**Common Issues and Resolutions in Databricks**

**Schema drift** and **schema evolution** handle **VARIANT data types**, what causes drift in Databricks Delta Lake.

**Databricks Delta Schema Issues Comparison Table (Bronze/Silver/Gold)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Schema Issue** | **Description** | **Where It Typically Arises** | **Delta Layer(s) Most Affected** |
| 1 | **Schema Drift** | Unexpected changes in incoming data structure (new, missing, renamed columns). | Ingestion (Auto Loader, streaming reads) | **Bronze** |
| 2 | **Schema Evolution** | Automatic incorporation of schema changes into Delta tables. | Writes to Delta tables with mergeSchema enabled | **Bronze**, Silver |
| 3 | **Data Type Conflicts** | Incoming data columns have types that don’t match the table schema. | Ingestion and append writes to Delta | **Bronze**, Silver |
| 4 | **Nullability Mismatches** | Changes in whether a column allows nulls, causing write or merge errors. | Delta table overwrites, schema enforcement | **Silver**, Gold |
| 5 | **Column Reordering** | Columns arrive in unexpected order, potentially breaking mapping. | Batch ingestion with schema inference | **Bronze** |
| 6 | **Nested Field Changes** | New fields or type changes inside nested structs or JSON columns. | Ingestion of semi-structured data | **Bronze**, Silver |
| 7 | **Column Dropping** | Expected columns disappear from incoming data, leading to nulls or errors. | Ingestion pipelines (e.g., Auto Loader) | **Bronze**, Silver |
| 8 | **Column Renaming** | Columns are renamed upstream without notice, breaking downstream queries. | Ingestion and transformations | **Bronze**, Silver |
| 9 | **Schema Inference Variability** | Automatic inference guesses inconsistent types across files. | Auto Loader and .option("inferSchema", "true") | **Bronze** |
| 10 | **Backward Incompatible Changes** | Schema changes that can’t be merged automatically, requiring manual fixes. | Table evolution (dropping or retyping columns) | **Silver**, **Gold** |

**Quick Explanation of Layers:**

* **Bronze:** Raw ingestion layer, accepts as much as possible, including drift.
* **Silver:** Cleaned and standardized data, schema must be consistent.
* **Gold:** Curated, aggregated, business-ready tables—very sensitive to schema mismatches.

**Example Interpretation:**

* **Schema Drift**: Mostly impacts **Bronze**, because that’s where new columns show up unexpectedly.
* **Data Type Conflicts**: Show up in Bronze if you merge raw files, and again in Silver when you try to standardize.
* **Backward Incompatible Changes**: Usually occur in **Silver or Gold**, when you evolve the curated schema in a way that breaks existing queries

**Schema Issues with Mitigation Strategies by Delta Layer**

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Schema Issue** | **Delta Layer(s) Affected** | **Mitigation Strategy** |
| 1 | **Schema Drift** | Bronze | . Use mergeSchema=true in Auto Loader to accept new columns. . Store raw JSON or text if drift is frequent. . Capture schema in schemaLocation for auditing. |
| 2 | **Schema Evolution** | Bronze, Silver | . Enable mergeSchema selectively. . Use explicit schemas in Silver to standardize. . Monitor schema changes with DESCRIBE HISTORY. |
| 3 | **Data Type Conflicts** | Bronze, Silver | . Cast fields to consistent types before writing. . Validate incoming types during ingestion. . Use raw ingestion in Bronze and parse safely in Silver. |
| 4 | **Nullability Mismatches** | Silver, Gold | . Define schemas explicitly in Silver. . Use ALTER TABLE to adjust nullability. . Validate nullability during transformations. |
| 5 | **Column Reordering** | Bronze | . Always define schemas instead of relying on inference. . Use column mapping by name rather than position. |
| 6 | **Nested Field Changes** | Bronze, Silver | . Store raw JSON in Bronze. . Parse nested structures carefully in Silver. . Cast nested types consistently. |
| 7 | **Column Dropping** | Bronze, Silver | . Implement data quality checks to detect missing columns. . Use default values or null fills in Silver. . Alert on unexpected column loss. |
| 8 | **Column Renaming** | Bronze, Silver | . Avoid renames in raw data. . Maintain mapping tables for renamed fields. . Standardize names during Silver transformations. |
| 9 | **Schema Inference Variability** | Bronze | . Avoid .option("inferSchema","true") for production. . Use explicit schemas. . Revalidate schemas periodically. |
| 10 | **Backward Incompatible Changes** | Silver, Gold | . Use versioned tables (VERSION AS OF) for recovery. . Plan schema changes carefully with stakeholders. . Document all changes and test impact. |

