Delta lake CDF

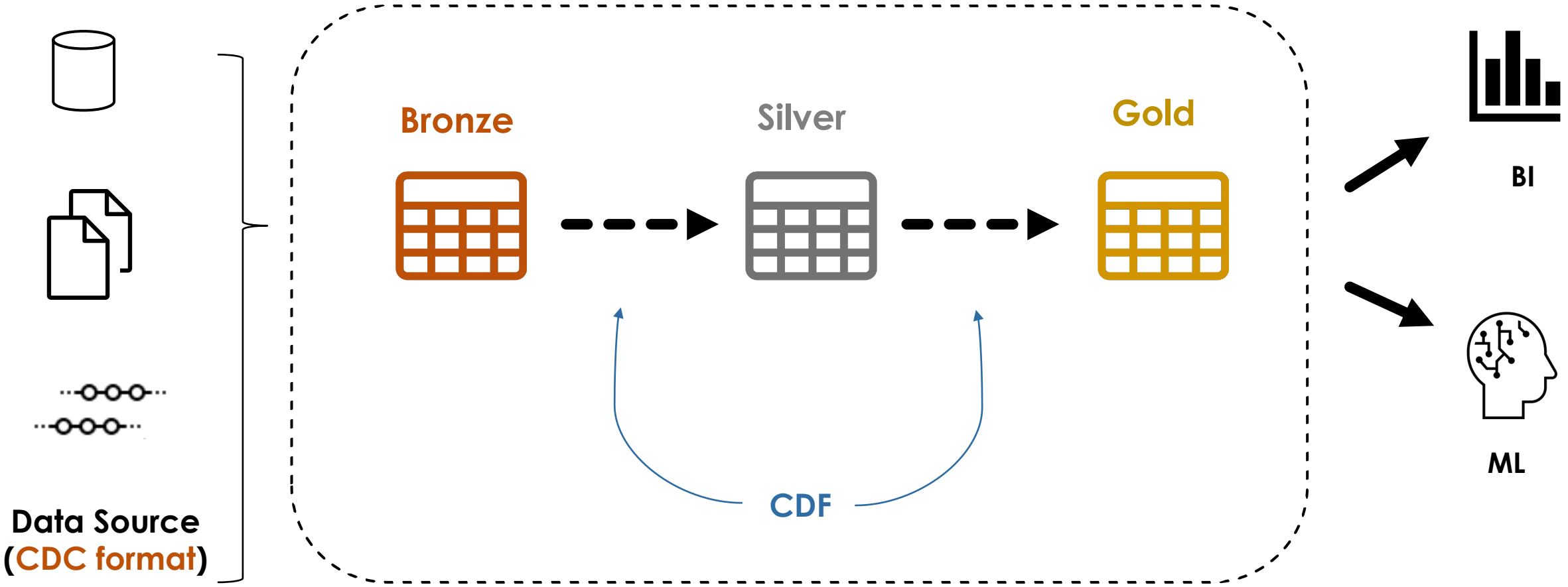
Learning Objectives

- What is CDF
- Enabling CDF
- When to use CDF

What is CDF

- Delta Lake Change Data Feed (CDF)
- Automatically generate CDC feeds about Delta Lake tables.
- Records row-level changes for all data written into a Delta table
  - Row data + metadata (whether row was inserted, deleted, or updated)

Multi-Hop Architecture



CDF

Delta table (v1)

Country ID	Country	Vaccination Rate
FR	France	0.7
CA	Canada	0.65
IN	India	0.6

CDF

Delta table (v2)

Country ID	Country	Vaccination Rate
FR	France	0.75
CA	Canada	0.65
IN	India	0.6
US	USA	0.5

CDF

Delta table (v2)

Country ID	Country	Vaccination Rate
FR	France	0.75
CA	Canada	0.65
IN	India	0.6
US	USA	0.5

table_changes

Country ID	Country	Vaccination Rate	Change Type	Time	Version
FR	France	0.7	update_preimage	07:00:00	2
FR	France	0.75	update_postimage	07:00:00	2
US	USA	0.5	insert	07:00:00	2

CDF

Delta table (v2)

Country ID	Country	Vaccination Rate
FR	France	0.75
CA	Canada	0.65
IN	India	0.6
US	USA	0.5

table_changes

Country ID	Country	Vaccination Rate	Change Type	Time	Version
FR	France	0.7	update_preimage	07:00:00	2
FR	France	0.75	update_postimage	07:00:00	2
US	USA	0.5	insert	07:00:00	2

CDF

Delta table (v3)

Country ID	Country	Vaccination Rate
FR	France	0.75
IN	India	0.6
US	USA	0.5

table_changes

Country ID	Country	Vaccination Rate	Change Type	Time	Version
FR	France	0.7	update_preimage	07:00:00	2
FR	France	0.75	update_postimage	07:00:00	2
US	USA	0.5	insert	07:00:00	2
CA	Canada	0.65	delete	08:00:00	3

Querying the change data

- SELECT *

  FROM **table_changes**('table_name', start_version, [end_version])

- SELECT *

  FROM **table_changes**('table_name', start_timestamp, [end_timestamp])

Enabling CDF

- New tables
  - ```
    CREATE TABLE myTable (id INT, name STRING)
    TBLPROPERTIES (delta.enableChangeDataFeed = true)
    ```
- Existing table
  - ```
    ALTER TABLE myTable
    SET TBLPROPERTIES (delta.enableChangeDataFeed = true)
    ```
- All new tables
  - `spark.databricks.delta.properties.defaults.enableChangeDataFeed`

CDF retention

- Follow the retention policy of the table
- When running VACUUM, CDF data is also deleted.

When to use CDF

Use CDF when

- Table's changes include updates and/or deletes
- Small franction of records updated in each batch (from CDC feed)

Don't use CDF when

- Table's changes are append-only
- Most records in the table updated in each batch