# Databricks Certified Data Engineer Associate

# Bigdata History

# Bigdata History

- 2004: Google published a paper on the Google File System, scalable & fault-tolerant distributed file system for large amounts of data.
- 2008: Hadoop, an open-source software framework for distributed storage and processing of big data, was released by Apache.
- 2009: Matei Zaharia started developing Spark as a research project at UC Berkeley's AMPLab.
- 2010: Spark was open-sourced under a BSD license.
- 2012: Spark became an Apache Software Foundation incubator project.
- 2013: Databricks is founded by the creators of Apache Spark to provide a cloud-based platform for big data processing and analytics.
- 2014: Delta Lake, an open-source storage layer that provides ACID transactions on top of data lakes, is introduced by Databricks.
- 2018: Databricks introduces Delta Engine, a high-performance query engine for Delta Lake, which allows queries to run up to 20x faster than Apache Spark on Delta Lake tables.
- 2020: Databricks introduces the concept of a "lakehouse", which combines the best aspects of data warehouses and data lakes, allowing users to have ACID transactions on large datasets, fast analytics, and data science capabilities in one platform.
- 2021: Databricks announces a $1 billion funding round, valuing the company at $28 billion

# Introduction to Databricks

# Introduction to Databricks

- Databricks is a cloud-based data engineering, data science, and machine learning platform built on top of Apache Spark.
- It was founded by the original creators of Spark in 2013 and is now used by thousands of companies, including Fortune 500 companies and startups.
- Provides a collaborative workspace where data engineers, data scientists, and machine learning engineers can work together to build data pipelines, run analytics, and build machine learning models
- tools for data ingestion, data processing, data analysis, and machine learning, as well as a suite of collaboration tools for sharing code, notebooks, and dashboards.
- One of the key benefits of Databricks is that it abstracts away many of the complexities of managing and scaling Spark clusters, allowing users to focus on building data pipelines and models.

# Data lake

- A data lake is a centralized repository that allows organizations to store all their structured and unstructured data at any scale.
- Built using Hadoop or cloud-based services like AWS S3, and provide a low-cost way to store large amounts of data in various formats.

# Delta Lake

- Delta Lake is an open-source storage layer that provides ACID transactions on top of data lakes.
- Delta Lake addresses some of the challenges of traditional data lakes by adding reliability, scalability, and performance.
- Delta Lake provides features like schema enforcement, versioning, data indexing, and ACID transactions, which makes it easier to manage large data sets
- Deltalake enables data engineers and data scientists to work collaboratively

