# **ETL Flow Documentation for Data Analytics Infrastructure**

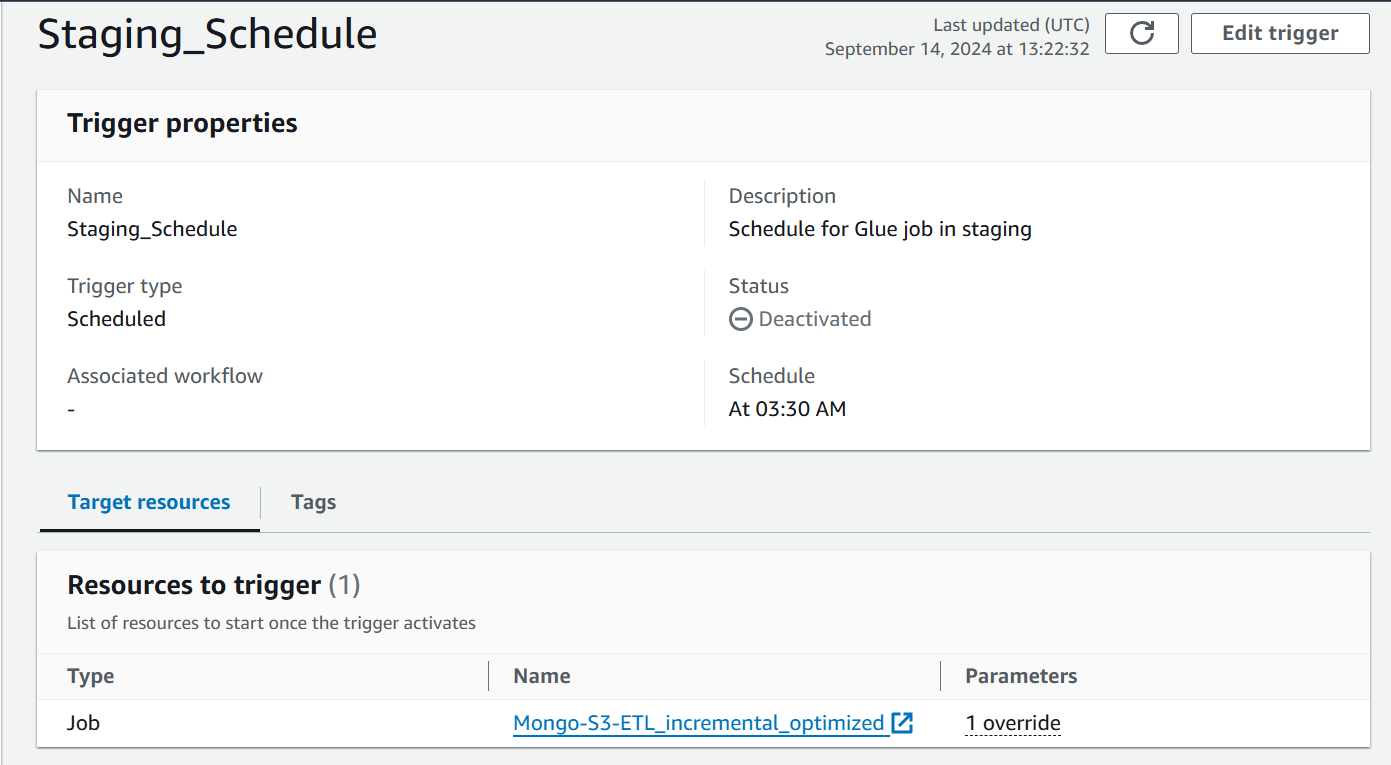
