Data Modelling

Data engineers primarily work on the storage and processing of Data
A list of activities Data engineers may do

- Ingest data from a data source
- Build and maintain a data warehouse
- Create a data pipeline
- Create an analytics table for a specific use case
- Migrate data to the Cloud
- Schedule and automate pipelines
- Backfill data
- Debug data quality issues
- Optimize Queries and design a database.

Conceptual Data modelling => Logical data modelling => Physical data modelling
Data modelling is an iterative process.

RDMS => Relational Database Model was invented at IBM by Edgar Codd(1970), it is defined as a digital database based on the relational model of data, or a software system used to manage relational databases.

Advantages of a Relational Database

- East to use
- Ability to do Joins
- Ability to do aggregations and analytics
- Smaller data volumes
- Easier to change business requirements
- Flexibility of queries
- Modelling data not modelling queries
- Secondary indexes
- ACID transactions => data integrity

ACID properties are properties of a database transaction to Guarantee validity even in the event of errors, power failures.
A => Atomicity

C => Consistency
I => Isolation
D => Durability

When not to use a Relational Database

- When there is a large amount of data
- When there is a need to store different data type formats
- When high throughput is needed =>fast reads
- When there is the need for a flexible schema => Not all Columns have to be used by a flexible row
- When there is a need for high availability
- When there is a need for horizontal scaling

Python wrapper for Postgres SQL => psycopg2

Common types of NOSQL Databases

- Apache Cassandra => Partition Row store
- MongoDB => Document Store
- DynamoDB => key-value store
- Apache HBase => wide column store
- Neo4l => Graph Database

Terminologies in Apache Cassandra

- Keyspace => collection of tables
- Table => A group of partitions
- Row => A single item
- Partition => A fundamental unit of access, collection of rows, How data is distributed
- Primary key => This is made up of partition keys and clustering columns

Apache Cassandra => Apache Cassandra has its own query language

When to Use NoSQL

- Large amounts of data
- If you need horizontal scalability
- If you need high throughput --fast reads
- if you need a flexible schema
- if there is a need for high scalability
- if there is a need to be able to store different types of data formats
- Users are distributed at low latency
- If there is a need for linear scalability => the more nodes are added the faster the performance

When creating Tables in Apache Cassandra always ask the question what queries will I be making, that way you understand How to set the Primary Key.

Relational Databases

OLAP =>Online Analytical Processes, Databases optimized for these workloads allow for complex analytical and ad hoc queries. These type of databases are optImized for the reads and would have lots of aggregations.
OLTP => Online Transactional Processes, Databases optimized for these workloads allow for less complex queries in large volumes. These type of queries for these databases are read, insert, update, delete and it would also have little aggregations.

Normalization: To reduce data redundancy and increase data integrity. Redundancy => reducing the copies of your data

Objectives of Normal form

- To free the database from unwanted insertions, updates and deletion dependencies.
- To reduce the need to refactor the database as new data types are introduced.
- To make the relational model more informative to users
- To make the database neutral to query statistics

The process of Normalization is a step by step process.

- First normal form (1NF)
- Second normal form(2NF)
- Third normal form(3NF)

There are up to six normal forms but most databases strive to achieve third Normal form.

First Normal form

- Atomic values: each cell contains unique and single values => remove any collections or list of values.
- Be able to add data without altering tables
- Separate different relations into different tables
- Keep relationships together between tables together with foreign keys.

Second Normal form

- Must have reached 1NF
- All columns in the table must rely on the primary key => determine the primary key and ensure it is Unique, the primary key might change in the new table, but they should usually have a thing in common.

Third Normal form

- Must have reached 2NF
- No transitive dependencies
- transitive dependencies you are trying to attain is that to go from A -> C, you want to avoid going through B

Denormalization: to be done on heavy workloads to increase performance. It is also the process of trying to increase read performance rather than write performance by adding redundant copies of the data

Facts tables consist of measurements, metrics or facts of a business process.

Dimension, a structure that orders facts, measures in order to enable users to answer business questions. Dimensions are people, products, place and time.

Two of the most important schemas for data warehouses are Star schema and snowflake schema. Star schema is the simplest schema of data mart schema.
In a star schema, the fact table is in the middle, and it is surrounded by dimension tables. A snowflake schema is a local arrangement of the table in a multi-dimensional database represented by centralized fact tables which are connected to multiple dimensions. Snowflakes have multiple levels of relationships and multiple parents.

When a group of columns are the primary key, they are called composition key.

The Upsert and Do NOTHING clause is used when there is a need to add more details to an existing customer or overwrite some details.

Non-relational Databases

Apache Cassandra is a technology that was created at Facebook and became a high-level technology in 2010. In Apache Cassandra every node is connected to every other node, it is a peer-to-peer database architecture.
Eventual consistency occurs when data in some locations are inconsistent.

CAP theorem: says that it is impossible for a Datastore to provide two out of the three guarantees of Consistency, Availability and Partition tolerance.
Consistency => Every read from the database get the latest and correct piece of data or an error.
Availability => Every request is given and response is received without a guarantee that the data is the latest update.
Partition tolerance => The system continues to work regardless of losing network connectivity between nodes.
Cassandra and other NoSQL databases are optimized for high availability and partition tolerance(A, TP), consistency is not that much of a thing. When using Apache Cassandra always think Queries first, denormalization is a must, One table per query is a great strategy. AC(Apache Cassandra) does not allow for JOINS between tables.

In CQL(Cassandra Query Language), there are no JOINs, no GROUP BY and no subqueries. A guide to choosing the primary key and the composite key is to usually choose the first key that will be the filter.
When setting up the Primary key for Cassandra, the first key is the partition key(They manage how data is distributed between nodes), and the other keys are clustering keys.

The primary key must be Unique not necessarily the Partition Key. Clustering column will sort the data in ascending order.
In AC, failure to include a WHERE clause will result in an error, it is highly discouraged.