ANNAPOORNIMA S

SMA Lab 12: Performing ETL using Databricks

225229101

In [0]:
```
# Import functions
from pyspark.sql.functions import col, current_timestamp

# Define variables used in code below
file_path = "/databricks-datasets/structured-streaming/events"
username = spark.sql("SELECT regexp_replace(current_user(), '[^a-zA-Z0-9]', '_'
table_name = f"{username}_etl_quickstart"
checkpoint_path = f"/tmp/{username}/_checkpoint/etl_quickstart"

# Clear out data from previous demo execution
spark.sql(f"DROP TABLE IF EXISTS {table_name}")
dbutils.fs.rm(checkpoint_path, True)

# Configure Auto Loader to ingest JSON data to a Delta table
(spark.readStream
  .format("cloudFiles")
  .option("cloudFiles.format", "json")
  .option("cloudFiles.schemaLocation", checkpoint_path)
  .load(file_path)
  .select("*", col("_metadata.file_path").alias("source_file"), current_timesta
  .writeStream
  .option("checkpointLocation", checkpoint_path)
  .trigger(availableNow=True)
  .toTable(table_name))
```

```
Out[1]: <pyspark.sql.streaming.query.StreamingQuery at 0x7ff5a4cfe820>
```

In [0]:
```
df = spark.read.table(table_name)
```

In [0]:
```
display(df)
```

action	time	_rescued_data	source_file	processing_time
Close	1469679568	null	/databricks-datasets/structured-streaming/events/file-49.json	2023-09-11T16:40:44.958+0000
Close	1469679568	null	/databricks-datasets/structured-streaming/events/file-49.json	2023-09-11T16:40:44.958+0000
Open	1469679569	null	/databricks-datasets/structured-streaming/events/file-49.json	2023-09-11T16:40:44.958+0000
Close	1469679571	null	/databricks-datasets/structured-streaming/events/file-49.json	2023-09-11T16:40:44.958+0000
Open	1469679571	null	/databricks-datasets/structured-streaming/events/file-49.json	2023-09-11T16:40:44.958+0000
Open	1469679571	null	/databricks-datasets/structured-streaming/events/file-49.json	2023-09-11T16:40:44.958+0000
Close	1469679572	null	/databricks-datasets/structured-streaming/events/file-49.json	2023-09-11T16:40:44.958+0000
Close	1469679573	null	/databricks-datasets/structured-	2023-09-