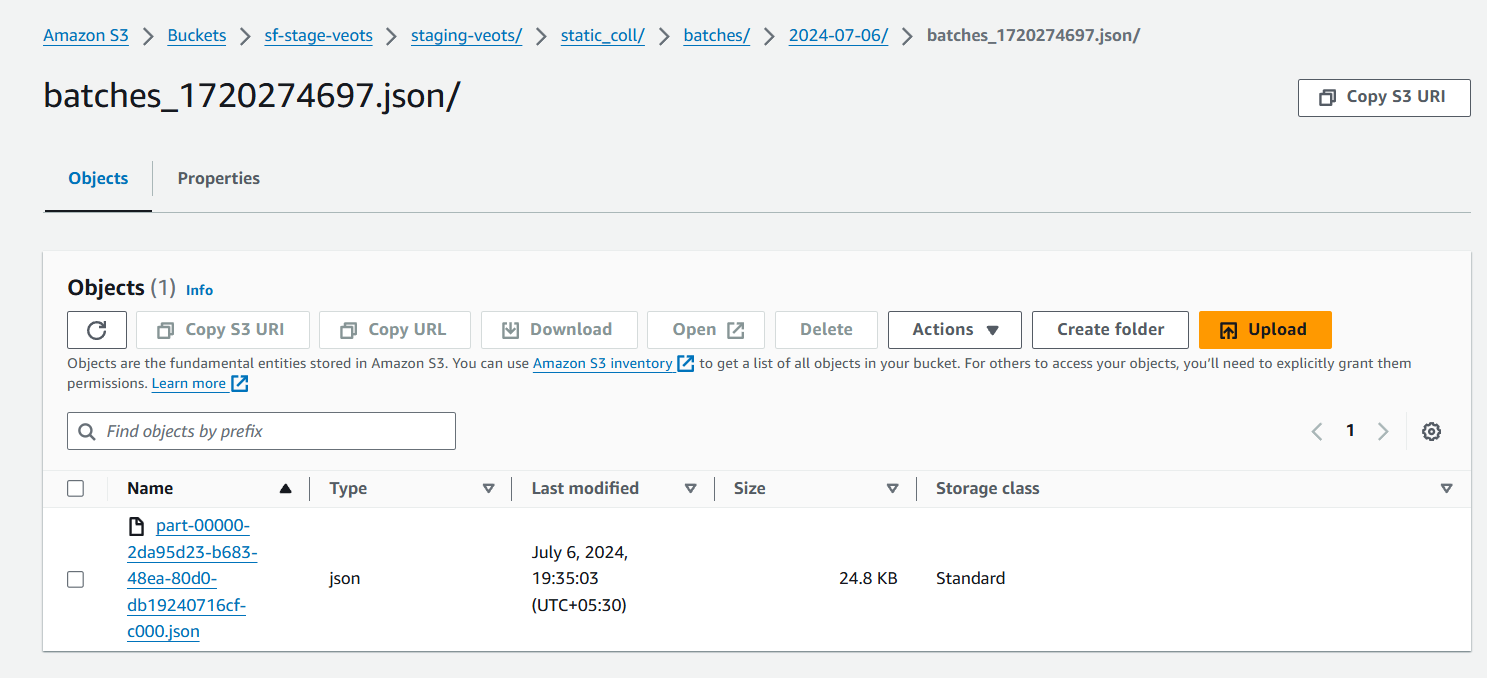
## **1. Introduction**

