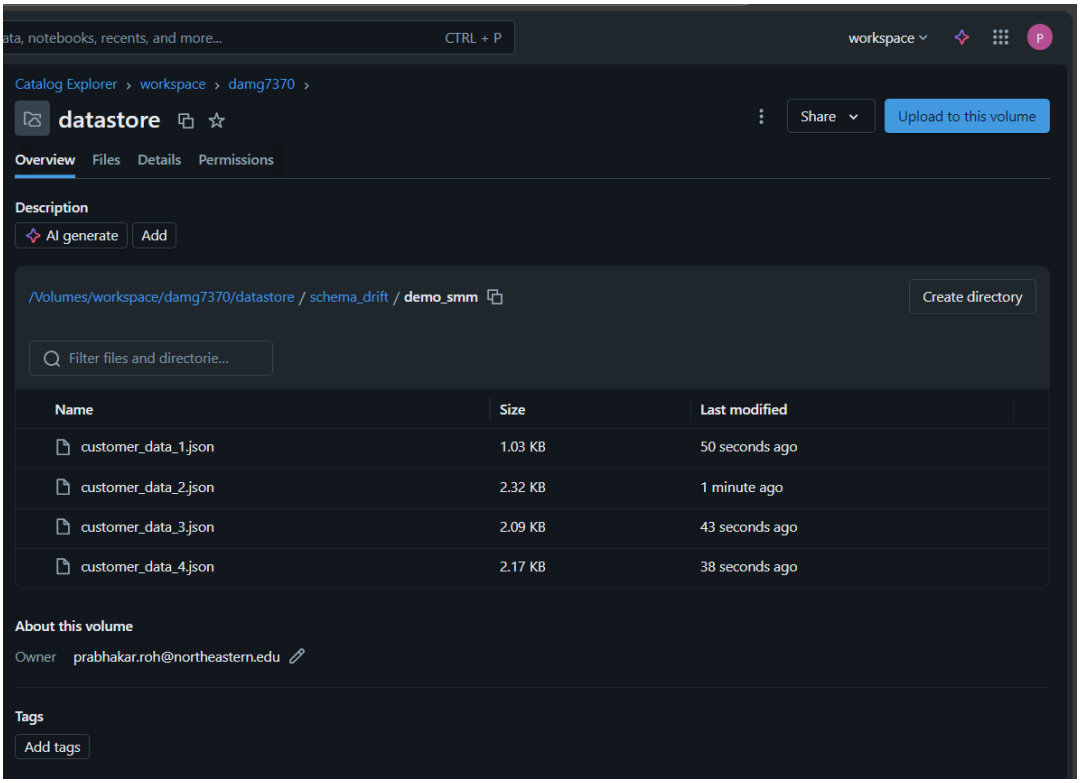
DAMG 7370 Designing Advanced Data Architectures for Business Intelligence

