Databricks develops and sells a cloud data platform based on a modern lakehouse data architecture. Databricks’ lakehouse is based on the open source Apache Spark framework that allows users to run analytical queries against semi-structured, schemaless data.

schema less means the database don't have fixed data structure, such as MongoDB

Apache Spark is an open-source parallel processing framework that supports in-memory processing to boost the performance of applications that analyze big data. Big data solutions are designed to handle data that is too large or complex for traditional databases. Spark processes large amounts of data in memory, which is much faster than disk-based alternatives.

A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions.

Delta lake: Delta Lake is an open-source storage layer that brings ACID (atomicity, consistency, isolation, and durability) transactions to Apache Spark and big data workloads.

Delta Lake uses Delta Engine for optimizing transactions on Data Lake. Delta Engine is highly compatible with Apache Spark and Spark APIs. And Databricks provides Delta Engine optimizations without additional costs as a built in feature. So it is recommended to use Delta Engine optimization features when processing data especially in Databricks or Apache Spark.

Data Live Tables:

As the amount of data. Data sources and data types increase in. the organization, data engineering teams face the pressure to transform raw, messy data into clean reliable data for data analytics data science and ML. In order to achieve this Data engineers, build complex, production grade pipelines which contain multiple tables and dependencies. Once these pipelines are built the data engineers should continuously monitor the data movement and if there are any errors, they should be able to rectify these errors. And again, if the new requirements come up this entire process should be repeated. Here rather than working on data the data engineers are working on the tools so the operational overhead will be increased.

**More easily build and maintain data pipelines:**

With Delta live table once you define the pipeline with the data source, transformation logic and the destination state they will handle everything by this way the operational overhead will be eliminated.

**Automatic testing:**

Prevent bad data from flowing into tables through validation and integrity checks and avoid data quality errors with predefined error policies (fail, drop, alert or quarantine data). In addition, you can monitor data quality trends over time to get insight into how your data is evolving and where changes may be necessary.

**Deep visibility for monitoring and easy recovery:**

Reduce downtime with automatic error handling and easy replay. Speed up maintenance with single-click deployment and upgrades.

**What is Auto Loader**

Auto Loader incrementally and efficiently processes new data files as they arrive in cloud storage without any additional setup.

**How does Auto Loader work?**

Auto Loader incrementally and efficiently processes new data files as they arrive in cloud storage. Auto Loader can load data files from AWS S3 (s3://), Azure Data Lake Storage Gen2 (ADLS Gen2, abfss://), Google Cloud Storage (GCS, gs://), Azure Blob Storage (wasbs://), ADLS Gen1 (adl://), and Databricks File System (DBFS, dbfs:/). Auto Loader can ingest JSON, CSV, PARQUET, AVRO, ORC, TEXT, and BINARYFILE file formats.

Auto Loader provides a Structured Streaming source called cloudFiles. Given an input directory path on the cloud file storage, the cloudFiles source automatically processes new files as they arrive, with the option of also processing existing files in that directory.

Structured Streaming is a fast, scalable, fault-tolerant, end-to-end, exactly-once stream processing API that helps users in building streaming applications without having to reason about it.

**Benefits of Auto Loader over using Structured Streaming directly on files:**

Scalability

Performance

Cost effective

Photon runtime:

Photon is part of a high-performance runtime that runs your existing SQL and DataFrame API calls faster and reduces your total cost per workload. Photon is used by default in Databricks SQL warehouses. Azure Databricks clusters Photon is available for clusters running Databricks Runtime 9.1 LTS and above.