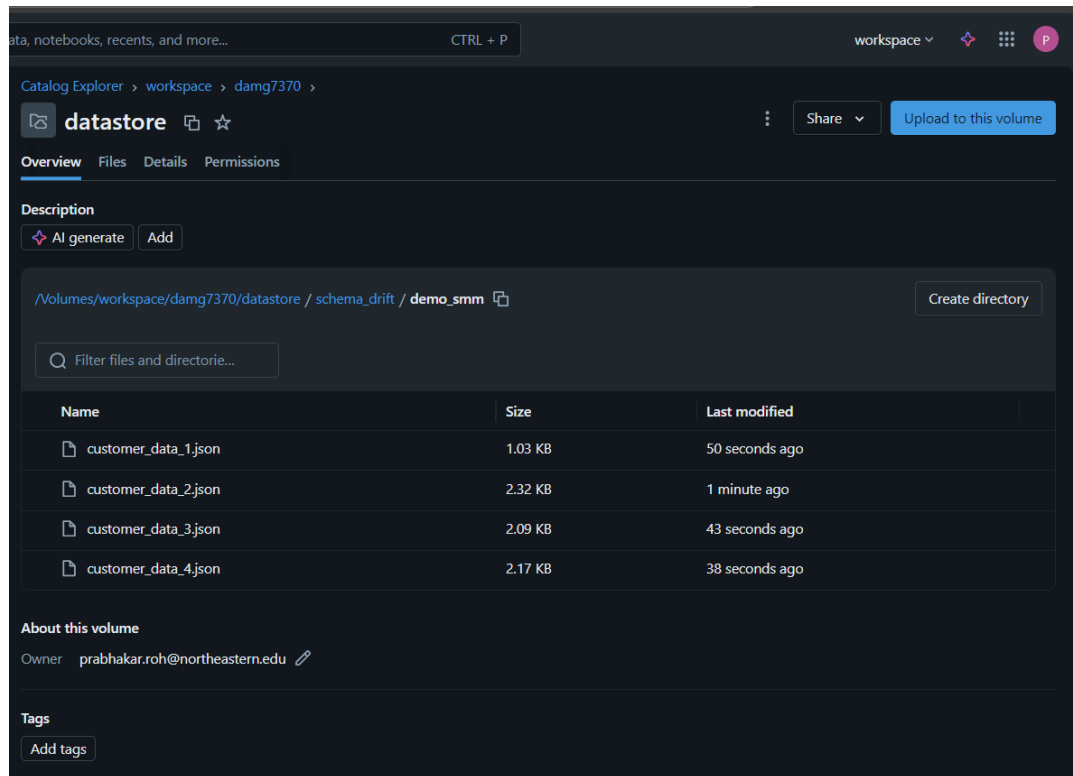


# DAMG 7370 Designing Advanced Data Architectures for Business Intelligence

## Schema Drift

### PART 1

Uploading json files to the volume



Testing pipeline to see data in bronze

schema drift

workspace: bronze

demo\_schema\_drift\_small\_table\_cleaned

Python Lakeflow Pipelines Editor: ON

6

```
# I am not sure how this works for streams. I have
# not done much exploration in this method. Hope it works
# Here as well we use _rescued_data column to check
# quality expectation for schema update
# PROS: No need to reload bronze layer table as
# _rescued_data has all desired changed and can be used to
# process
# CONS: Less code changes because _rescued_data
# doesn't need any additional logic to handle.
```

Pipeline graph

Streaming table: demo\_cust\_bronze\_sd

Output records: 39

demo\_cust\_silver\_sd

Output records: 37

2 exp

demo\_cust\_silver\_sd

Columns

	PhoneNumber	signupDate	_rescued_data	ingestion_datetime	source_filename
1	555-123-4567	2023-01-15	null	2025-11-15T22:01:31.208+00...	/Volumes/workspace/damg7370/datastore/sche...
2	555-234-5678	2023-02-20	null	2025-11-15T22:01:31.208+00...	/Volumes/workspace/damg7370/datastore/sche...
3	555-345-6789	2023-03-05	null	2025-11-15T22:01:31.208+00...	/Volumes/workspace/damg7370/datastore/sche...
4	555-456-7890	2023-04-12	null	2025-11-15T22:01:31.208+00...	/Volumes/workspace/damg7370/datastore/sche...
5	555-012-3456	2023-10-21	null	2025-11-15T22:01:31.208+00...	/Volumes/workspace/damg7370/datastore/sche...
6	555-116-7521	2023-02-28	null	2025-11-15T22:01:31.208+00...	/Volumes/workspace/damg7370/datastore/sche...
7	555-534-5537	2023-08-04	null	2025-11-15T22:01:31.208+00...	/Volumes/workspace/damg7370/datastore/sche...
8	555-524-5491	2023-05-24	null	2025-11-15T22:01:31.208+00...	/Volumes/workspace/damg7370/datastore/sche...
9	555-557-5139	2023-03-11	null	2025-11-15T22:01:31.208+00...	/Volumes/workspace/damg7370/datastore/sche...
10	555-384-8895	2023-04-05	null	2025-11-15T22:01:31.208+00...	/Volumes/workspace/damg7370/datastore/sche...
11					