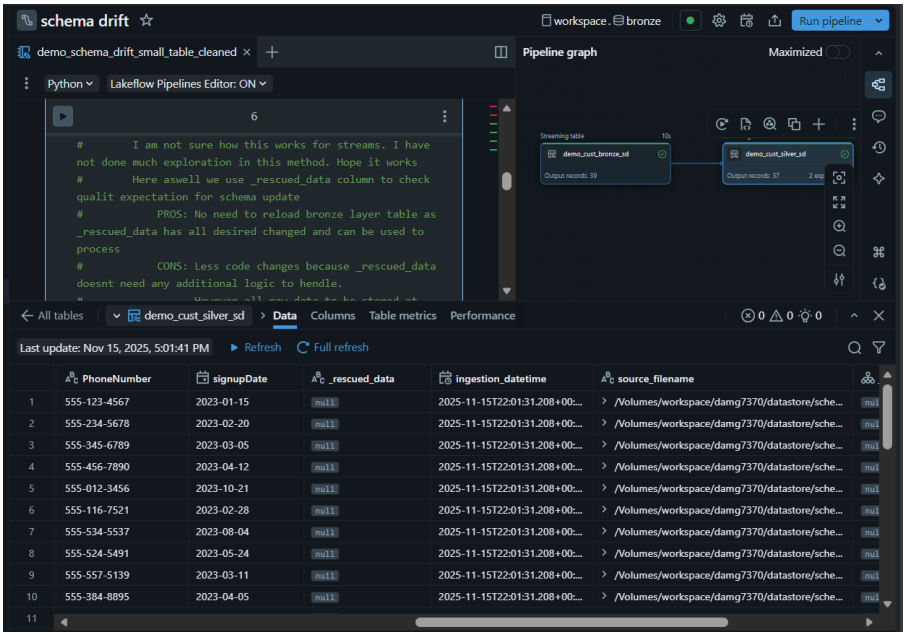
Schema Drift

PART 1

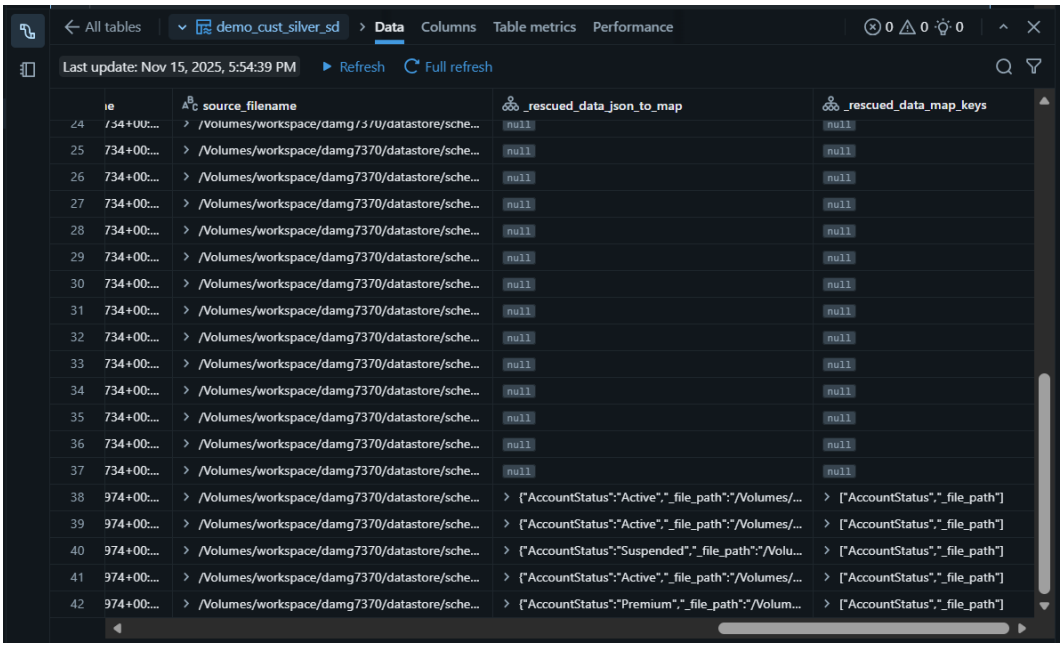
Uploading json files to the volume



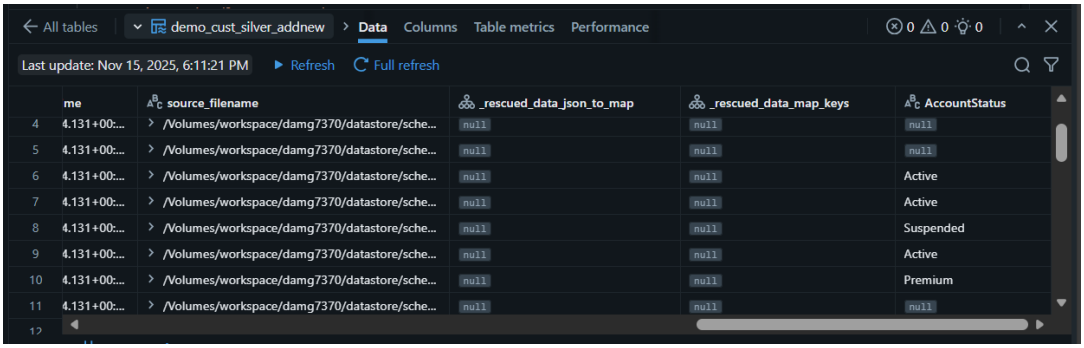
Testing pipeline to see data in bronze



Rescued_data columns are created in rescued silver table but no new column is added



