộng cao, tối ưu hóa cho tải lượng dữ liệu trong warehouse
* Apache Spark – là một hệ thống xử lý dữ liệu phân tán mã nguồn mở, hỗ trợ nhiều dạng ngôn ngữ và APIs (Java, Scala, Python, SQL)
* Azure Synapse Data Explorer – giải pháp phân tích dữ liệu cho hiệu suất cao, được tối ưu hóa cho truy vấn thời gian thực của log, và sử dụng dữ liệu từ xa sử dụng Kusto Query Language (KQL)
* Data engineer sử dụng để tạo giải pháp phân tích dữ liệu thống nhất, kết hợp lấy dữ liệu pipelines, data warehouse, và data lake thông qua một dịch vụ duy nhất
* Data Analysis sử dụng SQL và Spark pools thông qua notebooks tương tác để khám phá và phân tích dữ liệu, và tận dụng sự tích hợp với các dịch vụ như Azure Machine Learning và Microsoft Power BI để tạo mô hình dữ liệu và xuất cái nhìn bên trong dữ liệu

1. **Azure Databricks**

* Phiên ản tích hợp Azure của Databricks platform,
* Kết hợp Apache Spark xử lý dư liệu với csdl SQL và tích hợp giao diện quản lý cho phép phân tích dữ liệu có khả năng mở rộng
* DE sử dụng Databricks và Spark skills để tạo kho dữ liệu phân tích lưu trữ trong Azure Databricks .
* DA sử dụng notebook hỗ trọ trong Azure Databricks đẻ truy vấn và hiển thị hóa dữ liệu một cách dễ dàng thông qua giao diện web.

1. **Azure HDInsight:** Azure HDInsight là một dịch vụ Azure cung cấp các cụm được lưu trữ trên Azure cho các công nghệ xử lý dữ liệu lớn nguồn mở phổ biến của Apache, bao gồm:

* Apache Spark – hệ thống xử lý dữ liệu phân tán, hỗ trọ nhiều ngôn ngữ và API,..
* Apache Hadoop – hệ thống phân tán sử dụng MapReduce để xử lý khối lượng lớn dữ liệu hiệu quả thông qua nhiều cụm nodes. Mapreduce có thể được viết bằng Java hoặc bởi giao diện như là Apache Hive – API dựa trên SQL chạy trên Hadoop
* Apache Hbase: một hệ thống mã nguồn mở cho lưu trữ NoSQL có khả năng mở rộng và truy vấn.
* Apache Kafka – tin nhắn broker cho xử lý dữ liệu dòng
* DE sử dụng Azure HDInsight để hỗ trợ phân tích dữ liệu lớn dựa trên nhiều công nghệ mã nguồn mở.

1. **Azure Stream Analytics:**

* Xử lý dòng thời gian thực, bắt những dòng dữ liệu từ input, và ứng dụng truy vấn để xuất và xử lý dữ liệu từ dòng dữ liệu input, và viết kết quả ra cho vieejv phân tích và xử lý sau này.
* DE sử dụng

1. **Azure Data Explorer:**

* Cung cấp truy vấn với hiệu suất cao cho log và dữ liệu và dữ liệu từ xa tương tự Azure Data Explorer
* DA sử dụng đẻ truy vấn và phân tích dữ liệu bao gồm một thuốc tính timestamp, như là thường được thấy trong file log và IoT dữ liệu từ xa.

1. **Microsoft Purview:**

* Giải pháp cho quản trị dữ liệu doanh nghiệp rộng
* Tạo map cho dữ liệu và track dữ liệu thông qua nhiều nguồn và hệ thống dữ liệu,
* Dữ liệu đáng tin cậy cho phân tích và báo cáo
* DA sử dụng để quản trjij dữ liệu thông qua doanh nghiệp và đảm bảo tính thống nhất dữ liệu được sử dụng trong phân tích

1. **Power BI:**

* Platform cho mô hình phần tích dữ liệu và báo cáo mà DA sử dụng để tạo và chia sẻ hiện thị hóa dữ liệu tương tác. Power BI report đựco tạo ra bằng cách sử dụng PowerBI Desktop.