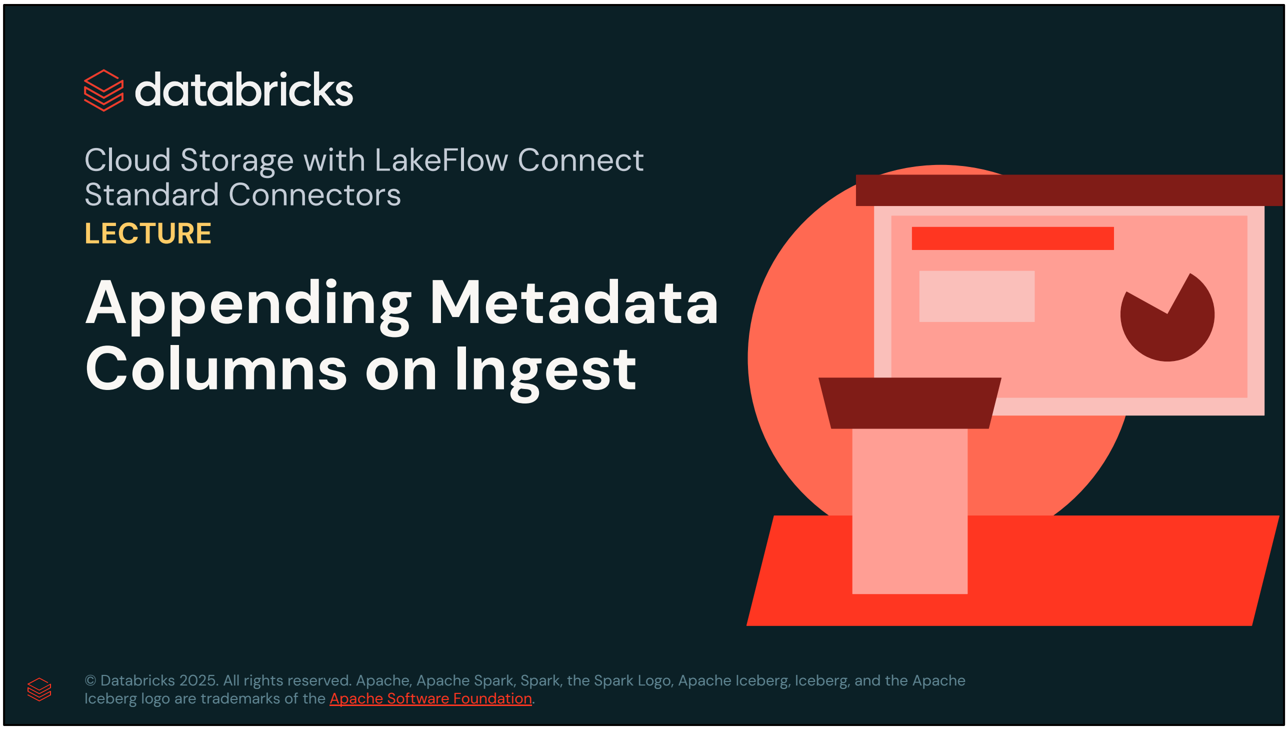


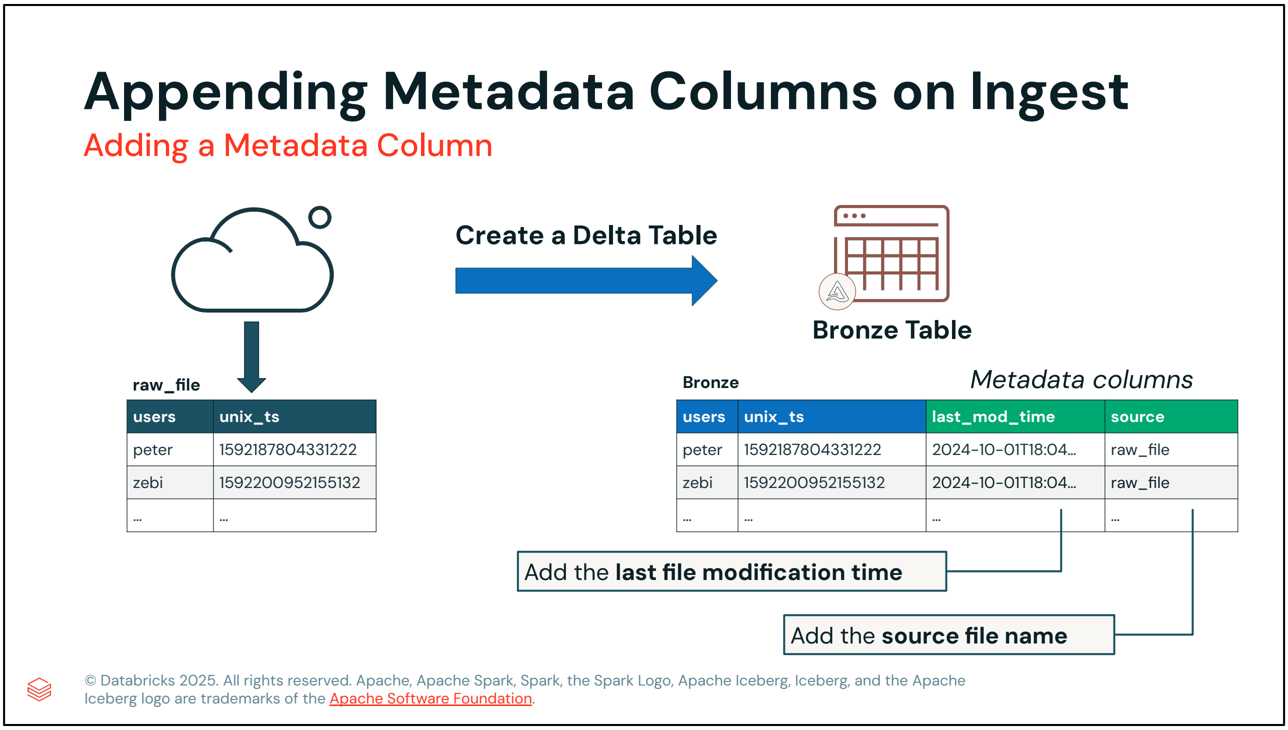
databricks

Cloud Storage with LakeFlow Connect Standard Connectors

LECTURE

Appending Metadata Columns on Ingest

In data ingestion from cloud storage, metadata columns like source file name and modification time can be appended using the _metadata column—capturing essential context for each row during table creation in the Lakehouse.