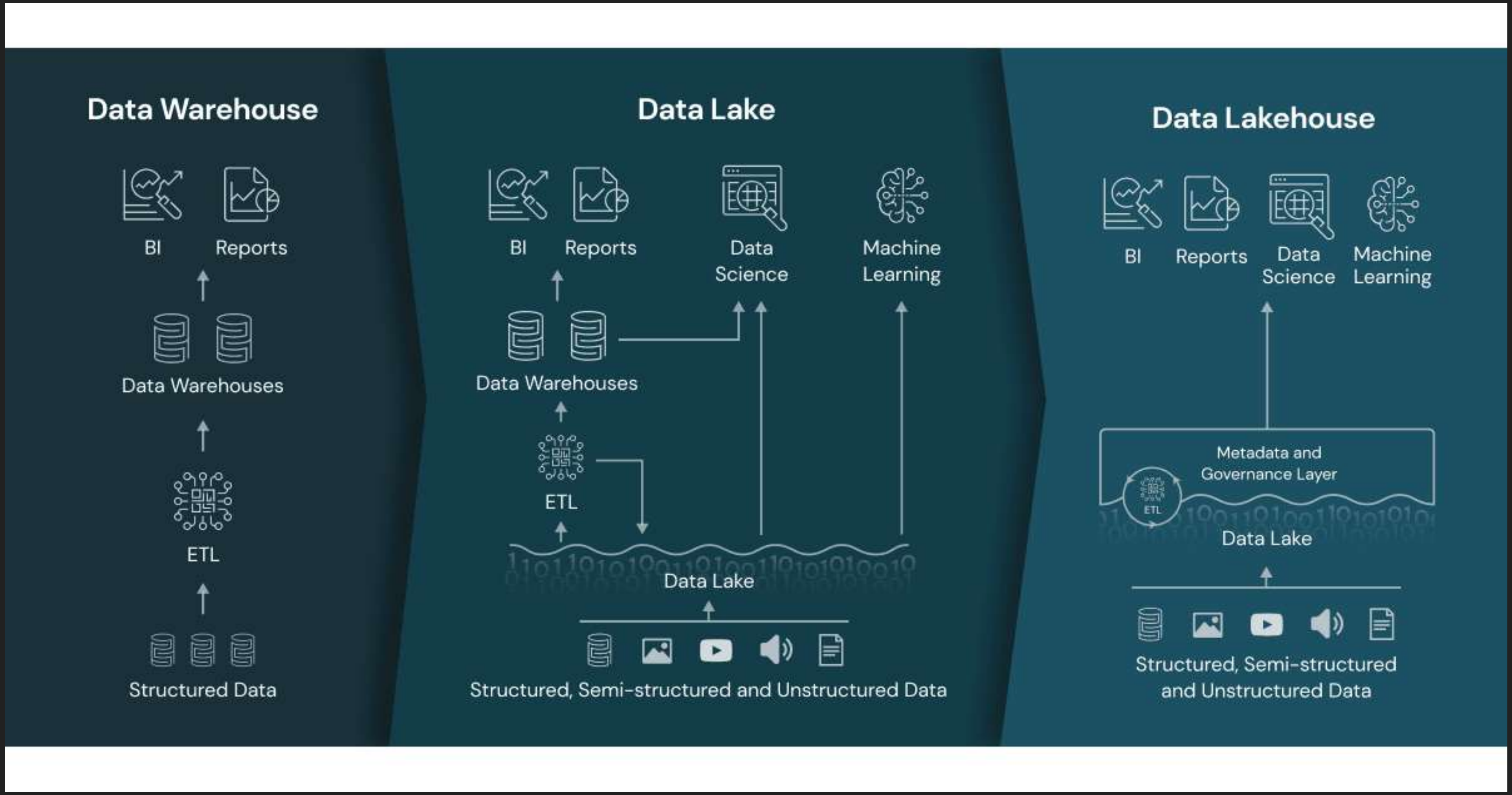
# Data Warehouse

- A data warehouse is a large, centralized repository of data that is used for reporting and analysis
- It is designed to support the storage and querying of large volumes of structured, semi-structured, and unstructured data from a variety of sources
- Data warehouses typically use a relational database management system (RDBMS) to store data
- It is optimized for query performance, scalability, and reliability
- They are often used in business intelligence (BI) applications to provide users with insights and trends from large datasets

# Lakehouse

- Lakehouse is a new concept introduced by Databricks in 2020, which combines the best features of
  - data warehouses
  - data lakes
- A lakehouse architecture is built on top of Delta Lake and allows users to have
  - ACID transactions on large datasets
  - fast analytics
  - data science capabilities in one platform

# What is a Data Lakehouse?



Credit: Databricks

# Key Technology Enabling the Data Lakehouse

- ➢ **There are a few key technology advancements that have enabled the data lakehouse:**

- ✓ Metadata layers for data lakes

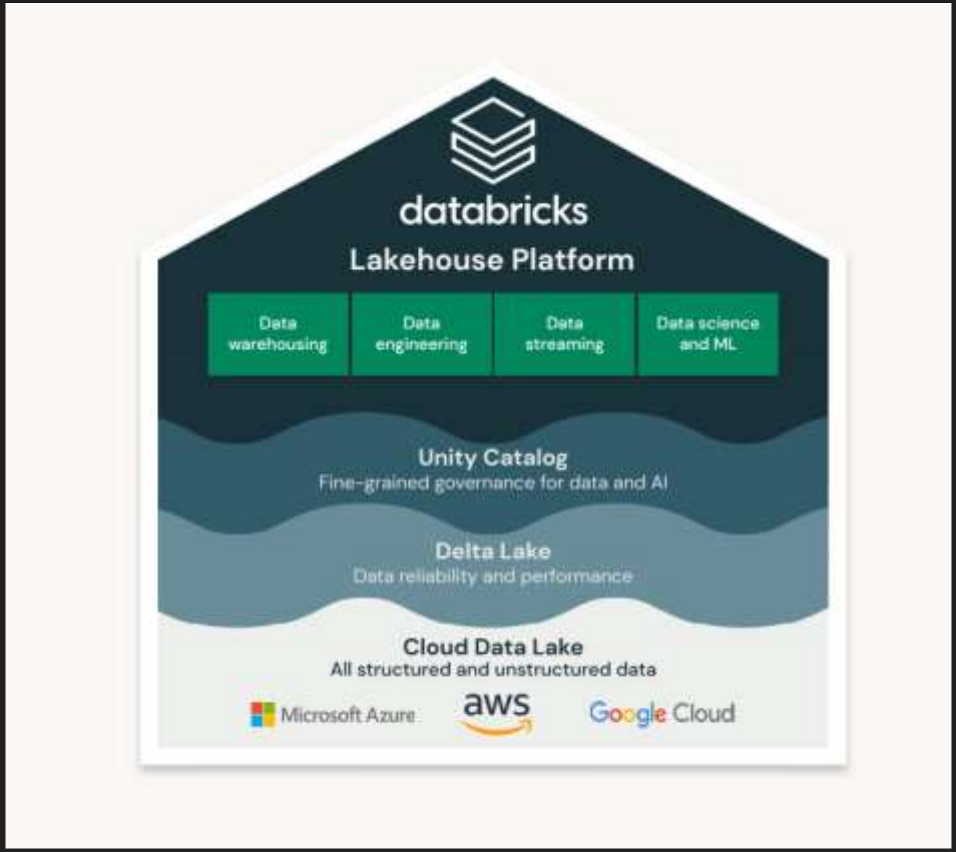
- ✓ New query engine designs providing high-performance SQL execution on data lakes

