LTC - ~~3/1/2024 2+2~~

~~vault8.com~~

Tableau

→ Data visualisation tool and Business intelligence (BI) tool

→ BI - process of collecting analysing and presenting business data so that companies can make better decisions.

→ Can handle Big Data when Excel struggles.

→ Tableau Desktop → used to create reports and dashboard.

→ Tableau Public → Free version to create and share dashboards online

Connecting Data

You can connect Tableau to Excel CSV, SQL, My SQL, Google Sheet etc.

Start Tableau → Connect to Data → choose File / Server → Import.

<h> <h>

Tableau Workspace.

→ Data Pane - In left side it shows dimensions and measures

→ Rows and Columns Shelf - Used to drag fields and build charts.

→ Marks Card - Controls color, size, labels, shapes, tooltips

→ Filters and Pages - Filter data or create animations

→ Dashboard Tab - Combine multiple sheets into one view

Dimensions - Categorical field
Measure - Numerical value

Filters - Drag a field to Filters to include / exclude data

Sort - Right click → Sort to ascending / descending

Calculated Field

→ Create new fields using formula
→ Right click anywhere in the Data Pane → Select Create → Calculated Field

Dashboard

Combine multiple work sheets By dragging charts into dashboard

Extension used is .twbx

Data Ingestion

→ Process of gathering managing and utilizing data efficiently

→ Foundamental step in data preprocessing

→ It involves the seamless importation transfer or loading of raw data from diverse external source into a centralized system.

→ Then it awaits for further processing and analysis

Key steps

1) Data collection - Gather raw data from various sources

2) Data transformation - Clean, normalize and enrich (light processing)

3) Data loading - Move the transformed data into target system

Types of Data Ingestion

Batch ingestion

→ Data is collected and processed at scheduled intervals

eg:- Payroll, Daily reports.

Real-time Ingestion

→ Data is ingested as soon as it is generated

eg:- → It provide instant insights

eg:- Banking apps ingesting real time

transaction logs

Micro - Batching

→ A hybrid approach - small batches processed frequently

Eg:- Semi - live analytics.

Batch Processing

→ Technique where data is collected, stored over a period of time

→ Processed together as a single unit or batch instead of being processed immediately

Data Collection - Data from various sources is collected.

Batch Formation - Grouped based on time

Processing - Process the entire batch at once

Output / Storage - Processed results are stored in a database, file or sent to other system for further use.

Common Batch Processing Tools

→ Apache Spark
→ Talend
→ AWS Glue

Stream Processing

→ Continuous and real-time processing of data as it flows in - one record or event at a time

→ No delay between data arrival and processing

~~Data in cont~~

Data ingestion - Continuously ingested from sources like Kafka, sensors

Real-time Processing - Each item is processed as it arrives

Output Delivery - Results are pushed immediately to database

Feature Engineering

→ Broad process that includes, creating new features, transforming existing ones, and selecting the most relevant features.

→ Enhance the dataset comprehensively to improve model performance.

→ Involves a variety of methods, including selection, extraction, transformation and creation of features.

Feature Selection

→ A specific technique within feature engineering focused on selecting the most relevant features from the

dataset

→ Simplify the model and improve performance by reducing the no: of features

→ Involves methods like filter, droppers and embedding

~~Feature~~ Feature Extraction

→ Specific techniques within feature engineering aimed at transforming existing features into new, more informative features.

→ Involves methods like PCA, ICA and polynomial features

Feature Selection Methods

Filtering Method

Filters selected features based on statistical measures and don't involve any ml algorithm

Common techniques

→ Correlation Coefficient
Purpose - Identify the linear relationship between features and the target variable

Chi-Squared test
→ Test the independence between categorical features and the target variable

ANOVA
→ Compare the mean of different groups and determine if the mean of a feature differs significantly from others

Techniques for Feature Extraction

Principal Component Analysis (PCA)

Reduces the dimensionality of the data by transforming it into a set of uncorrelated variable called principal component.

Independent Component Analysis

Find statistically independent components not just uncorrelated

Feature Construction

Creating new features based on existing ones, such as polynomial features interaction terms

Binning

Convert continuous variable to categorical bins

Kafka

Publish / subscribe with message queues.

Messaging design pattern used in software system to enable loose coupling (connected in such a way that they depend on each other as little as possible) between components

Publishers - Produce messages
→ Don't know who recieves them

Subscribers - Consume message
- They don't know who sent them

Kafka

Message / Event streaming platform
Producers (Publishers) push message to Kafka
Consumers (Subscribers) listen and recieve messages

Kafka Data functions

→ Collection
→ Storage
→ Compaction
→ Distribution
→ Tracking

Kafka Message

→ known as an event in Kafka
→ Unit of data
  • Row, record, map

→ Treat all events as byte array

Message Content

→ Key - defined by producer
  Need not be unique, not mandatory
  used for Partitioning

→ Value (Byte array)
  Content of the message
  user defined

→ Time stamp
  Automatically time stamped

Topics

→ Messages are stored in Topics
eg:- Table in a database which contain records which is sent.
Partitions — made — Sub topics.

