In-person Databricks training

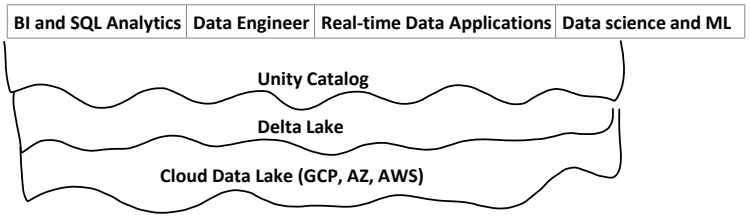
Tuesday, 30 August 2022 8:54 AM

ML pipeline? Retrain?
BI-SHAP explainer

Day 1 (30 Aug 2022): SQL/ Datawarehouse (Yasamin Mokri (Trainer), Wibur Tong (CS))

1. **Introduction**

- Architect:



- Lakehouse: Delta lake with fine grained governance and security with Unity Catalog (Files, Blobs, Table ACLs) .
- Unity Catalog:
  - Data assets: Warehouses, Tables, Cols, DataLake, Files, ML/Models, Dashboards/ Notebooks
  - Data Lineage
  - Attribute-based access control
  - Security policies
  - Table and column level tags
  - Auditing
  - Data sharing. For eg: BI app can access the data
- SQL workloads: Native SQL interface, BI tools supports, can also do BI dashboard within the platform.
- Delta Lake optimizes performance, using Spark APIs -> .format("delta").
- You can do Time Travel with Delta Lake

```
%sql
VERSION AS OF version
#Rollback
RESTORE db TIMESTAMP AS OF timestamp
```

- Auto Optimize And Auto Compact properties:
  - Set delta properties .autoOptimize.optimizeWrite=True;
  - autoOptmize.autoCompact=true;
- Can Enable CDF (Delta Change Data Feed) for tables:

```
%sqp
ALTER TABLE
SET TBLPROPERTIES (delta.enableChangeDataCapture=True);
```

  Change Data Capture captures row-level changes from any data sources DBR, Cloud Storage, or DBFS; can handle out-of-order events

- Delta Live Table: CREATE INCREMENTAL LIVE_TABLE table AS SELECT *FROM cloud_File
- Data Quality validation and monitoring with DLT by defining `EXPECT`. If `ON VIOLATION`, can fail, drop, alart, quarantine th e data pipeline.
- Databricks table name convention:
  - Bronze is raw ingestion and history (should be Delta format)
  - Silver is the filtered and cleaned data, but still fine-grained -> DS can work on this data
  - Gold is business-level aggregates, ready for BI & Streaming Analytics
- Databricks lakehouse is 12x better in terms of performance cost than cloud DWH (Snowflake), and with high concurrency - 3x QpH performance

2. **Data Engineering lab**

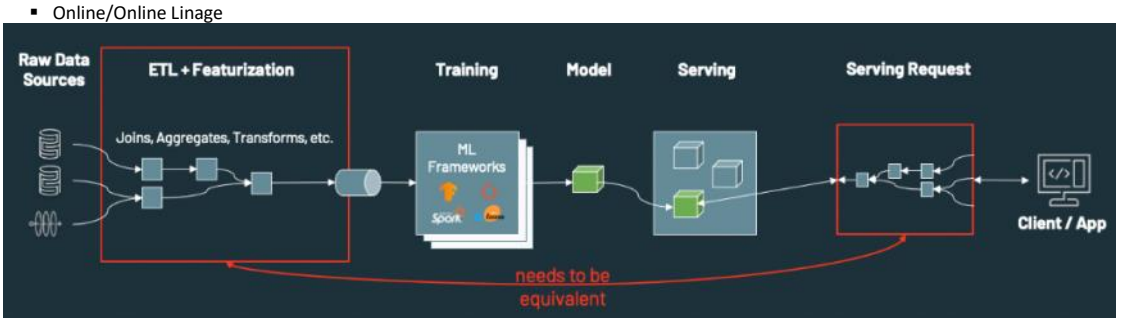
You can create a copy of an existing Delta table at a specific version using the clone command. Clones can be either deep or shallow.

- A deep clone is a clone that copies the source table data to the clone target in addition to the metadata of the existing table.
- A shallow clone is a clone that does not copy the data files to the clone target. The table metadata is equivalent to the source. These clones are cheaper to create, but they will break if original data files were not available

Git repo: https://github.com/zivilenorkunaite/apjbootcamp2022

Day 2 (31 Aug 2022): Machine Learning

- Pre-packaged libs on ML cluster: Tf, Keras, XGBoost, Spark, Sklear, Koalas, Mfflow, Mpleap, Hyperopt
- Feature Store:
  - Discoverability and Reusability
  - Versioning
  - Online/Online Linage



- Model Serving:

Model Serving and Model Registry