Data Repositories

Organizations today generate enormous amounts of data from their everyday activities, like student registrations, financial transactions, hospital visits, supermarket purchases, mobile apps, websites, sensors, CCTV cameras, and social media. Because this data is so valuable, it must be stored in places that make it easy to use, protect, analyze, and understand. These storage places are called data repositories. Different types of repositories serve different purposes, depending on the nature of the data and are discussed below.

a) Database

Commonly, a database stores data in an organized, structured format, usually as tables with rows and columns. It is designed to handle day-to-day operations where data must be added, updated, or retrieved quickly and accurately. For instance, when a student registers for a course, when a bank processes a withdrawal, or when a retail shop updates its stock, the data goes into a database. Databases support fast transactions and are the backbone of operational systems. They give organizations a reliable and consistent way to manage current, real-time data.

b) Data warehouse

While a database focuses on today's activities, a data warehouse focuses on the long-term history of the organization. It stores large amounts of cleaned, well-organized data collected from various departments over many months or years. A data warehouse is not meant for frequent updates; instead, it is used for analysis, reporting, and decision-making. It helps organizations understand trends, patterns, and performance over time. For instance, a data warehouse can answer questions like: How have our sales changed over the past five years? or Which faculties have the highest graduation rates? Data warehouses power dashboards and business intelligence tools used by managers and executives.

c) Data lake

Unlike databases and data warehouses, which require information to be structured and organized, a data lake can store any type of data in its raw state. This includes tables, text files, scanned documents, videos, audio files, social media posts, emails, web logs, app logs, and sensor readings. Nothing has to be formatted first; everything is stored as it comes. This makes data lakes extremely useful for advanced data science, machine learning, IoT analytics, and big data exploration. Because a data lake does not force structure on the data, it can handle massive, diverse data from many sources at once.

d) Data mart

Instead of storing the entire organization's data, a data mart contains information for a single department such as finance, admissions, sales, or human resources. This makes access faster and more focused. E.g., the finance department may only need budget, expense, and revenue data, not student records or staff profiles. Data marts help departments retrieve exactly what they need without searching through large, organization-wide systems.

Roles of Data Professionals

This section looks into major data roles and the specific storage systems each one works with.

a) The Data Analyst

A data analyst's job is to look at data and explain what is happening. They answer questions like: How many students registered this semester? How many patients visited today? What were last month's sales? Because they need neat, organized information, they mostly use databases for current data (today's activities) and Data warehouses for older, cleaned-up data from previous months or years. They turn this data into charts, tables, and simple summaries. They do not usually work with raw or messy data.

b) The Data Engineer

A data engineer is like the builder and fixer of the data system. They make sure data flows smoothly from one place to another. They work with all data repositories. They create the pipes that move data automatically every minute or every day. Without them, analysts and scientists would have nothing to work with.

c) The Database Administrator (DBA)

A DBA takes care of databases only. They make sure the database is fast to access, secured, backed up and always available. They do not work with warehouses or lakes. They focus on the central system where daily transactions happen.

d) Data Scientist

A Data Scientist is a professional who uses data to discover patterns, answer important questions, and make predictions. Their main focus is to understand the data deeply and uncover the hidden meaning inside it. They spend much of their time exploring data, cleaning it, arranging it, and testing different ideas to see what insights it can reveal.

A data scientist often works with large collections of information that may come from different places such as a database, a data warehouse, or a data lake, because they need a mixture of tidy data and raw data to experiment with. Using mathematics, statistics, and computer programs, they try to find relationships in the data, such as why something is happening or what is likely to happen next.

e) Machine Learning Engineer

A Machine Learning Engineer is the professional who takes the models created by the data scientist and makes them work in real life. Their job is more practical and software-focused. If the data scientist builds the model in a notebook or test environment, the ML engineer builds the system that uses the model. They design how the model will connect to a website, an app, a school portal, a hospital system, or a banking platform. They make sure the model runs quickly, produces accurate results, and can handle many users at once.

Machine learning engineers also monitor the model to make sure it continues to perform well. If new data arrives or user behavior changes, they update the model or rebuild parts of the system so it stays accurate and reliable. They use storage systems such as data lakes or lakehouses for training large models, and they use databases when the model needs to make real-time predictions.