Rescued\_data columns are created in rescued silver table but no new column is added

demo\_cust\_silver\_sd

Columns

	source_filename	_rescued_data_json_to_map	_rescued_data_map_keys
24	/Volumes/workspace/damg7370/datastore/sche...	null	null
25	/Volumes/workspace/damg7370/datastore/sche...	null	null
26	/Volumes/workspace/damg7370/datastore/sche...	null	null
27	/Volumes/workspace/damg7370/datastore/sche...	null	null
28	/Volumes/workspace/damg7370/datastore/sche...	null	null
29	/Volumes/workspace/damg7370/datastore/sche...	null	null
30	/Volumes/workspace/damg7370/datastore/sche...	null	null
31	/Volumes/workspace/damg7370/datastore/sche...	null	null
32	/Volumes/workspace/damg7370/datastore/sche...	null	null
33	/Volumes/workspace/damg7370/datastore/sche...	null	null
34	/Volumes/workspace/damg7370/datastore/sche...	null	null
35	/Volumes/workspace/damg7370/datastore/sche...	null	null
36	/Volumes/workspace/damg7370/datastore/sche...	null	null
37	/Volumes/workspace/damg7370/datastore/sche...	null	null
38	/Volumes/workspace/damg7370/datastore/sche...	{"AccountStatus": "Active", "_file_path": "/Volumes/...	["AccountStatus", "_file_path"]
39	/Volumes/workspace/damg7370/datastore/sche...	{"AccountStatus": "Active", "_file_path": "/Volumes/...	["AccountStatus", "_file_path"]
40	/Volumes/workspace/damg7370/datastore/sche...	{"AccountStatus": "Suspended", "_file_path": "/Volum...	["AccountStatus", "_file_path"]
41	/Volumes/workspace/damg7370/datastore/sche...	{"AccountStatus": "Active", "_file_path": "/Volumes/...	["AccountStatus", "_file_path"]
42	/Volumes/workspace/damg7370/datastore/sche...	{"AccountStatus": "Premium", "_file_path": "/Volum...	["AccountStatus", "_file_path"]

For add new\_columns the column is added directly

	me	source_filename	_rescued_data_json_to_map	_rescued_data_map_keys	AccountStatus
4	4.131+00:...	> /Volumes/workspace/damg7370/datastore/sche...	null	null	null
5	4.131+00:...	> /Volumes/workspace/damg7370/datastore/sche...	null	null	null
6	4.131+00:...	> /Volumes/workspace/damg7370/datastore/sche...	null	null	Active
7	4.131+00:...	> /Volumes/workspace/damg7370/datastore/sche...	null	null	Active
8	4.131+00:...	> /Volumes/workspace/damg7370/datastore/sche...	null	null	Suspended
9	4.131+00:...	> /Volumes/workspace/damg7370/datastore/sche...	null	null	Active
10	4.131+00:...	> /Volumes/workspace/damg7370/datastore/sche...	null	null	Premium
11	4.131+00:...	> /Volumes/workspace/damg7370/datastore/sche...	null	null	null
12					

