Data Engineering essentials using SQL, Python, PySpark

01/06/2023

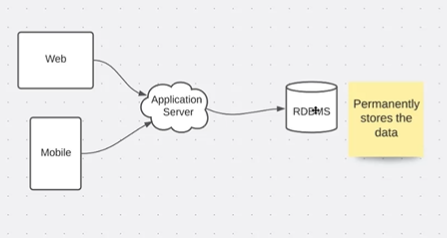
**Section 1: introduction**

11/06/2023

**Section 2: Getting Started with SQL for Data Engineering**

1. Setup postgres and pgAdmin (IDE)
2. To understand architecture of a Mobile application:

* Consider a retail application.
* Data have to be permanently stored in RDBMS
* The data will be stored in the form of tables in the database.
* Web and mobile will not interact with database directly, it will need application server
* Application server will take care of processing data and that will be rendered to the web, this is a typical mobile application architecture.



1. There are different types of databases available like Oracle , Mysql, Postgress, MS SQL server Sybase these are all RDBMS where as Snowflake, Databricks, Teradata, GCP BigQuery , AWS Redshift, Azure Synapse these are all data warehouse technologies.

12/06/2023

1. Purpose based database are nothing but databases which will be used for a specific purpose and development. There are different categories in purpose based databases.

* RDBMS
* Data warehouse or MPP
* NoSQL
* MongoDB
* Cassandra
* SoIR (Elastic search)
* Neo4j (Graph based technology)

Depending upon the specific requirement we use database.

* E.g. if we consider LinkedIn profile, for activity page we don’t need RDBMS, for that we can use NoSQL database but if we go to my premium, we see other details for these we can use RDBMS
* Purpose based databases is the process of micro service architecture
* Purpose based databases will be used as a down streaming data processing, particularly we use data warehouse for down streaming data processing.

1. Data warehouse & Data Lakes

* MPP (Massively parallel processing) these are also databases.
* Both data lake and data warehouse work together to deal with reports and dashboards of the users.
* Typically, we use BI tools for dashboards (Tableau/PowerBI)
* Data lake is nothing but a storage. (Cloud storage), based on the cloud platform relevant technology is used. AWS(S3), Azure (Azure storage), GCP
* Data lake is the cloud storage and data warehouse which provides computing capacity on top data lake.

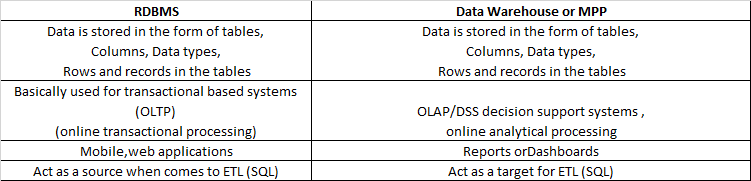
1. Usage of RDBMS and Data warehouse technologies

* E.g. buy a product from retail company
* Tracking the progress of the order
* Cancel the order and get the refund

For above operations we use RDBMS databases, it can be ORACLE, postgres etc.

* What is the profitability of the organization, what is the revenue trend over this year, which category is performing well, these are famous problem statements which can be considered for data lakes and data warehouse

1. Differences and similarities in RDBMS and Data Warehouse



* We typically use task related to RDBMS and data warehouse like, validation of the data, troubleshooting and bug solving, setting up databases for Development
* As application users we don’t use SQL directly, we translate the requests to SQL.