For add new_columns the column is added directly



KEY OBSERVATIONS:

1. **RESCUE MODE (demo_cust_bronze_sd)**:

- AccountStatus appears in _rescued_data as: {"AccountStatus":"Active"}
- Need to query _rescued_data to see the values
- Columns NOT automatically added to table schema

2. **ADDNEWCOLUMNS MODE (demo_cust_bronze_addnew)**:

- AccountStatus appears as actual column with value "Active"
- Can directly SELECT these columns
- _rescued_data is ALWAYS NULL
- Table schema grows automatically

3. WHEN TO USE EACH:

RESCUE MODE :

- Production environments requiring governance
- Need to validate before accepting schema changes
- Want audit trail of schema evolution
- Compliance/regulatory requirements

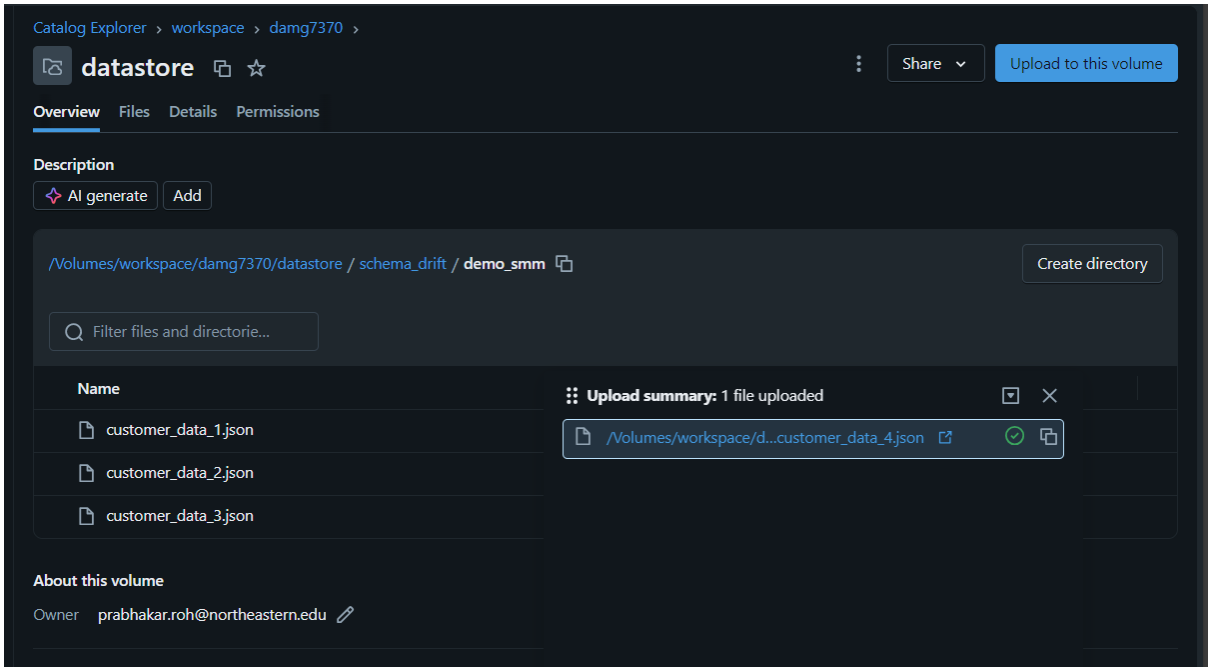
ADDNEWCOLUMNS MODE :

