**Name : Kavya Gundala**

**Data Engineering**

**Data Engineer:**

* A data engineer makes design, build the platforms for collecting multiple data sources, processing and storing for analytical purpose.

**Data Scientist:**

* A data scientist analyses from complex data to create new data insights from raw data.

**Data Engineering:**

* Data Engineering is the process of Designing, building and scaling systems that organize data for further analytics .
* These systems provide to find practical and accurate applications of the data.

**Data Classification:**

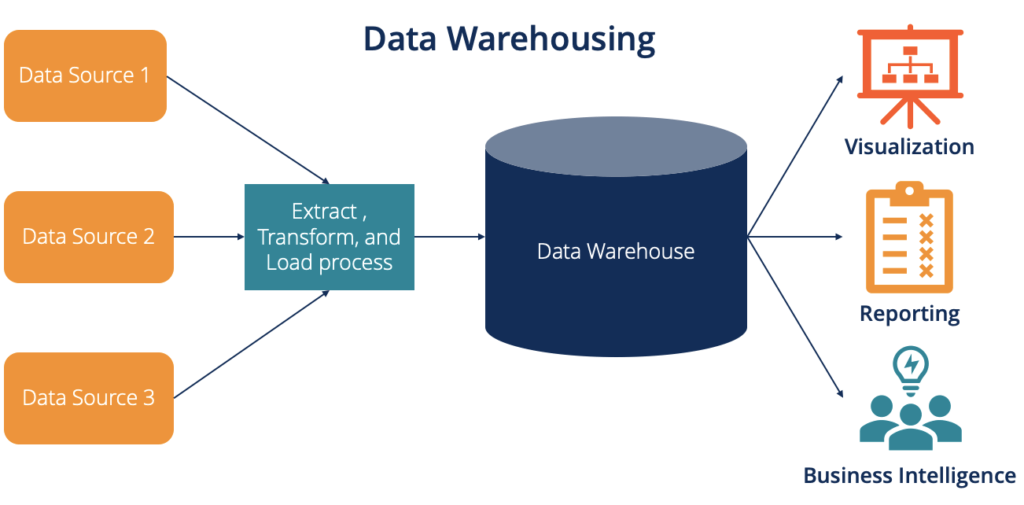
* **Raw Data:** This is the unprocessed data directly obtained from data source and no schema is applied.
* **Processed Data:** The raw data with schema is called Processed data. These data are stored in tables as relations.
* **Cooked Data:** This is the unique summarized data obtained from Processed data.

**Big Data Properties:**

* **Volume:** This describes how much data is stored .
* **Velocity:** This describes how fast the data is processing.
* **Variety:** This describes how different the data is.
* **Veracity:** This describes how consistent the data is.

**Data Warehouse:**

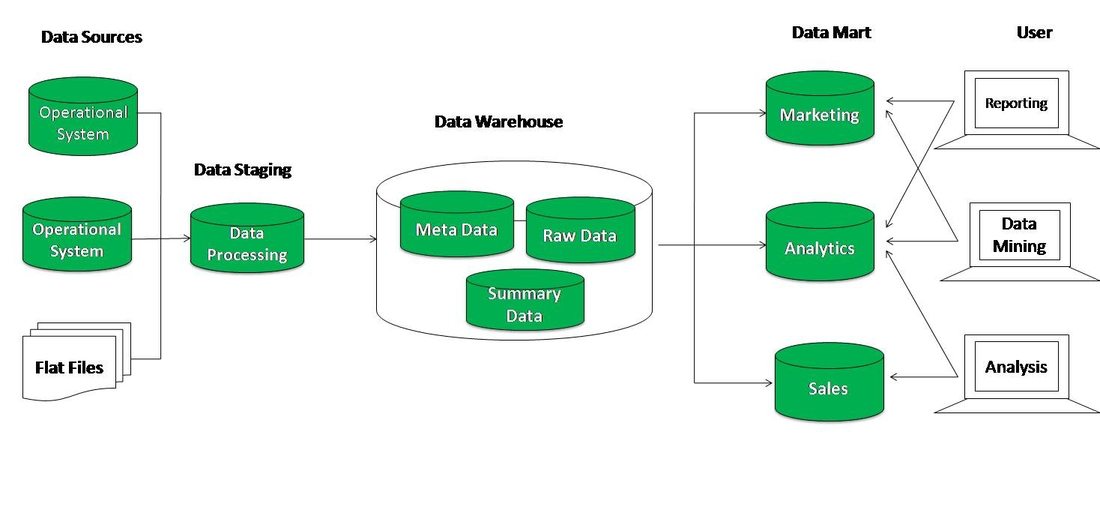
* Data Warehouse is a subject oriented, integrated, time-variant, non-volatile collection of structured and unstructured data.
* This data is helpful for decision making process to provide a coherent data from uncoherent, static data in business organizations.
* The Data Warehouse is a database used for data reporting and analysis.
* It is designed to support business intelligence (BI) and analytical reporting activities by providing a consolidated and optimized view of data.



