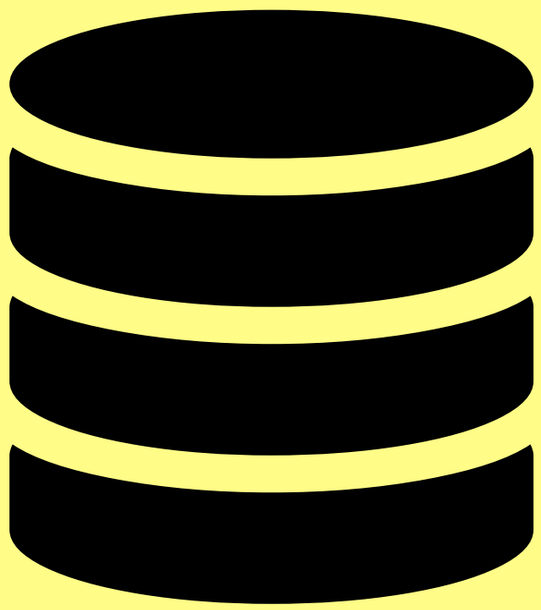
brijpandeyji

Data Storage Decoded

Demystifying the Differences Between a Database, Data Lake, Data Warehouse, Data Mart and Delta Lake



swipe >

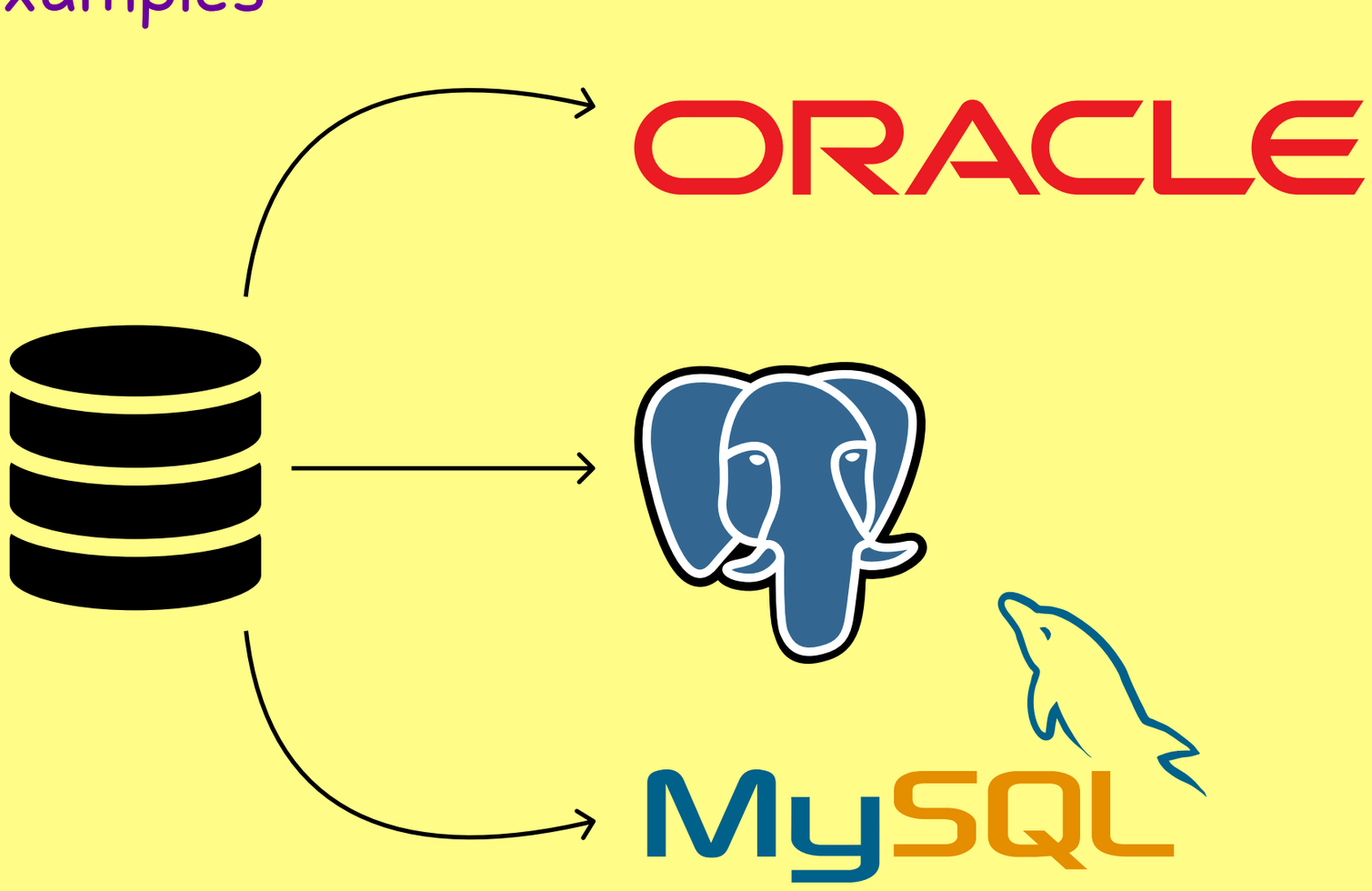
brijpandeyji

Database

A database is a structured collection of data that is stored in a specific format and can be queried and manipulated using a specific language, such as SQL.

Examples of databases include MySQL, Oracle, and PostgreSQL.

Here are some examples of a database:



swipe >

brijpandeyji

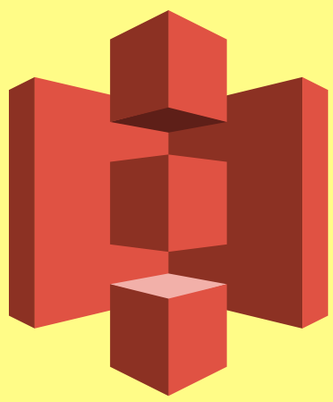
Data Lake

A data lake is a large, centralized repository that can store structured and unstructured data at any scale. The data is