- Development environments
- Rapid prototyping

- Trusted data sources

- When flexibility > control

PART 2

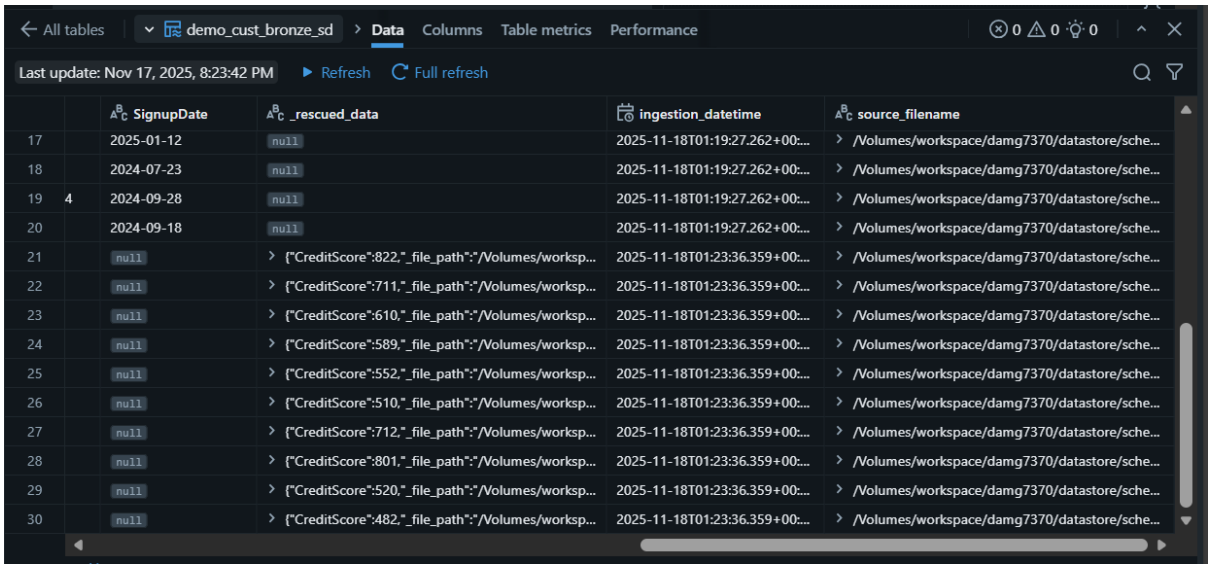


Problem in current pipeline:

Streaming DataFrames represent continuous, unbounded data flows with no defined endpoint, while .collect() is designed for finite datasets where Spark can read all rows and return them to the driver. When you call .collect() on a batch DataFrame, Spark knows there are exactly N rows to materialize and can complete the operation. However, streaming DataFrames are conceptually infinite - data keeps arriving indefinitely, so there's no "last row" to signal completion.

As a result, Spark either throws an error, returns an empty result as a safeguard, or hangs indefinitely waiting for stream termination that never comes. In your original function, the check if not df.isStreaming else [] evaluates to True for streaming DataFrames, so zit returns an empty list [], the for loop iterates zero times, and no columns get added - causing silent failure.

