**Task 2**

Name : Ayesha Bibi

Date: 15/3/2023

1. **Data Marts:**

A data mart is a subset of a data warehouse focused on specific line of business, or built for specific group of users. Data marts make specific data available

**Need of Making Data Marts:**

The following are the main reasons for making data mart:

1. **Security:**

It secure data from unauthorized accesses. Because in data ware houses, we have the data of different departments of the organization so the chances of cracking data are more. But if have separate data marts for every department like we have HR data mart, for HR department, marketing data mart for marketing department, and restrict them to their marts only so in this we make our data more secure.

1. **Efficiency:**

Data marts increase the efficiency of accessing data. Like if we made data marts and want to access the data of specific department then we don’t need to search the entire data warehouse for specific department data we just use that specific data mart and perform the analysis on it which we want.

1. **Quicker access to insights:**

As teams identify and extract valuable data in a shorter space of time, the enterprise benefits from accelerated business processes and higher productivity. Like if company wants to check sales then they don’t search the whole data warehouse just use the Sales data mart to check the sales ratio in different time periods to get the insight that they want.

1. **Simpler data maintenance:**

Data marts focus on a single line, which leads to less clutter and easier maintenance.

**Types of data marts:**

There are three types of data marts that differ based on their relationship to the data warehouse and the respective data sources of each system.

1. **Dependent data marts**:

Dependent data marts use top-down approach. In which we first design data warehouse and then made data marts on top of it. Mostly large companies use dependent data marts because they want to fast down their process of querying.

1. **Independent data marts**:

Independent data marts act as a standalone system that doesn't rely on a data warehouse. Analysts can extract data on a particular subject or business process from internal or external data sources, process it, and then store it in a data mart repository until the team needs it. Then Data collected in data mart is used for designing the data warehouse down the line. Good for medium sized companies as entire data warehouse takes a lot of time which could be saved if data marts are created first.

1. **Hybrid data marts:**

Hybrid data marts combine data from existing data warehouses and other operational sources. This approach is Important if the new line of business arises then for speeding up the process instead of loading that data into data warehouse, we make a separate data mart of it and load that data into it.

1. **Data Mesh:**

Data mesh is a segregation of data-by-data domains and they are hosted and served as separate data domains and they are also owned by different teams. The ownership lies to different peoples and every domain dataset act as a product.

In the data mesh paradigm, domains host and serve their data products for easy consumption, rather than dataflowing through an ETL-type process to some centralized data warehouse.

**Why Need Data Mesh:**

Operational systems change or develop over time. At the same time, the pipelines implemented in the ETL (or ELT) value chain are strongly coupled steps that support analytical business needs. So, when the operational systems modified in order to propagate the changing needs. So, we need to made changes throughout all steps of the processes in which pipelines involved. Some automation can certainly be implemented, but this is one of the (main) reasons why traditional solutions have a tendency to become inflexible and time consuming to maintain in the face of an increasing pace of change in business.

1. **Data Lake:**

A data lake is a centralized repository for storing, and processing large amounts of data in its raw form, including structured, unstructured, and semi-structured data at any scale.

**Why do you need a data lake?**

We need data lake for storing all type of data like log files, data from click-streams, social media, and internet connected devices. Then perform different analytics on that data to generate business value from data. This helped organizations to identify and act upon opportunities for business growth faster by attracting and retaining customers, boosting productivity, proactively maintaining devices, and making informed decisions.

**Benefits:**

1. Increase operational efficiencies
2. A data lake makes it easy to store, and run analytics on machine-generated IoT data to discover ways to reduce operational costs, and increase quality.
3. It Helps for fast ingestion of new data.

**The Challenges of Data Lakes:**

The main challenge with a data lake architecture is that raw data is stored with no oversight of the contents. For a data lake to make data usable, it needs to have defined mechanisms to catalog, and secure data. Without these elements, data cannot be found, or trusted resulting in a “data swamp." Meeting the needs of wider audiences require data lakes to have governance, semantic consistency, and access controls.

1. **OLTP VS OLAP:**

|  |  |
| --- | --- |
| OLAP | OLTP |
| Provides historical data for reporting and planning | Manages day to day operations |
| Use large and complex queries like aggregation | Use simple and standard queries |
| Use select query for fetching the data | Use Insert, Update and Delete statement |
| Users are mostly knowledgeable workers like business analytics, etc. | Used by end users like clerks, cashiers and DBA are mostly knowledgeable worker, etc. |
| In this data is denormalized to improve query performance | In this data is stored in third Normal form. |
| Requires significantly large space | Requires relatively smaller space |
| Backups are rarely needed | Data is backed up regularly |
| In this design changes according to reporting subjects. Snowflake model-based design. | Design changes as per industry requirements. ER model-based design. |

1. **DWH vs Data Lake:**

|  |  |  |
| --- | --- | --- |
| Parameters | Data Lakes | Data Warehouse |
| Data | Data lakes store everything. | Data Warehouse focuses only on Business Processes. |
| Processing | Data are mainly unprocessed | Highly processed data. |
| Type of Data | It can be Unstructured, semi-structured and structured. | It is mostly in tabular form & structure. |
| Agility | Highly agile, configure and reconfigure as needed. | Compare to Data lake it is less agile and has fixed configuration. |
| Storage | Data lakes design for low-cost storage. | Expensive storage that gives fast response times are used |
| Security | Offers lesser control. | Allows better control of the data. |
| Replacement of EDW | Data lake can be source for EDW | Complementary to EDW (not replacement) |
| Schema | Schema on reading (no predefined schemas) | Schema on write (predefined schemas) |
| Data Processing | Helps for fast ingestion of new data. | Time-consuming to introduce new content. |
| Data Granularity | Data at a low level of detail or granularity. | Data at the summary or aggregated level of detail. |
| Tools | Can use open source/tools like Hadoop/ Map Reduce | Mostly commercial tools. |

**Task 3**

**Can a Database be used as DWH?**

Yes, Database can be used as DWH if we structured it to do analytics not just transactions because a data warehouse is a type of database that integrates copies of transaction data from disparate source systems and provisions them for analytical use.

**Major Dereferences between Structured and Unstructured data:**

Structured data is highly specific and is stored in a predefined format in the form of table having rows and columns, where unstructured data is a conglomeration of many varied types of data that are stored in their native formats e.g., images, audio, video, etc. This means that structured data takes advantage of schema-on-write and unstructured data employs schema-on-read.

**Duties of Data Engineer:**

Here is the list of roles and responsibilities, Data Engineers are expected to perform:

**1. Work on Data Architecture**

**2. Collect Data**

**3. Conduct Research**

**4. Create Models and Identify Patterns**

**5. Automate Tasks**