==============================

Data Analyst Concepts/Terms

==============================

Data Lake: It is a storage repository that can store a vast amount of raw data (Structured, unstructured, semi structured) offers massive scalability and is usually based on distributed and often open-source Technologies.

Data Warehouse: Centralized Repository designed for Query and Analysis, it integrates data from multiple sources and is optimized for data analytics. Ex a retail company can use Data Warehouse to combine sales data, inventory data, customer information to optimize stock level and improve customer purchases.

Data Wrangling: Which is the process of cleaning transforming and organizing raw data into a structure and usable format. It involves tasks like removing duplicates handling missing values and converting data types for example transforming raw sales data into a standardized format with consistent units allows businesses to compare performance and make informed decisions.

Data integration: which is the process sources of combining data from different sources formats and systems into a unified and consistent view it could involve merging customers data from online sense and in-store transactions thereby helping returns gain a holistic View and enhancing personalized marketing.

Data modeling: which is the process of creating a structural representation of real-world data and is by designing a blueprint that defines how data should be organized stored and accessed for example in an e-commerce company data modeling helps Define how customers’ orders products and reviews are related and stored in a database and this obviously makes it easy to retrieve and analyze information efficiently.

Data mining: which is the process of discovering patterns correlations anomalies and statistically significant information from large data sets companies can use data mining to analyze customer purchase history and identify products preferences and so leading to targeted marketing.

Data visualization: which is the art of presenting complex data in visual forms like charts and graphs to make it easy to understand the apps professional grasp Trends patterns and relationships in data and so this is where we can use tools like Tableau and power bi to create powerful data visualizations to them present to the leadership then we have data schema which defines the structure and organization of data in a database or data set it outlines the data types relationships and constraints ensuring consistency and efficient querying a well-defined schema streamlines data management and for example in a customer database a schema could specify that each record includes Fields like the name email and the purchase history ensuring it's easy to consume data then we have structure and structure and semi-structured data so structured data is organized in rows and columns and often in databases so an example is a relational database like SQL with the clear tables and schemas unstructured data lacks specific formatting organization and so examples are emails videos and social media posts and then semi-structured data as some organizational properties but not S3 test structure data and so examples is a Json or XML files with tags but flexible data representations.

Descriptive statistics: which involves summarizing and presenting data to gain insights in the industry is used to understand Trends patterns and key characteristics of the data set using for example methods like the mean median and mode this is different from inferential statistics which involves drawing conclusions and making predictions about a population based on a sample of data he helps us make informed decision by analyzing a subset of the data to infer insights about the entire population and so a car manufacturer might use inferential statistics to estimate the average fuel efficiency of all vehicles produced based on a sample of cars tested in a controlled environment.

Skewness: so a measure of a symmetry of the probability mission of a random variable about its mean and so if the curve is shifted to the left or to the right is said to be skewed positive skewness scene indicates the distribution that is skewed towards to the left while negative skewness indicates that distribution that is Q to the right.

Correlation: which measures the strength and direction of a linear relationship between two variables so it helps uncover correlation between different factors for example a positive correlation between advertising spending and sales indicates that higher spending often leads to an increase in sales.

Causation: that refers to a cause-and-effect relationship between two variables where a change in one variable directly influence a change in another and so in an e-commerce company if we find that free shipping increases sales then we can use this causal insight to drive growth into four-hour sales and so correlation means two variables change together while causation implies that changes in one variable directly in influence the other.

Regression analysis: which is a statistical method to model the relationship between a dependent variable and one or more independent variables often used for prediction and understanding variable relationships and so an example if a company obsess that increase advertising which is the independent variable usually leads to higher sales which is the dependent variable they can use regression analysis to quantify this relationship and predict future sales from plan advertising expenditures next is outlier which refers to a data point that significantly deviates from the overall pattern of a data set and so outliers can impact analysis and decisions and so in finance and your unusually high stock price of a certain day might be an outlier affecting the accuracy of the average stock calculations when we have normalization which is a data pre-processing technique that scales features to a common range for example between zero and one and so in a customer database normalizing income and age helps Bank evaluate loan applications fairly preventing one feature from dominating the decision making process due to its larger scale.