The bronze captures the change in rescued data

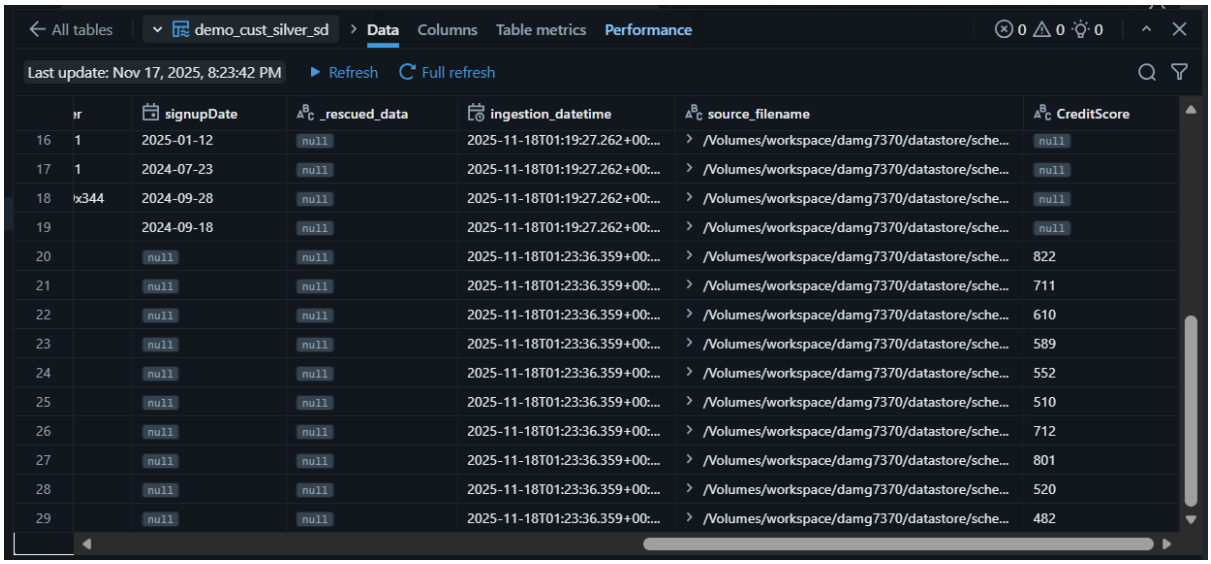


		SignupDate	_rescued_data	ingestion_datetime	source_filename
17		2025-01-12	null	2025-11-18T01:19:27.262+00:...	/Volumes/workspace/damg7370/datastore/sche...
18		2024-07-23	null	2025-11-18T01:19:27.262+00:...	/Volumes/workspace/damg7370/datastore/sche...
19	4	2024-09-28	null	2025-11-18T01:19:27.262+00:...	/Volumes/workspace/damg7370/datastore/sche...
20		2024-09-18	null	2025-11-18T01:19:27.262+00:...	/Volumes/workspace/damg7370/datastore/sche...
21		null	{"CreditScore":822,"_file_path":"/Volumes/worksp...	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...
22		null	{"CreditScore":711,"_file_path":"/Volumes/worksp...	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...
23		null	{"CreditScore":610,"_file_path":"/Volumes/worksp...	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...
24		null	{"CreditScore":589,"_file_path":"/Volumes/worksp...	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...
25		null	{"CreditScore":552,"_file_path":"/Volumes/worksp...	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...
26		null	{"CreditScore":510,"_file_path":"/Volumes/worksp...	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...
27		null	{"CreditScore":712,"_file_path":"/Volumes/worksp...	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...
28		null	{"CreditScore":801,"_file_path":"/Volumes/worksp...	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...
29		null	{"CreditScore":520,"_file_path":"/Volumes/worksp...	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...
30		null	{"CreditScore":482,"_file_path":"/Volumes/worksp...	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...

Giving expected columns for the new column to be updated is the solution or to make both the tables batch to do proper schema inference


```python
def process__rescue_data_new_fields(df):

    # Hardcode ALL fields that might appear in _rescued_data
    expected_rescued_fields = [
        "CreditScore"

    ]

    # Parse _rescued_data to map
    df = df.withColumn(
        "_rescued_map",
        from_json(
            col("_rescued_data"),
            MapType(StringType(), StringType())
        )
    )

    # Extract each field from _rescued_data
    for field_name in expected_rescued_fields:
        # Only add if column doesn't already exist
        if field_name not in df.columns:
            df = df.withColumn(
                field_name,
                when(
                    col("_rescued_map").isNotNull(),
```

Credit Score being populated in the Silver table with corrected code



	ır	signupDate	_rescued_data	ingestion_datetime	source_filename	CreditScore
16	1	2025-01-12	null	2025-11-18T01:19:27.262+00:...	/Volumes/workspace/damg7370/datastore/sche...	null
17	1	2024-07-23	null	2025-11-18T01:19:27.262+00:...	/Volumes/workspace/damg7370/datastore/sche...	null
18	bx344	2024-09-28	null	2025-11-18T01:19:27.262+00:...	/Volumes/workspace/damg7370/datastore/sche...	null
19		2024-09-18	null	2025-11-18T01:19:27.262+00:...	/Volumes/workspace/damg7370/datastore/sche...	null
20		null	null	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...	822
21		null	null	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...	711
22		null	null	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...	610
23		null	null	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...	589
24		null	null	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...	552
25		null	null	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...	510
26		null	null	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...	712
27		null	null	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...	801
28		null	null	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...	520
29		null	null	2025-11-18T01:23:36.359+00:...	/Volumes/workspace/damg7370/datastore/sche...	482