Basic things need to know

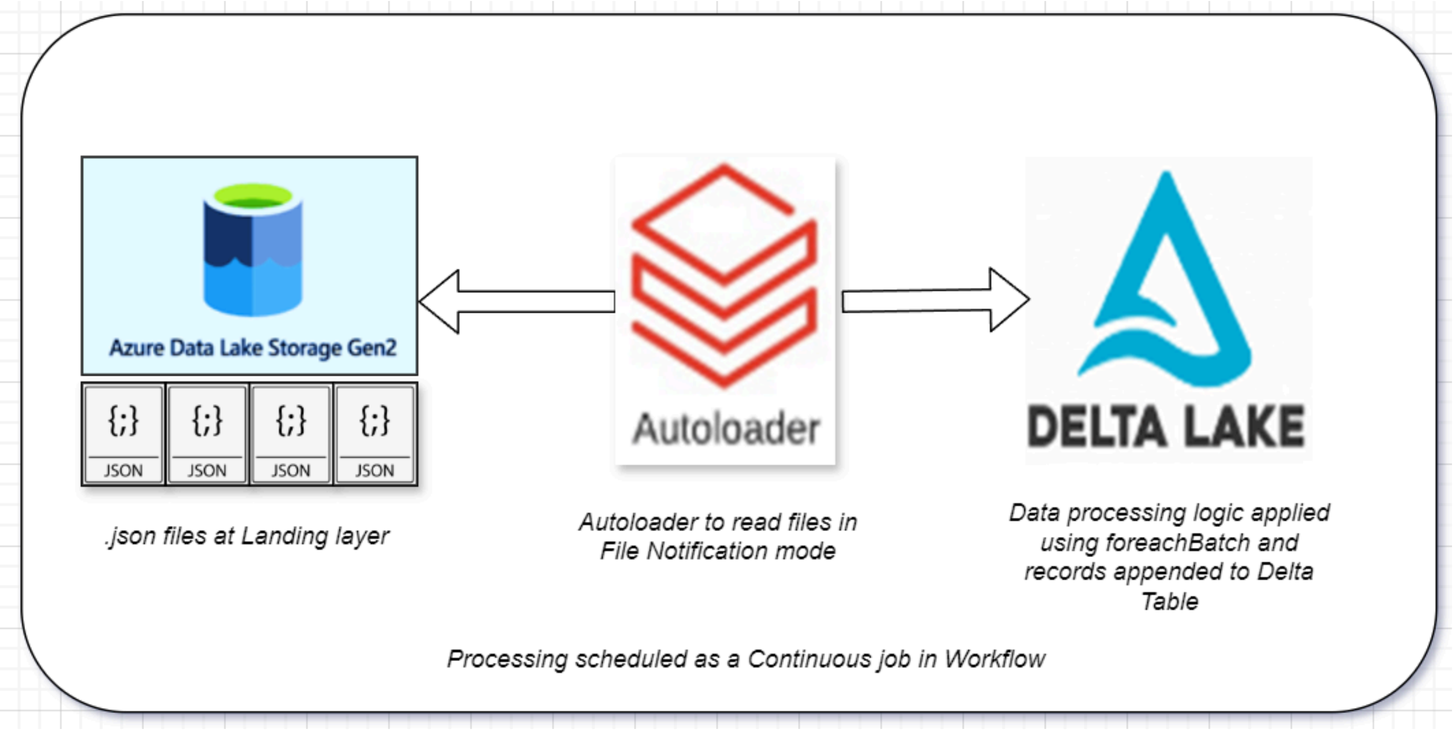
Structured Streaming

Structured Streaming in Databricks is a real-time data processing framework that enables continuous and incremental processing of streaming data.

Problems with Structrued Streaming

- Slow in performance [Spark relies on file listing (listFiles()) to check for new files]
- No dynamically schema inference
- No Schema Evolution
- Cost reduction by reducing API calls

Autoloader


Azure Data Lake Storage Gen2
JSON
JSON
JSON
JSON
.json files at Landing layer
Autoloader
Autoloader to read files in File Notification mode
DELTA LAKE
Data processing logic applied using foreachBatch and records appended to Delta Table
Processing scheduled as a Continuous job in Workflow

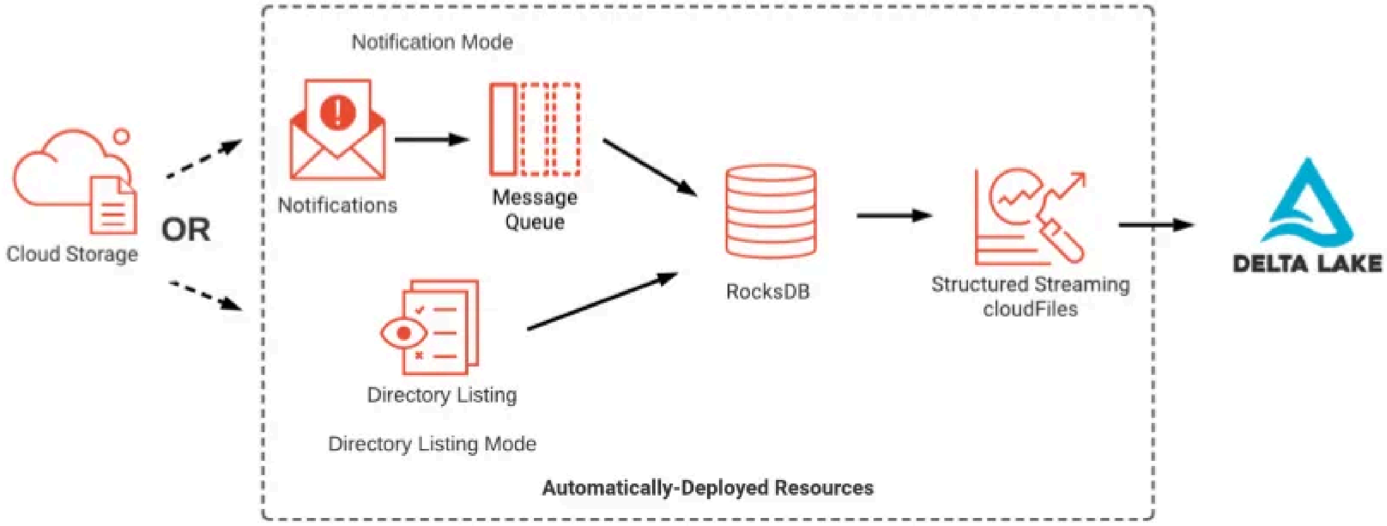
Autoloader is a Databricks feature built on Structured Streaming that allows incremental ingestion of new files from cloud storage.

Key Features

- **Incremental Ingestion:** Detects and processes new files without reloading old ones.

- **Schema Evolution:** Supports automatic handling of new columns.
- **Checkpointing:** Ensures fault tolerance by tracking processed files.

Modes



1. Directory Listing Mode

- The default mode in Autoloader for detecting new files.
- Autoloader lists files incrementally instead of re-scanning all files.
- Uses a checkpoint location to track processed files.
- New files are discovered efficiently, avoiding reprocessing.

2. File Notification Mode

- Creates Event Notification Object in Cloud
- The event notification informs Autoloader that a new file has been uploaded or created in the specified storage location.
- The event notification is placed in a message queue
- The queue holds the notification until it's ready for processing
- Autoloader listens to the message queue for new notifications.
- When a new notification arrives Autoloader picks it up.
- Autoloader reads the new file based on the event notification.