Appending Metadata Columns on Ingest

Adding a Metadata Column

raw_file

users	unix_ts
peter	1592187804331222
zebi	1592200952155132
…	…

Bronze

Metadata columns

users	unix_ts	last_mod_time	source
peter	1592187804331222	2024-10-01T18:04…	raw_file
zebi	1592200952155132	2024-10-01T18:04…	raw_file
…	…	…	…

Add the **last file modification time**

Add the **source file name**

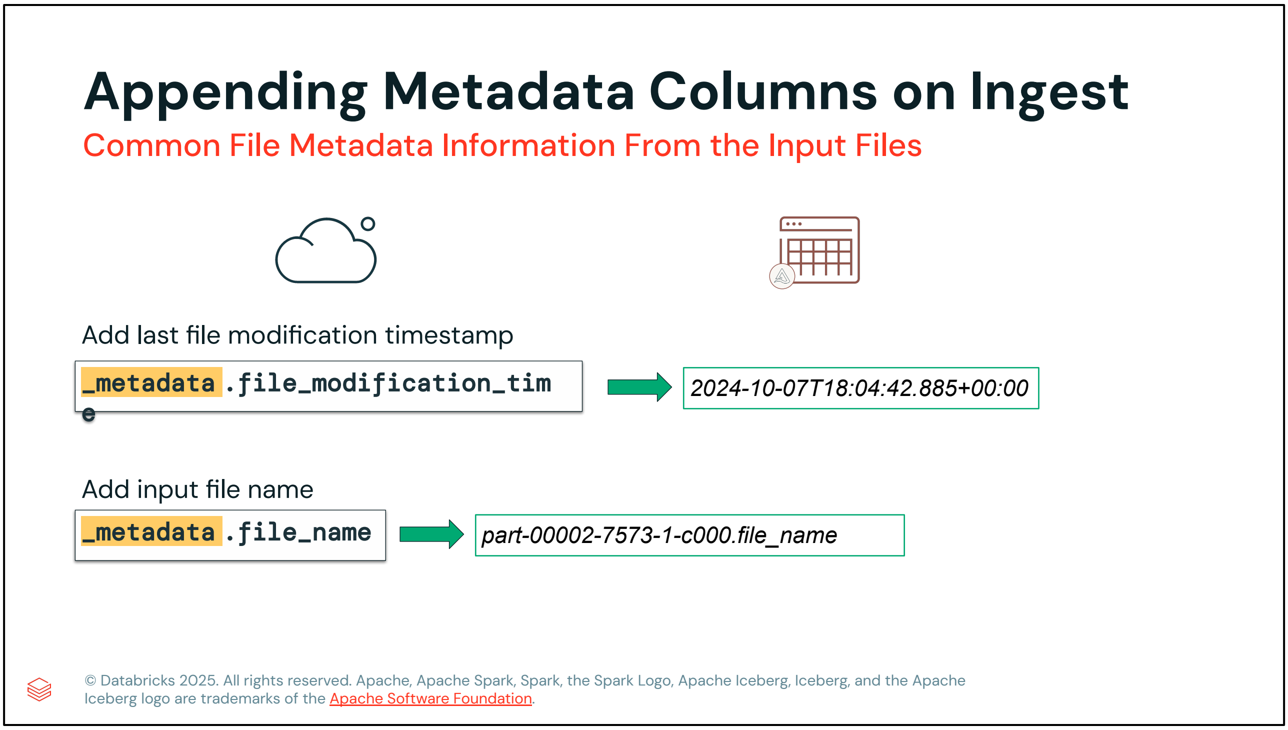
You can append metadata column information from input data source files when creating a table.

For example, suppose your cloud storage location contains a set of raw files (these could be CSV, TXT, JSON, or other formats) and you want to ingest those files into a bronze-level table.

As part of that ingestion, you may want to add specific metadata columns to each row in the table. These columns can include information from the ingestion source, such as:

- The last modification time of the file the row originated from
- The source file name

This helps preserve important context about the data's origin, which can be valuable for auditing, lineage, and debugging purposes.



Appending Metadata Columns on Ingest

Common File Metadata Information From the Input Files

Add last file modification timestamp

`_metadata.file_modification_time` → *2024-10-07T18:04:42.885+00:00*

Add input file name

`_metadata.file_name` → *part-00002-7573-1-c000.file_name*

To add metadata columns during ingestion, you can use the special **_metadata** column. This is a hidden column that is available for all input file formats. To include the **_metadata** column in the returned DataFrame, you must explicitly select it in your read query when specifying the source.

The _metadata column contains a variety of useful fields. Two common fields include:

- **_metadata.file_modification_time** – provides the last modification timestamp of the input file

- **_metadata.file_name** – returns the name of the source file for each row

Documentation: https://docs.databricks.com/aws/en/ingestion/file-metadata-column

databricks

Thank you for completing this lesson and continuing your journey to develop your skills with us.