```
#Define schema for our sample streaming
from pyspark.sql.types import
StructType,StructField,IntegerType,StringType
schema_defined=StructType([StructField('File',StringType(),True),
                           StructField('Shop',StringType(),True),

StructField('Sale_count',IntegerType(),True)
                           ])

dbutils.fs.mkdirs("/FileStore/tables/stream_checkpoint/")
dbutils.fs.mkdirs("/FileStore/tables/stream_read/")
dbutils.fs.mkdirs("/FileStore/tables/stream_write/")
```

```
True
```

```
df = spark.readStream \
    .format("csv") \
    .schema(schema_defined) \
    .option('header', True) \
    .option("sep", ",") \
    .load('/FileStore/tables/stream_read/')

df1 = df.groupBy('Shop').sum('Sale_count')

dbutils.fs.ls("/FileStore/tables/stream_read/")
```

```
[FileInfo(path='dbfs:/FileStore/tables/stream_read/
online_sales_1.csv', name='online_sales_1.csv', size=338,
modificationTime=1751523041000),

FileInfo(path='dbfs:/FileStore/tables/stream_read/online_sales_2.csv',
name='online_sales_2.csv', size=338, modificationTime=1751523041000),

FileInfo(path='dbfs:/FileStore/tables/stream_read/online_sales_3.csv',
name='online_sales_3.csv', size=338, modificationTime=1751523041000),

FileInfo(path='dbfs:/FileStore/tables/stream_read/online_sales_4.csv',
name='online_sales_4.csv', size=332, modificationTime=1751523041000),

FileInfo(path='dbfs:/FileStore/tables/stream_read/online_sales_5.csv',
name='online_sales_5.csv', size=335, modificationTime=1751523041000)]
```

```
dbutils.fs.ls("/FileStore/tables/stream_read/")
```

```
[FileInfo(path='dbfs:/FileStore/tables/stream_read/
online_sales_1.csv', name='online_sales_1.csv', size=338,
modificationTime=1751523041000),

FileInfo(path='dbfs:/FileStore/tables/stream_read/online_sales_2.csv',
name='online_sales_2.csv', size=338, modificationTime=1751523041000),

FileInfo(path='dbfs:/FileStore/tables/stream_read/online_sales_3.csv',
```

```
name='online_sales_3.csv', size=338, modificationTime=1751523041000),

FileInfo(path='dbfs:/FileStore/tables/stream_read/online_sales_4.csv',
name='online_sales_4.csv', size=332, modificationTime=1751523041000),

FileInfo(path='dbfs:/FileStore/tables/stream_read/online_sales_5.csv',
name='online_sales_5.csv', size=335, modificationTime=1751523041000)]
```

```
df4 = df.writeStream \
    .format("parquet") \
    .outputMode("append") \
    .option("path", "/FileStore/tables/stream_write/") \
    .option("checkpointLocation",
"/FileStore/tables/stream_checkpoint/") \
    .start()

display(
    spark.read.format("parquet")
    .load("/FileStore/tables/stream_write/")
)

df_result = spark.read.parquet("/FileStore/tables/stream_write/")
df_result.show()
```

```
+----+-------------------+----------+
|File|               Shop|Sale_count|
+----+-------------------+----------+
|1011|2025-07-01 10:50:00|       500|
|1012|2025-07-01 10:46:00|       501|
|1013|2025-07-01 09:27:00|       502|
|1014|2025-07-01 09:20:00|       503|
|1015|2025-07-01 10:50:00|       504|
|1021|2025-07-01 09:30:00|       500|
|1022|2025-07-01 10:01:00|       501|
|1023|2025-07-01 09:51:00|       502|
|1024|2025-07-01 10:37:00|       503|
|1025|2025-07-01 09:46:00|       504|
|1001|2025-07-01 10:48:00|       500|
|1002|2025-07-01 10:05:00|       501|
|1003|2025-07-01 10:46:00|       502|
|1004|2025-07-01 10:54:00|       503|
|1005|2025-07-01 10:36:00|       504|
|1006|2025-07-01 09:17:00|       500|
|1007|2025-07-01 10:03:00|       501|
|1008|2025-07-01 10:40:00|       502|
|1009|2025-07-01 10:54:00|       503|
|1010|2025-07-01 10:38:00|       504|
+----+-------------------+----------+
only showing top 20 rows
```