Brokers

→ Worked Brain
→ Recieve messages from producer and stores locally as logs

→ Multiple Kafka brokers can be clustered together to form a single Kafka cluster

→ Kafka instance - one Kafka broker process running on a server.

→ Kafka broker instance act as the active controller for the cluster

Logs

→ Physical files for storing data

Partition

→ Each topics can have 1 to n Partition
→ allow Kafka to scale
→ Each partition has a leader broker (Broker Power)
In order to write a specific partition the message needs to be sent towards corresponding leader

→ Enable consumers to share workload through consumer group

→ Each published message gets stored in one partition
→ Message ordering guaranteed only within a partition

→ ~~Send~~ the partition for a message is determined by the message key

→ Kafka uses a hashing function to allocate a partition based on the message key

→ Messages with the same key will always

always end up in the same partition

Consumer Groups

→ A group of consumers who share a topic workload

→ Each message goes to only one consumer in a group

→ Consumers splits splits workload through partitions
  → Num of partition >= No of consumers

If there are more consumers than partitions If then there will be consumers with no work.

~~When to~~
Consumer Offset Management

Offset

Number to track message consumption by consumer and partition

→ Broker keeps track of what is sent and acknowledge using 2 offset value

→ Current offset - last message sent to a given consumer

→ Committed offset - last message acknowledged by consumer

→ Broker resend message if the message is not received after certain time

→ Ensure atleast one delivery

Pandas

- dt.day_name()
  Pandas date time accessor

When you have a column of datetime value .dt lets you extract parts of the date time.

data.loc[:, 'day_of_week'] ← which column to select.
↑ which rows to select

- .loc – label based indexing used to select rows and columns
- : – all rows

Visualisation Techniques

Discrete value

Bar Plots

→ Usage: Ideal for showing frequency of each category in discrete data.

Count Plots

Similar to bar plots, but ~~usually~~ specifically designed to show the count of occurrences of each category directly

Tools

Matplotlib : use plt.bar() for bar plot

Seaborn : use sns.barplot()
sns.countplot()

Continuous Data

Histogram

useful for displaying the distributions of continuous data. It groups the data into bins and shows the frequency of data points

Density Plots

-> Similar to histograms
-> Provide a smoothed version of distributions

Tools

plt.hist() - histogram
plt.fill_between() - density plots
sns.histplot - histogram
sns.kdeplot - density.

ETL - (Extract, Transform, Load)

→ Traditional technique of extracting raw data, transforming it as required for the users and storing it as required for the users

→ Extract - It is the process of extracting raw data from all available data source

→ Transform - Then Extracted data is immediately transformed as required by the users.

→ Load - The transformed data is then loaded into the data warehouse

→ Major drawback is that once the data is transformed and stored the original raw data is lost.

→ But for structured and smaller dataset

ELT - (Extract, Load Transform)

→ Handles structured semistructured and unstructured data

→ Flexible allowing transformation based on business needs.

→ Retain original data.

→ In ELT data from source system is first loaded into the data warehouse without full transformation

Apache Airflow

Directed Acyclic Graph (DAGs)

→ Workflow blueprints written in Python
→ Define tasks with clear, ordered steps
→ DAGs prevent loops, ensuring forward progress

Operators

→ Specialised tools for tasks
→ Pre-defined templates for jobs

Defining Tasks

→ Task is a step in the workflow
→ Each task usually runs a small Python function

The scheduler

→ Continuously monitors all defined workflows

→ Triggers tasks based on dependencies and schedules.

Workers / Executor

Executor assigns specific tasks to the workers

Data Warehousing OLAP vs OLTP

→ Subject oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process

Need for Data Warehousing

→ Handles large data volumes

→ Provides centralized storage

→ Enables trend analysis using historical data.

Characteristics of Data Warehousing

→ Subject-Oriented

OLTP (online Transaction Processing)
Characteristics

→ Fast, frequent Insert
Designed for day to day operations
transaction can be adding and updating documents

→ Fast, frequent INSERT, UPDATE, DELETE operations.
→ Highly normalized
→ Data is current

OLAP

→ Supports decision making and analysis
→ Complex SELECT queries
→ users are few but highly analytical users.
→ Data is historical summarized multi dimensional data

OLAP operations

R

Roll up

→ Reduce dimension

2) Drill down
→ opposite of roll up.
→ Increase detail by descending a hierarchy.

3) Slice
Selects a single dimension from the cube

4) Dice
Select 2 or more dimensions to create a more specific sub-cube

5) Pivot
→ Reorient the data cube to gain new perspective