* This is the documentation to explain the entire flow of the data analytics ETL and automation.
* We have configured the Schedule for AWS Glue job daily using AWS triggers so that the new data gets processed everyday once.
* Once the Glue job runs, we have configured Stream and task in snowflake to Load Data in the Data Warehouse (Snowflake).
* We load data into snowflake tables in a append fashion. And we have created views for analytics which take data from these tables.
* These views are pulled into Power BI and the reports are created.

## **2. Data Sources**

* MongoDB is the only data Source in our infrastructure.
* We fetch the data from the source daily in an incremental fashion where only the new data which has arrived after the load time of our last Data Refresh is processed.

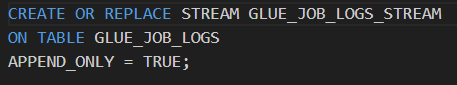
## **3. Data Ingestion Process**

* **Data Collection Mechanism**:
  + We are loading the data from source (MongoDB) into AWS S3 bucket (sf-stage-veots) in JSON format.
  + We are not doing any transformations in the data in this part of the process, as we are taking the ELT approach where we do the transformations on the data after loading it into Data Warehouse.
  + Refer to [Final Optimized ETL job.docx](https://ksjrudhprivatelimited-my.sharepoint.com/:w:/r/personal/rajat_sawant_veots_com/_layouts/15/Doc.aspx?sourcedoc=%7B78278115-4CD2-4A56-AAD2-38D3E4457DBD%7D&file=Final%20Optimized%20ETL%20job.docx&action=default&mobileredirect=true&DefaultItemOpen=1&ct=1726319758907&wdOrigin=OFFICECOM-WEB.START.EDGEWORTH&cid=c46d781d-ad19-4620-aca5-e031f6c120ee&wdPreviousSessionSrc=HarmonyWeb&wdPreviousSession=d7e45120-510c-464a-a3c2-164ba8d2cedc) for detailed explanation on the ETL job.
* **Triggering Mechanism**:
  + To schedule the Glue Job, we have used AWS Triggers which runs the glue job at a scheduled time every day.  
    
  + The above trigger runs the Glue job at 3:30 AM (UTC) that is 9:00 AM (IST)
* **Storage in AWS**:
  + The data extracted from MongoDB is stored in AWS S3 bucket (sf-stage-veots).
  + The folder structure in S3 is [sf-stage-veots]/[database name]/[static/dynamic collection]/[collection name]/[Date]/[File name]/partitions  
    In the above example, partitions of batches table are stored. In this case we only have one partition.

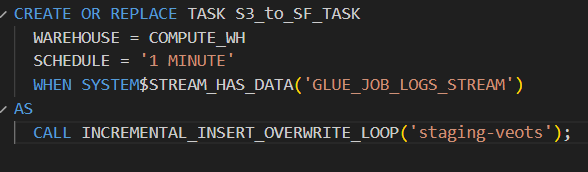
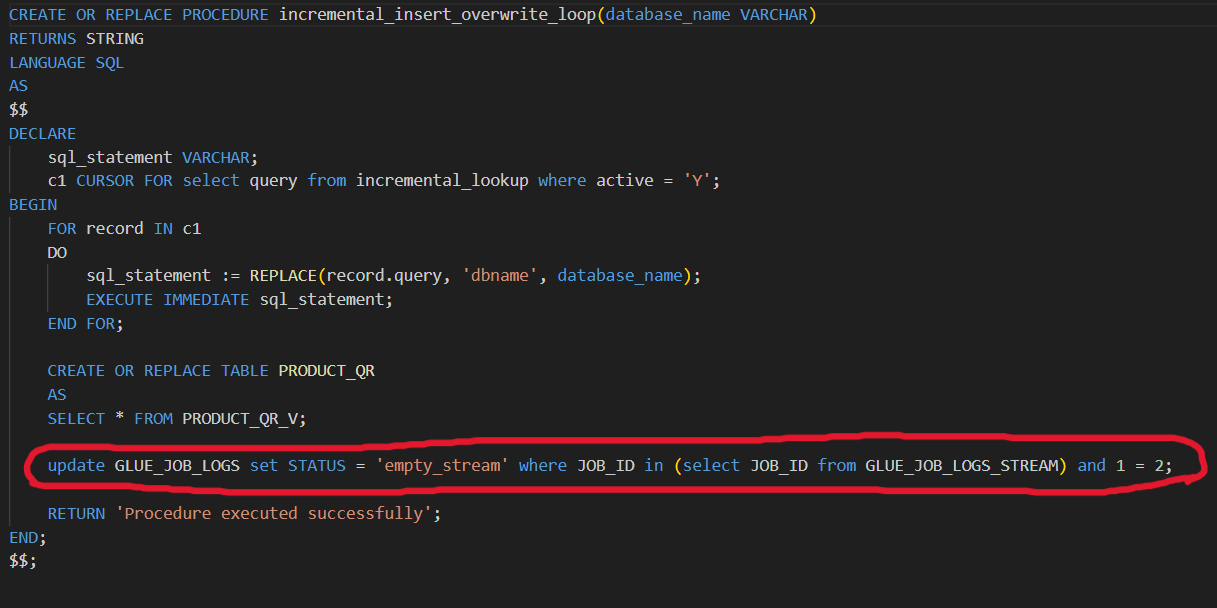
## **4. Data Transformation & Processing**

* **Transformation Process**:
  + We are not doing any transformation on the data in the Glue job, except for the filtering of the new records from the MongoDB for incremental load.
  + Data is stored in S3 bucket in JSON format. This JSON Files are converted into Tables only after ingesting it into snowflake.