Data governance: which involves creating and enforcing rules processes and policies for managing data across an organization and so it ensures data quality data security and compliance so super important department within a data team.

ETL: which stands for extract transform and load it's a process used in the industry to gather data from various sources and so we have extra then clean structure and enrich the data then transform before finally loading it into a storage or database for analysis which is the load part.

Time series: analysis which involves studying data points collected at a successive evenly spaced intervals over time and so a utility company might analyze historical energy consumption patterns to predict demand ensuring they allocated resources efficiently.

Baby testing: which is a method where two versions A and B of something are compared to determine which performs better it is used often to optimize user experiences and marketing strategies and so for example a website can run on a B test by showing one group of users the original design which is a and another group a new design which is B and then analyze which version leads to uh for example higher click through rate next up is hypothesis testing which is a statistical method used to determine if there is enough evidence to support or reject a claim about a population based on a sample and show a food delivery app might use hypothesis testing to determine if a new discount strategy significantly boosts all the numbers ensuring affecting business strategies.

Predictive modeling: which involves using historical data to create a model that predicts future outcomes and so again for instance an insurance company can use predictive modeling to assess the risk factors and optimize your future claims helping them set appropriate premium rates and manage Financial stability next up is SQL or SQL which is the structure query language and is a domain-specific language used for managing and querying relational databases and it's crucial for data retrieval and manipulation and so for example I use SQL on a daily basis to query data and find insights on a large scale.

Python libraries: these are collections of modules or functions that provide specific functionalities of feature they basically help avoiding the Reinventing the Wheel by providing pre-written code for a common task and so in data analytics the most popular python libraries are NumPy for numerical operations so for example arrays and mathematical functions pandas for data manipulation and Analysis then matplotlib for visualization and plotting scikit learn for machine learning tasks and TensorFlow for deep learning Frameworks.

Month over Month and Elvia: so these are comparisons of data between consecutive months or the same month in different years so for example if a company had sales of hundred dollars in January and 110 in February the month of a month sales growth will be 10 instead if a company has changed of 100 in February 2022 and 120 in February 2023 the year of asa's growth of February would be 20.

YTD all year today: that refers to the period of time from the beginning of the current calendar year up to the present date and so here today is used for tracking financial performance and Trends and so for example imagine that today it's August 1st 2023 so the year today it refers to the time from January 1st to August 1st 2023.

----

kpi: which is a measurable value that reflects how effectively a business is achieving its objectives for example in an e-commerce company might track kpis like conversion rate and average all the value to coach website performance.

<u>Clustering</u>: so clustering is a machine learning technique that group similar data points together based on their characteristics and its use for segmenting and pattern recognition for example a marketing team can apply clustering to group customers with similar purchase behaviors allowing tailored marketing strategies to Boost customer engagement and sales.

<u>Classifications</u>: which is a machine learning task where the goal is to predict the category or class of an object based on its features and so it's used for various applications like flow detections and so a bank can employ classifications to automatically identify fraud ring transactions by analyze transaction patterns and safeguarding customers Financial Security next up is feature engineering so feature engineering is the process of selecting transforming and creating new features from raw data to improve machine learning models' performance for example in email spam classifications features like the email length the number of exclamation marks and presence of certain keywords can be engineered to help the model distinguish between spam and legitimate emails.

<u>Ensemble learning</u>: so it's a technique that combines multiple machine learning models to improve predictive performance and reduce overfitting and so in a medical diagnosing system combining predictions from multiple algorithms like the random Forest the gradient boosting and neural networks can provide a more reliable diagnosis by reducing individual model biases and errors and then last super important concept is NLP or natural language processing and so it is a field of artificial intelligence focused on enabling computers to understand intermittent and generate human language in the industry and so NLP is used for tasks like sentiment analysis chatbots and language translation.

===

===