stored in its raw format and can be used for a variety of purposes, including batch and real-time processing, analytics, and machine learning.

Data lakes are often built on top of Hadoop or cloud platforms like AWS S3.

Here are some platforms where datalake can be built:

(HDFS - Hadoop Distributed File System)



(AWS S3)

swipe >

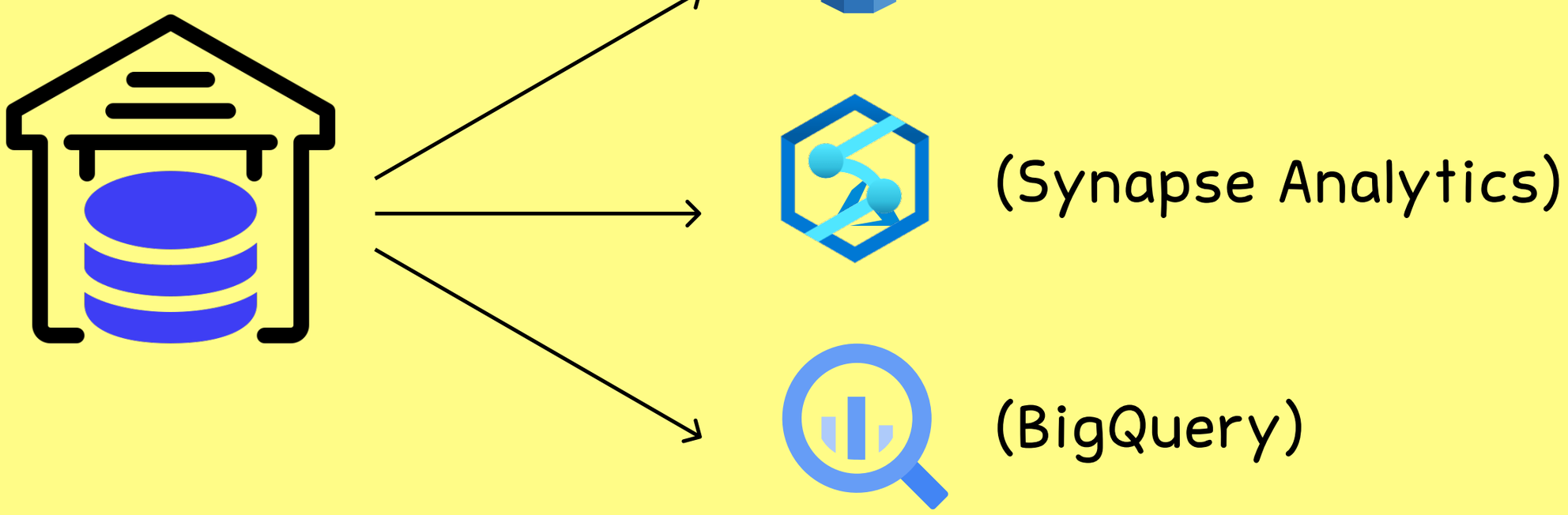
brijpandeyji

Data Warehouse

A data warehouse is a large, centralized repository that is optimized for reporting and analytics. It stores structured data that has been extracted, cleaned, and transformed from various sources. The data is stored in a specific format and can be queried and analyzed using tools like SQL and BI tools.

