**Databricks for Data Science**

In previous eras companies primarily used structured data from operational systems for their software products and decision-making processes. Now today's products often use AI technologies such as computer vision, speech models, and text mining. This shift demands a new approach to data management (DM) systems. It emphasizes scalable cloud-based platforms to handle all organizational data efficiently.

**Evolution of Data Management:**

**A) Relational Databases (For early Days):**

* Relational databases were the traditional method for data collection, storage, and analysis. By using Structured Query Language (SQL) it’s flexible to manage highly structured data.
* These databases were simple and reliable but led to **data silos**, where different parts of an organization stored data in fragmented systems.

**B) Data Warehouses:**

* To integrate these silos, data warehouses were developed to unite different databases. It allowed the data for business intelligence (BI) and reporting.
* Some limitations included high maintenance costs, lack of support for machine learning (ML) use cases, and issues with scalability and flexibility.

**C) Data Lakes:**

* As data volumes grew and the need for managing unstructured data became crucial, **Data Lakes** came into picture. Here we could store raw data in various formats.
* **Apache Hadoop** and **Spark** were key technologies in this space, with Spark offering significant speed improvements and a unified analytics engine for large-scale data processing, SQL analytics, and ML.

**D) Delta Lake:**

* An open-source storage layer that enhances **Data Lakes** by offering **ACID transactions**, scalable metadata handling, and unified streaming and batch data processing with the help of Spark, the processing technology for data reliability and scalability.
* Delta Lake on Databricks enables efficient data pipelines, reducing computer times and costs with scalable cloud runtime and optimized Spark clusters.

**Modern Data Management with Lakehouse:**

**A) Challenges of Data Lakes:**

* Data lakes faced issues like difficulty in appending and modifying data, job failures, handling large metadata, and poor performance.
* These challenges led to the development of the Lakehouse architecture, which combines the benefits of data warehouses and data lakes.

**B) Lakehouse Architecture:**

* Lakehouse is a new DM architecture. It supports BI, SQL analytics, data science, and ML on a single platform.
* It uses low-cost object storage for data lakes and provides features like transactions, data quality enforcement, and consistency, making it suitable for various data applications.

**C) Advantages of Lakehouse:**

* **Unified Data Teams:** Brings together data engineers, scientists, and analysts on one platform.
* **Breaking Data Silos:** Centralizes all data, making it accessible and manageable.
* **Real-time Data Processing:** Supports batch and streaming data, keeping data up to date.
* **Reduced Vendor Lock-in:** Uses open formats, allowing easy data transfer between different technologies.

**Data Bricks:**

It is a unified **Data Analytics Platform**.

* It integrates with various tools (**Spark, Python, R, ML libraries**) and provides a platform for both BI and non-BI workloads.
* It simplifies the ML lifecycle with **MLflow**, an open-source platform for managing the complete ML process, ensuring enterprise reliability and security.

The **Lakehouse** approach, powered by **Databricks**, provides a modern, scalable, and efficient solution for data management. It supports diverse use cases from BI to advanced ML. This approach reduces costs, improves security, and unifies data teams, making it an essential tool for today's data-driven organizations.