**Purpose of Data Warehouse:**

* The purpose of Data Warehouse is to provide centralized repository for storing, organizing and analysing large volumes of data.
* It provides **Data Integration:** Data warehouses integrate data from multiple sources, such as operational databases, log files and flat files. This integration creates a unified view of the data.
* **Decision Support System**: Data warehouses provides to make decisions from data.
* **Data Quality and Consistency:** Data warehouses include processes for data cleansing and transformation, improving data quality and consistency.
* **Performance Optimization:** Data warehouses ensures that analytical queries can be executed efficiently, even on large datasets.

**Data Warehouse Architecture:**



**Data sources:**

**Operational database:** Operational database is accessed by Operational system to carry out regular operations of an organization.

**External source:** Data originates from various operational systems such as transactional databases, logs, and external data sources.

**Data Staging:**

**Data Processing:** Extracted data is temporarily stored in a staging area. This area allows for data validation, cleansing, and transformation before it is loaded into the data warehouse.

**Data Warehouse:**

* Data Warehouse is a subject oriented, integrated, time-variant, non-volatile collection of structured and unstructured data.
* This data is helpful for decision making process to provide a coherent data from uncoherent, static data in business organizations.
* The Data Warehouse is a database used for data reporting and analysis.

**Data Mart:**

* The data marts are subsets of data warehouse from a specific organization.
* The data in data warehouses is stored in the form of data marts.

**Operational Data Store:**

* Operational database is accessed by Operational system to carry out regular operations of an organization.
* Operational database uses OLTP(Online Transaction Processing) architecture for faster transaction processing.

**Benefits of OLTP:**

* Simplicity and Efficiency
* Data Integrity

**Pitfalls of OLTP:**

* Needs instant update.
* Data from OLTP is not suitable for data analysis.
* Even for simple transactions, need to query multiple tables using joins.

**OLTP vs Data Warehouse:**

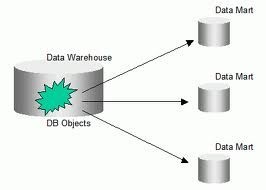
|  |  |
| --- | --- |
| **Data Warehouse** | **OLTP** |
| 1. It is a technique that collects data from different sources into central repository. | 1. It is a technique that is used for detailed day to day transaction data which keep chaining on everyday. |
| 1. It is designed for decision making process. | 2. It is designed for business transaction process. |
| 1. It stores large amount of data or historical data. | 3. It holds current data. |
| 1. It used for analyzing the business. | 4. In Online transaction processing, the size of data base is around 10MB-100GB. |
| 1. In Data warehousing, the size of database is around 100GB-2TB . | 5. In Online transaction processing, normalized data is present. |
| 1. It is subject-oriented. | 6. It is application-oriented. |
| 1. In Data warehousing, data redundancy is present. | 7. In Online transaction processing, there is no data redundancy. |

**Applications of OLTP:**

1. ATM machines: An example of an OLTP transaction would be at an ATM. In this scenario, two people share a joint bank account. They go to ATMs in separate locations at about the same time and try to withdraw all the money in the shared account. The OLTP system handles this transaction in real time. It allows the withdrawal from the ATM that finishes the authentication process first and then deals with, and ultimately denies, the second request.
2. Online bookings (ticketing, reservation systems, etc.)
3. Credit card payment processing (both online and in-store).

**Data Marts:**

* The data marts are subsets of data warehouse from a specific organization.
* The data in data warehouses is stored in the form of data marts.
* Data marts benefits to access data much more faster without having to go through complex data warehouse.



**Data Marts vs Data Warehouse:**

|  |  |  |
| --- | --- | --- |
|  | **Data Warehouse** | **Data Marts** |
| Definition | Data Warehouse consists of data from multiple heterogenous data and external files. | Data Marts are the subsets obtained from Data Warehouse. |
| Amount of data | vast data | Small amounts of data. |
| Lifetime | Long life | Short life in comparison. |
| Complexity of data | Data is contained in detailed form. | Data is contained in summarized form. |
| Size | Size is somewhere between 100GB and 1TB+. | Less than 100GB. |
| Access | Complicated process of creating schemas and views. | Easy design process of creating schemas and views. |