Examples of data warehouses include Amazon Redshift, Microsoft Azure Synapse Analytics, and Google BigQuery.

Here are some examples of a data warehouse:



swipe >

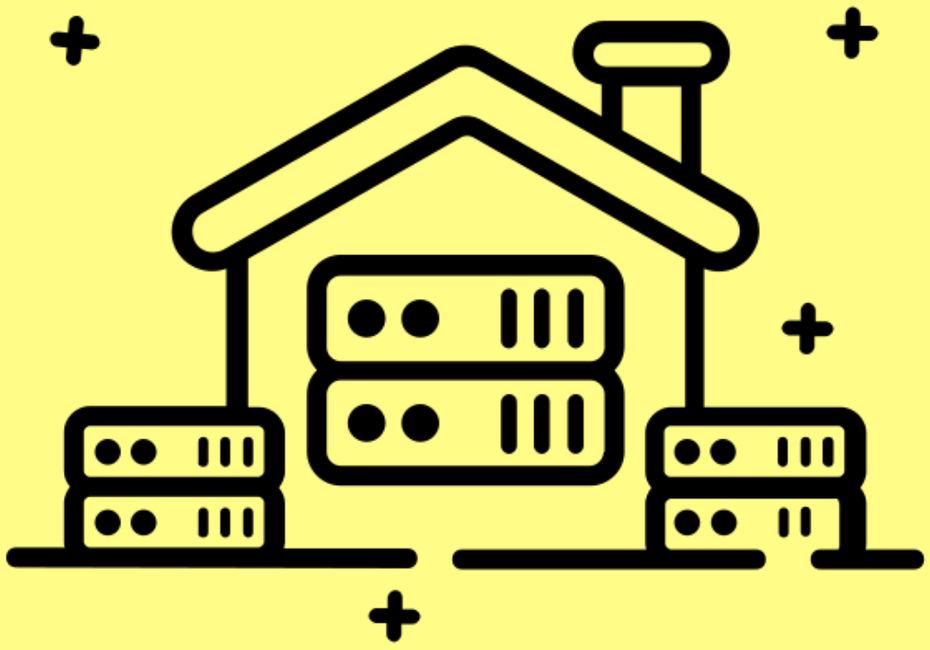
brijpandeyji

Data Mart

A data mart is a subset of a data warehouse that is focused on a specific business function or department. It contains a subset of data that is relevant to that specific area and is used to support decision making and reporting for that department.

For example, a data mart for the sales department might contain data on customers, products, and sales, while a data mart for the finance department might contain data on financial transactions and budgets.

Here is a visual representation of a data mart:



swipe >

Delta Lake

A delta lake is a storage layer that sits on top of data lakes, it allows for data to be stored in its original form, without the need to put it into a predefined structure, it also provide ACID transaction(Atomicity, Consistency, Isolation and Durability) for data lake ensuring data integrity and providing reliable data for analytics.

Here is a visual representation of a delta lake:

(Delta Lake)

swipe >

Summary

In summary, each of these systems serves a different purpose and has its own unique characteristics.

A database is a structured collection of data that is stored in a specific format and can be queried and manipulated using a specific language.

01

02

A data lake is a large, centralized repository that can store structured and unstructured data at any scale.

swipe >

brijpandeyji

Summary

03

A data warehouse is a large, centralized repository that is optimized for reporting and analytics.

A data mart is a subset of a data warehouse that is focused on a specific business function or department.

04

05

And finally, Delta Lake is a storage layer that sits on top of data lakes and provides ACID transactions for data lakes ensuring data integrity and providing reliable data for analytics.

swipe >

brijpandeyji

For More Interesting Content

Brij Kishore Pandey

Follow Me On LinkedIn

https://www.linkedin.com/in/brijpandeyji/