- ✓ Optimized access for data science and machine learning tools

# Data lakes, Data Lakehouse & Data Warehouses

| | Data lake | Data lakehouse | Data warehouse |
|---|---|---|---|
| **Types of data** | All types: Structure, semi-structured, unstructured (raw) | All types: Structure, semi-structured, unstructured (raw) | Structured data only |
| **Cost** | $ | $ | $$$ |
| **Format** | Open format | Open format | Closed, proprietary format |
| **Scalability** | Scales to hold any amount of data at low cost, regardless of type | Scales to hold any amount of data at low cost, regardless of type | Scaling up becomes exponentially more expensive due to vendor costs |
| **Intended users** | Limited: Data scientists | Unified: Data analysts, data scientists, machine learning engineers | Limited: Data analysts |
| **Reliability** | Low quality, data swamp | High quality, reliable data | High quality, reliable data |
| **Ease of use** | Difficult: Exploring large amounts of raw data can be difficult without tools to organize and catalog the data | Simple: Provides simplicity and structure of a data warehouse with the broader use cases of a data lake | Simple: Structure of a data warehouse enables users to quickly and easily access data for reporting and analytics |
| **Performance** | Poor | High | High |

# Databricks Lakehouse Platform

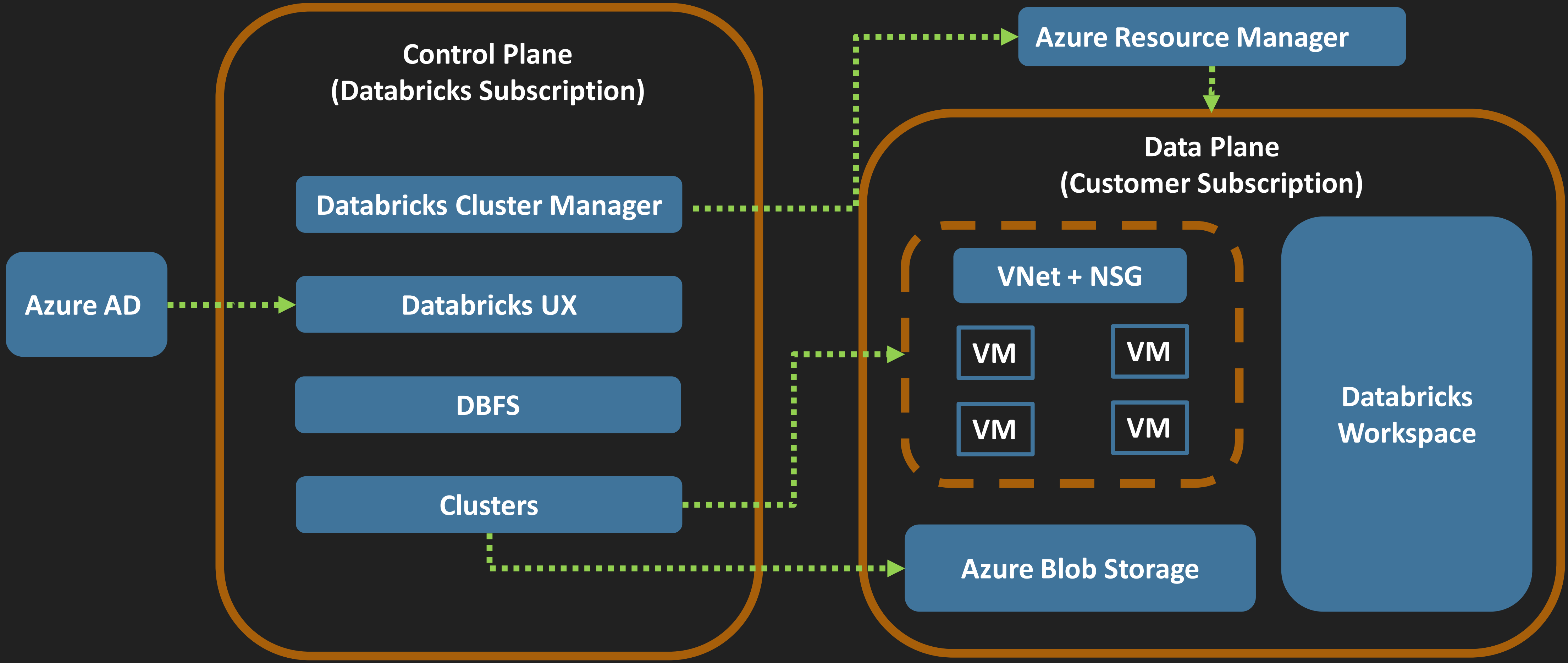


Credit: Databricks

Databricks provides a wide range of integrations with other popular data tools and platforms, such as AWS, Azure, and Google Cloud Platform, making it easy to integrate into existing data ecosystems

# Azure Databricks Architecture Overview

# Azure Databricks Architecture



# THANK YOU