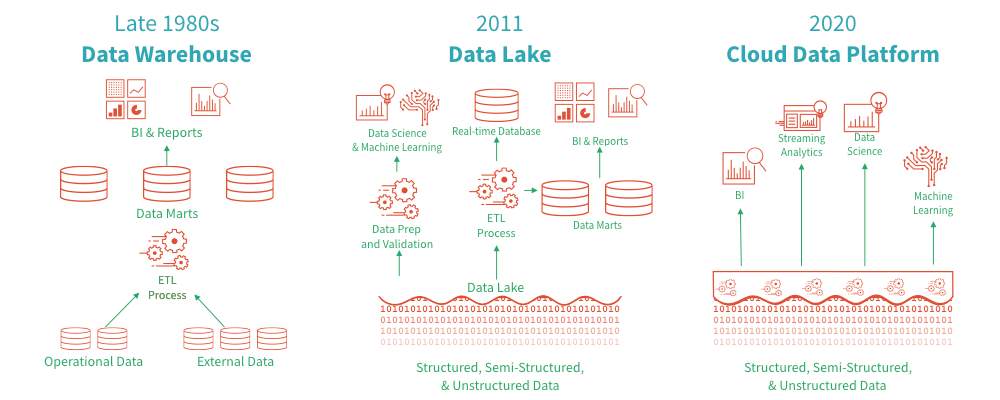
Over the past few years at Databricks, we’ve seen a new data management paradigm that emerged independently across many customers and use cases: the **cloud data platform**.



The evolution of data management.

**Data Warehouses**

Data warehouses have a long history in decision support and business intelligence applications.

Since its inception in the late 1980s, data warehouse technology continued to evolve. Massively parallel processing (MPP) architectures led to systems able to handle larger data sizes.

While warehouses excel in handling structured data, most enterprises have to deal with unstructured, semi-structured, and data with high variety, velocity, and volume. Data warehouses are not suited for many of these use cases, and they are certainly not the most cost-efficient.

**Data Lakes**

As companies began to collect large amounts of data from many different sources, architects began envisioning a single system to house data for many different analytic products and workloads.

About a decade ago companies began building **data lakes** – repositories for raw data in a variety of formats.

While suitable for storing data, data lakes lack some critical features. Data lakes do not support ACID transactions, do not enforce data quality, and their lack of consistency/isolation makes it almost impossible to mix appends and reads, and batch and stream jobs.

**Diverse Data Applications**

Companies require systems for diverse data applications including:

* SQL analytics
* real-time monitoring
* data science
* machine learning and artificial intelligence

Most of the recent advances in the modern data landscape have been building better models to process unstructured data (text, images, video, audio). These are precisely the types of data for which a data warehouse is sub-optimal.

**A Multitude of Systems**

A common approach to building these systems for diverse data applications is to use multiple systems:

* a data lake
* several data warehouses
* streaming systems
* time-series, graph, and image databases

Having a multitude of systems introduces complexity, and more importantly, introduces delay as data professionals invariably need to move or copy data between different systems.

**What is a cloud data platform?**

Recent innovations in data system design make the cloud data platform design pattern possible. This design is to implement data warehouse data structures and data management features directly on the kind of low-cost storage used for data lakes.

*A data lakehouse is what you would get if you redesigned data warehouses in the modern world, now that cheap and highly reliable storage (in the form of object stores) are available.*

**A data lakehouse has the following key features:**

* storage is decoupled from compute
* open storage formats, tools, and processing engines
* support for diverse data types ranging from unstructured to structured data
* support for diverse workloads, including SQL and analytics, data science, and machine learning
* transaction support
* end-to-end streaming