## **5. Data Loading into Snowflake**

* **Snowflake Staging**:
  + We create an external stage in Snowflake which points to the AWS S3 bucket that has the data that came from MongoDB.
  + We take this data into Snowflake tables using stored procedure (INCREMENTAL\_INSERT\_OVERWRITE\_LOOP). Detailed process explained in [Data\_Analytics\_Overview\_Documentation.docx](https://ksjrudhprivatelimited-my.sharepoint.com/:w:/r/personal/rajat_sawant_veots_com/_layouts/15/Doc.aspx?sourcedoc=%7BAA4039D1-A70A-4524-90DA-61DB5622F14F%7D&file=Data_Analytics_Overview_Documentation.docx&action=default&mobileredirect=true&DefaultItemOpen=1&ct=1726322421452&wdOrigin=OFFICECOM-WEB.START.EDGEWORTH&cid=c33e4395-99c2-48e9-9bad-73a88af7d0c4&wdPreviousSessionSrc=HarmonyWeb&wdPreviousSession=d7e45120-510c-464a-a3c2-164ba8d2cedc)
* **Using Streams**:
  + We have created a table in Snowflake (GLUE\_JOB\_LOGS) and created a stream (GLUE\_JOB\_LOGS\_STREAM) on the table.  
    
  + When the Glue Job completes it inserts a row in the GLUE\_JOB\_LOGS table with the details about the glue job.
  + We are using Append –only stream so that the stream only tracks changes only when the data is inserted.
  + The GLUE\_JOB\_LOGS\_STREAM stream tracks this change in the GLUE\_JOB\_LOGS table and shows the new data in the stream.

## **6. Task Automation in Snowflake**

* **Snowflake Tasks**:
  + We have created a task with a condition that if the stream has data the task should run the procedure which loads all that data that is present in S3 bucket and appends the data in snowflake tables.  
    
  + So, whenever the stream has data, the task will execute the procedure that loads all the data from S3 into snowflake tables.
  + The stream will have data only when any record is inserted in the GLUE\_JOB\_LOGS table. Record is inserted in the GLUE\_JOB\_LOGS table when the Glue job that loads data from MongoDB to S3 completes run.
* **Consuming Data from Stream:**
  + In summary, we are using streams and tasks to automatically load the data from S3 to snowflake when the Glue job loads data from MongoDB to S3.
  + Once the task is executed, we must empty the stream so that it can track the changes when the next glue job is run.
  + To empty the stream, in the end of the procedure (INCREMENTAL\_INSERT\_OVERWRITE\_LOOP) we are running a query that consumes the data from the stream.  
      
    This query is logically wrong. So it does not make any changes in database but the query runs successfully.
  + Once the update query in the end of the procedure runs successfully the stream is now empty.

## **7. End-to-End Data Flow with AWS and Snowflake**

* The data flow in our infrastructure begins with extracting incremental data from MongoDB, which is stored in AWS S3 in JSON format.
* The data ingestion process is automated using AWS Glue, scheduled to run daily via AWS Triggers. The Glue job extracts new records from MongoDB and places them into S3 for processing.
* Once the data lands in the S3 bucket, Snowflake's external stage connects to S3 and ingests the data using a stored procedure.
* The procedure is triggered by a Snowflake task, which is activated by a stream that tracks changes in the Glue job log table.
* This automation ensures that as soon as new data is loaded into S3, it is seamlessly ingested into Snowflake tables in an append-only fashion. From there, data is available in Snowflake views, which are used to power analytics dashboards in Power BI.
* Through this streamlined ETL process, we maintain a fully automated, scalable data pipeline that processes and analyzes daily updates efficiently.