**Week 1**

**Data Models**

**Online Transaction Processing (OLTP) database** -  Optimized for data processing instead of analysis. Designed to effectively store transactions and help ensure consistency. optimized to read, write and update single rows of data to ensure that business processes go smoothly

**Online Analytical Processing (OLAP) database** - Optimized for analysis in addition to processing and can analyze data from multiple databases

**Primary key** – An identifier in a database that references a column or

a group of columns in which each row uniquely identifies each record in the table.

**Dimensional model** - Relational model that has been optimized to quickly retrieve data from a data warehouse

* Facts -  measurement or metric
* Dimensions -  provides more detail and context regarding that fact (who, what, where, when, why and how)
* Attribute - characteristic or quality of data

**Schemas**

**Star schema** - one fact table that references any number of dimension tables - designed to monitor data instead of analyzing it.  ideal for high scale information delivery, making output more efficient because of the limited number of tables and clear direct relationships

**Snowflake schema** - extension of a star schema with additional dimensions and, often, sub dimensions designed for lightning-fast data processing.

**Flattened schemas** -single table in which each record is represented by a single row of data

### Semi-structured schemas

* **Document schemas**
* **Key-value schemas**
* **Wide-column schemas**
* **Graph schemas**

**Database types**

**Row based databases** - Each row in a table is an instance or an entry in the database and details about that instance are recorded and organized by column.

**Columnar databases** - databases organized by columns which process data quickly, only retrieving information from specific columns

**Single-home databases** -All data is stored in the same physical location distributed databases - collection of data systems distributed across multiple physical locations

**Combined systems** -  database systems that store and analyze data in the same place

**Considerations for building efficient data warehouses**

**Business needs** - questions the organization wants to answer or the problems they want to solve these needs

**Shape and volume of the data** - rows and columns and its arrangement

**Model** the data warehouse will follow

**Elements a database schema should include**

* Relevant data
* Names and data types for each column and each table
* Consistent formatting across data entries
* Unique keys for every database entry and object

Gather information from stakeholders

Create a user-centered design where all of the requirements for the entire team are met

Establish the metrics and what data the target table should contain early on

**ETL**

* transformed before delivery
* Filters out data before TL
* Calculations replace or revise existing columns
* Sensitive information can be redacted
* Good with smaller structured datasets with longer load times but analysis is faster

**ELT**

* Transformed in the destination system
* load all of the data, allowing users to choose which data to analyze at any time
* Calculations are added directly to the existing dataset
* Data has to be uploaded before data can be anonymized, making it more vulnerable
* Good with large amounts of both structured and unstructured data, with fater loading but slower analysis

**Factors of database performance**

* Workload -The combination of transactions, queries, data warehousing analysis, and system commands being processed by the database system at any given time.
* Throughput - The overall capability of the database’s hardware and software to process requests.
* Resources - The hardware and software tools available for use in a database system.
* Optimization - Maximizing the speed and efficiency with which data is retrieved in order to ensure high levels of database performance.
* Contention - When two or more components attempt to use a single resource in a conflicting way.

**Optimize database performance**

* Indexes - Use the keys from the database tables to very quickly search through specific locations in the database
* Partitions - organized by logical groupings
* Logical queries

**Week 3**

**Quality testing**

* Completeness: Does the data contain all of the desired components or measures?
* Consistency: Is the data compatible and in agreement across all systems?
* Conformity: Does the data fit the required destination format?
* Accuracy: Does the data conform to the actual entity being measured or described?
* Redundancy: Is only the necessary data being moved, transformed, and stored for use?
* Timeliness: Is the data current?
* Integrity: Is the data accurate, complete, consistent, and trustworthy?

**Data dictionaries** - a collection of information that describes the content, format, and structure of data objects within a database, as well as their relationships.

**Data lineages** - describes the process of identifying the origin of data, where it has moved throughout the system, and how it has transformed over time.

Queries:

PARTITION BY

    RANGE\_BUCKET(Year, GENERATE\_ARRAY(2015,2022,1))