## 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 :

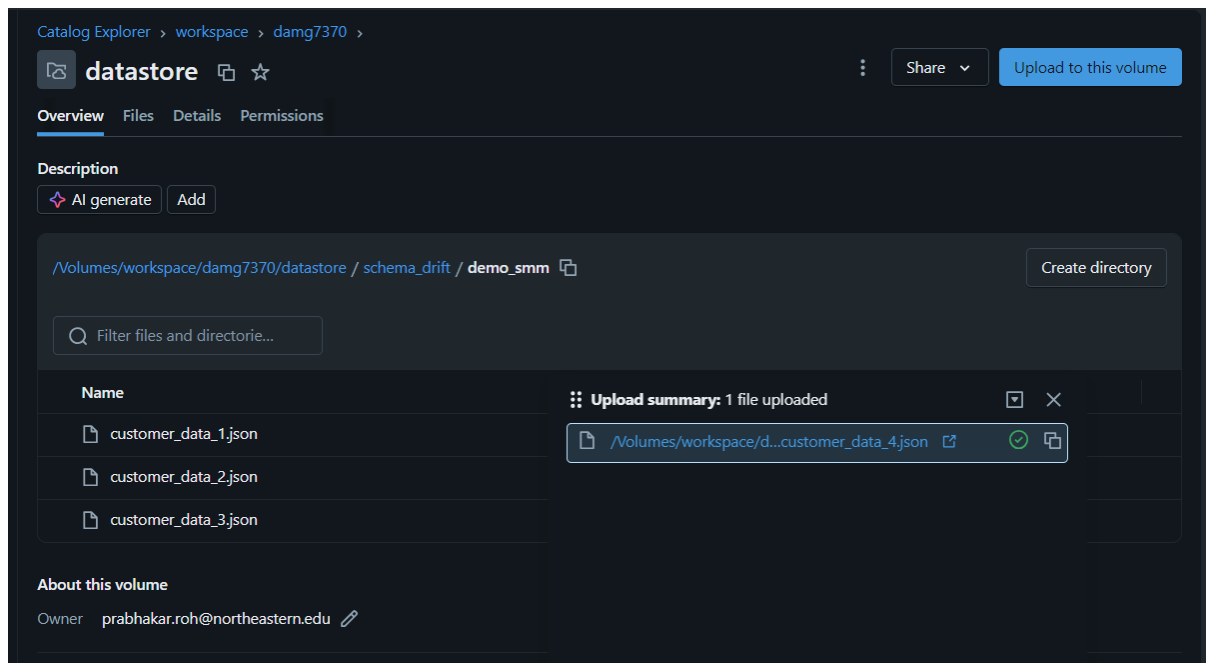
- Production environments requiring governance
- Need to validate before accepting schema changes
- Want audit trail of schema evolution
- Compliance/regulatory requirements

#### ADDNEWCOLUMNS MODE :

- Development environments
- Rapid prototyping

- Trusted data sources
- When flexibility > control

## PART 2

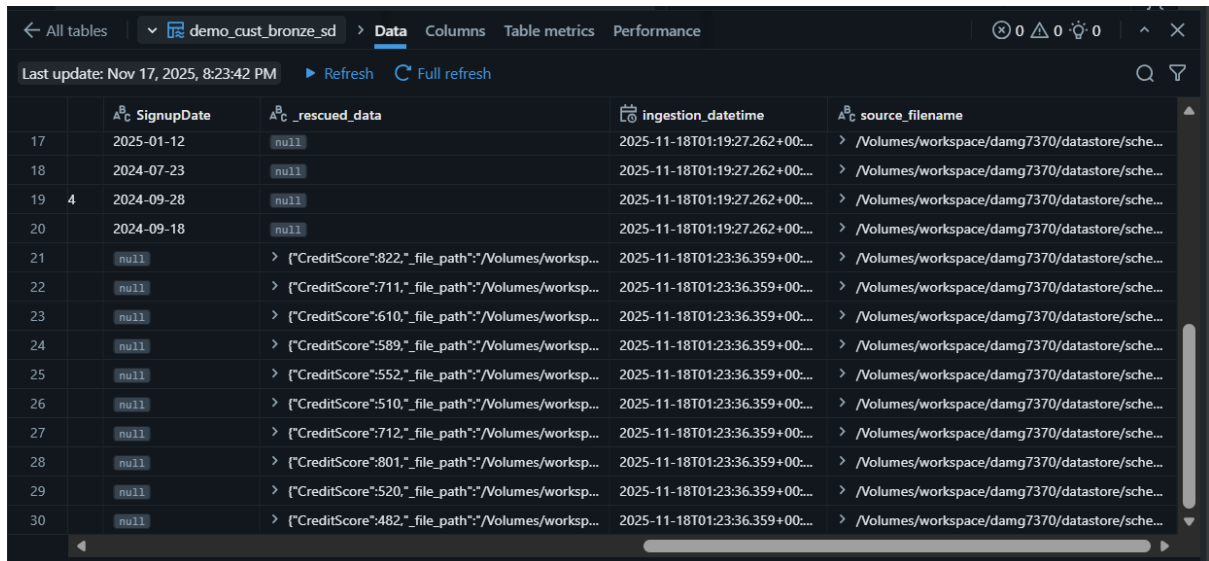


### Problem in current pipeline:

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	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

```
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(),  
                    col(f"_rescued_map.{field_name}"),  
                    null()  
                )  
            )
```

Credit Score being populated in the Silver table with corrected code

All tablesdemo\_cust\_silver\_sdDataColumnsTable metricsPerformance

Last update: Nov 17, 2025, 8:23:42 PMRefreshFull refresh

sr	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	bc344	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