**Schema Issues, Mitigation Strategies, and Code Examples (by Delta Layer)**

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Schema Issue** | **Delta Layer(s)** | **Mitigation Strategy & Code Example** |
| 1 | **Schema Drift** | **Bronze** | . **Use mergeSchema to accept new columns** python df = (spark.readStream.format("cloudFiles") .option("cloudFiles.format", "json") .option("mergeSchema","true") .load("/mnt/raw/")) . **Store raw JSON** python df = spark.readStream.format("cloudFiles").option("cloudFiles.format","text").load("/mnt/raw/") |
| 2 | **Schema Evolution** | **Bronze, Silver** | . **Merge schema when appending** python df.write.option("mergeSchema","true").format("delta").mode("append").save("/mnt/delta/bronze") . **Track schema changes** sql DESCRIBE HISTORY delta.`/mnt/delta/bronze` |
| 3 | **Data Type Conflicts** | **Bronze, Silver** | . **Cast types before writing** python df = df.withColumn("id", col("id").cast("string")) . **Store raw text and parse in Silver** python df = spark.readStream.format("cloudFiles").option("cloudFiles.format","text").load("/mnt/raw/") |
| 4 | **Nullability Mismatches** | **Silver, Gold** | . **Adjust nullability** sql ALTER TABLE silver\_table ALTER COLUMN user\_id DROP NOT NULL |
| 5 | **Column Reordering** | **Bronze** | . **Define schema explicitly** python schema = StructType([ StructField("id", StringType()), StructField("value", StringType()) ]) df = spark.read.schema(schema).json("/mnt/raw/") |
| 6 | **Nested Field Changes** | **Bronze, Silver** | . **Store raw JSON in Bronze** python df = spark.readStream.format("cloudFiles").option("cloudFiles.format","text").load("/mnt/raw/") . **Parse safely in Silver** python parsed\_df = df.withColumn("data", from\_json(col("value"), schema)) |
| 7 | **Column Dropping** | **Bronze, Silver** | . **Fill defaults in Silver** python df = df.withColumn("event\_type", coalesce(col("event\_type"), lit("unknown"))) . **Alert on missing columns** python expected = {"user\_id","event\_type"} actual = set(df.columns) if missing := expected - actual: print("Missing:", missing) |
| 8 | **Column Renaming** | **Bronze, Silver** | . **Map renamed fields** python df = df.withColumnRenamed("old\_name","new\_name") |
| 9 | **Schema Inference Variability** | **Bronze** | . **Avoid inference** python schema = StructType([...]) df = spark.read.schema(schema).json("/mnt/raw/") |
| 10 | **Backward Incompatible Changes** | **Silver, Gold** | . **Time travel recovery** python df = spark.read.format("delta").option("versionAsOf",3).load("/mnt/delta/silver") . **Test schema changes** sql DESCRIBE HISTORY delta.`/mnt/delta/silver` |

. **Quick Layer Recap:**

* **Bronze**: Ingest raw—accept drift, capture everything.
* **Silver**: Clean—apply standard schema, handle types and nullability.
* **Gold**: Curate—no tolerance for instability.

**Example Flow: Handling Drift + Type Conflicts**

**Bronze (Raw JSON as text):**

python

CopyEdit

bronze\_df = spark.readStream.format("cloudFiles").option("cloudFiles.format","text").load("/mnt/raw/")

bronze\_df.writeStream.format("delta").option("checkpointLocation","/mnt/checkpoints/bronze").start("/mnt/delta/bronze")

**Silver (Parse and cast):**

python

CopyEdit

from pyspark.sql.functions import from\_json, col

schema = StructType([

StructField("user\_id", StringType()),

StructField("event\_timestamp", TimestampType())

])

silver\_df = bronze\_df.withColumn("jsonData", from\_json(col("value"), schema))

silver\_df = silver\_df.select(

col("jsonData.user\_id").alias("user\_id"),

col("jsonData.event\_timestamp").alias("event\_timestamp")

)

silver\_df.write.format("delta").mode("overwrite").save("/mnt/delta/silver")

. **This pattern:**

* Tolerates drift (raw text in Bronze).
* Applies explicit schema in Silver.
* Avoids unexpected ingestion failures.

