(https://databricks.com)

Problem description

Starting a structured stream when the specified checkpoint location does **not yet** exist works fine
Restarting that same stream when the checkpoint location **DOES** exist is no longer possible, while that is the whole point of having a checkpoint location.
This worked fine up to ~2024-09-10 when the issue was first detected on our environment while restarting running streams.

DBR 14.3 LTS

Demo

3

```
# The source table is an empty delta table for this test.
# It can be any source. The same issue occurs when streaming from Azure Event Hub, for example.
src_table = "bronze_dev.ignition.historian"
spark.sql(f"delete from {src_table}")
spark.table(src_table).printSchema()
```

```
root
 |-- ingest_partition: long (nullable = true)
 |-- eh_id: string (nullable = true)
 |-- id: string (nullable = true)
 |-- eh_endpoint: string (nullable = true)
 |-- eh_namespace: string (nullable = true)
 |-- eh_name: string (nullable = true)
 |-- ingest_ts: timestamp (nullable = true)
 |-- partition: integer (nullable = true)
 |-- offset: long (nullable = true)
 |-- data: string (nullable = true)
 |-- insert_id: long (nullable = true)
 |-- update_id: long (nullable = true)
 |-- insert_ts: timestamp (nullable = true)
 |-- update_ts: timestamp (nullable = true)
```

4

```
# Functions used in this demo.
from pyspark.sql import DataFrame

# The checkpoint directory, located in an azure storage account made available through unity catalog.
checkpoint_dir = f"abfss://checkpoint@xxx.dfs.core.windows.net/streams/test_dwight_checkpointing"

def noop(df: DataFrame, batch_id: int):
    """No-op function that will be called by the structured stream.
    Takes the input df and does nothing really."""
    df.write.format("noop")

def start_stream():
    """Starts streaming from the source table."""
    (
        spark.readStream
        .format("delta")
        .options(**{
            "ignoreDeletes": "true",
            "skipChangeCommits": "true",
            "readChangeData": "false"
        })
        .table("bronze_dev.ignition.historian")
        .writeStream.format("delta")
        .foreachBatch(noop)
        .option("checkpointLocation", checkpoint_dir)
        .start()
    )
```

5

```
# Start the stream while the checkpoint directory doesn't exist.
dbutils.fs.rm(checkpoint_dir, True)
start_stream()
# stream starts fine = OK
```

▸ 3025ff5b-2b53-42d8-8f08-c98c2dd4ed14 *Last updated: 29 minutes ago*

6

```
# cancel the running command above to stop the stream
# restart the stream once stopped, this time NOT removing the checkpoint dir
start_stream()
# stream fails with a FileAlreadyExistsException = NOK
```

7