|  |
| --- |
|  |
| **Databricks Notebook: Handling Schema Drift + Evolution**  **1. Notebook Configuration**  python  CopyEdit  # Notebook configuration  bronze\_path = "/mnt/delta/bronze\_user\_events"  silver\_path = "/mnt/delta/silver\_user\_events"  gold\_path = "/mnt/delta/gold\_user\_events"  checkpoint\_bronze = "/mnt/checkpoints/bronze\_user\_events"  checkpoint\_silver = "/mnt/checkpoints/silver\_user\_events"  **2.Bronze Ingestion (Raw JSON as Text)**  **Why?** This step captures all data—**no ingestion failure if schema drifts**.  python  CopyEdit  from pyspark.sql import SparkSession  spark = SparkSession.builder.getOrCreate()  bronze\_df = (  spark.readStream  .format("cloudFiles")  .option("cloudFiles.format", "text") # Read raw text  .load("/mnt/raw/user\_events/") # Change to your source  )  bronze\_write = (  bronze\_df  .writeStream  .format("delta")  .option("checkpointLocation", checkpoint\_bronze)  .outputMode("append")  .start(bronze\_path)  )  . **Tip:** Wait until files arrive and confirm Bronze writes are working.  **3. Define Parsing Schema for Silver**  **Why?** Here you enforce the schema you expect.  python  CopyEdit  from pyspark.sql.types import StructType, StructField, StringType, TimestampType, StructType  parsed\_schema = StructType([  StructField("user\_id", StringType()),  StructField("device", StructType([  StructField("type", StringType()),  StructField("os", StringType())  ])),  StructField("event\_timestamp", TimestampType())  ])  **4. Silver Processing (Parsing + Casting)**  **Why?** This ensures consistent columns and data types.  python  CopyEdit  from pyspark.sql.functions import from\_json, col, lit, coalesce  bronze\_static\_df = spark.read.format("delta").load(bronze\_path)  silver\_df = (  bronze\_static\_df  .withColumn("jsonData", from\_json(col("value"), parsed\_schema))  .select(  col("jsonData.user\_id").alias("user\_id"),  col("jsonData.device.type").alias("device\_type"),  col("jsonData.device.os").alias("device\_os"),  col("jsonData.event\_timestamp").alias("event\_timestamp")  )  .withColumn("device\_type", coalesce(col("device\_type"), lit("unknown"))) # Fill nulls  )  # Preview  silver\_df.display()  **Note:** If you want this step as **streaming**, use:  python  CopyEdit  bronze\_stream\_df = spark.readStream.format("delta").load(bronze\_path)  and then replace bronze\_static\_df with bronze\_stream\_df.  **5. Write Silver Table**  **Batch Write (static example):**  python  CopyEdit  silver\_df.write.format("delta").mode("overwrite").save(silver\_path)  **Streaming Write Example:**  python  CopyEdit  silver\_stream = (  silver\_df.writeStream  .format("delta")  .outputMode("append")  .option("checkpointLocation", checkpoint\_silver)  .start(silver\_path)  )  **6. Query Silver Table**  python  CopyEdit  silver\_query\_df = spark.read.format("delta").load(silver\_path)  silver\_query\_df.display()  Now you have:   * Schema drift **tolerated** in Bronze. * Schema **enforced** in Silver. * Consistent data types.   **7.Optional Gold Table (Aggregation Example)**  python  CopyEdit  gold\_df = (  silver\_query\_df  .groupBy("device\_os")  .count()  .orderBy("count", ascending=False)  )  gold\_df.display()  # Save Gold  gold\_df.write.format("delta").mode("overwrite").save(gold\_path)  **This completes your pipeline:**   * **Bronze:** Raw JSON as text * **Silver:** Parsed & cleaned schema * **Gold:** Aggregated outputs   **8. Validate Schema Changes**  Check schema history:  sql  CopyEdit  DESCRIBE HISTORY delta.`/mnt/delta/bronze\_user\_events`  Inspect schema evolution:  python  CopyEdit  spark.read.format("delta").load(bronze\_path).printSchema()  Time travel example:  python  CopyEdit  df\_old = spark.read.format("delta").option("versionAsOf",0).load(silver\_path)  df\_old.display()  **Recap of Benefits**  You **never fail ingestion** if schema drifts. You **control and enforce schema** where you want (Silver). You can **backfill or reprocess** easily. |