Athena Querying for Parquet Files

1. \*\*Athena Table for Parquet Files\*\*  
 - Here's how to create the Athena table for your Parquet files:  
   
 ```sql  
 CREATE EXTERNAL TABLE IF NOT EXISTS your\_log\_table (  
 IGI\_LOG\_REFERENCE STRING,  
 IGI\_LOG\_TIME STRING,  
 IGI\_MACHINE\_NME STRING,  
 IGI\_MACHINE\_IP\_ADDRESS STRING,  
 IGI\_LOG\_TEXT STRING,  
 IGI\_LOG\_LEVEL STRING  
 )  
 PARTITIONED BY (year STRING, month STRING, day STRING)  
 STORED AS PARQUET  
 LOCATION 's3://your-bucket/log/';  
 ```  
   
2. \*\*Querying Parquet Files Based on Date\*\*  
 To query logs for a specific day, such as January 20th, 2025:  
   
 ```sql  
 SELECT IGI\_LOG\_REFERENCE, IGI\_LOG\_TIME, IGI\_MACHINE\_NME, IGI\_MACHINE\_IP\_ADDRESS, IGI\_LOG\_TEXT, IGI\_LOG\_LEVEL  
 FROM your\_log\_table  
 WHERE day = '2025-01-20' AND year = '2025' AND month = '01';  
 ```  
  
3. \*\*Querying by `IGI\_LOG\_REFERENCE`\*\*  
 To filter the logs by `IGI\_LOG\_REFERENCE`:  
   
 ```sql  
 SELECT IGI\_LOG\_REFERENCE, IGI\_LOG\_TIME, IGI\_MACHINE\_NME, IGI\_MACHINE\_IP\_ADDRESS, IGI\_LOG\_TEXT, IGI\_LOG\_LEVEL  
 FROM your\_log\_table  
 WHERE day = '2025-01-20' AND year = '2025' AND month = '01'  
 AND IGI\_LOG\_REFERENCE = 'specific-reference-id';  
 ```  
  
4. \*\*Querying Parquet Files by Time Range (Optional)\*\*  
 To filter logs by a time range, for example:  
   
 ```sql  
 SELECT IGI\_LOG\_REFERENCE, IGI\_LOG\_TIME, IGI\_MACHINE\_NME, IGI\_MACHINE\_IP\_ADDRESS, IGI\_LOG\_TEXT, IGI\_LOG\_LEVEL  
 FROM your\_log\_table  
 WHERE day = '2025-01-20' AND year = '2025' AND month = '01'  
 AND IGI\_LOG\_TIME >= '2025-01-20T10:00:00' AND IGI\_LOG\_TIME <= '2025-01-20T12:00:00';  
 ```  
  
5. \*\*Querying Specific Files in the `log` Folder\*\*  
 If you want to query only specific files from the `log` folder for a specific day:  
   
 ```sql  
 SELECT IGI\_LOG\_REFERENCE, IGI\_LOG\_TIME, IGI\_MACHINE\_NME, IGI\_MACHINE\_IP\_ADDRESS, IGI\_LOG\_TEXT, IGI\_LOG\_LEVEL  
 FROM your\_log\_table  
 WHERE "$path" LIKE 's3://your-bucket/log/2025/01/20/%'  
 AND IGI\_LOG\_REFERENCE = 'specific-reference-id';  
 ```  
  
6. \*\*Optimize Athena Querying with Partitioning and Parquet\*\*  
 - \*\*Partitioning by Date\*\*: Helps Athena read only the relevant data.  
 - \*\*Columnar format (Parquet)\*\*: Parquet files are efficient for querying.  
 - \*\*Compression\*\*: If your Parquet files are compressed, it further reduces the amount of data read.  
  
7. \*\*Summary of Queries\*\*  
 - \*\*Query logs for a specific day\*\*:  
 ```sql  
 SELECT \* FROM your\_log\_table WHERE day = '2025-01-20' AND year = '2025' AND month = '01';  
 ```  
  
 - \*\*Query logs by `IGI\_LOG\_REFERENCE`\*\*:  
 ```sql  
 SELECT \* FROM your\_log\_table WHERE day = '2025-01-20' AND year = '2025' AND month = '01' AND IGI\_LOG\_REFERENCE = 'specific-reference-id';  
 ```  
  
 - \*\*Query logs by `IGI\_LOG\_TIME` range\*\*:  
 ```sql  
 SELECT \* FROM your\_log\_table WHERE day = '2025-01-20' AND year = '2025' AND month = '01' AND IGI\_LOG\_TIME >= '2025-01-20T10:00:00' AND IGI\_LOG\_TIME <= '2025-01-20T12:00:00';  
 ```  
  
 - \*\*Query specific files from `log` folder\*\*:  
 ```sql  
 SELECT \* FROM your\_log\_table WHERE "$path" LIKE 's3://your-bucket/log/2025/01/20/%' AND IGI\_LOG\_REFERENCE = 'specific-reference-id';  
 ```  
  
8. \*\*Final Recommendations\*\*  
 - \*\*Use Partitioning\*\*: Partitioning by `year`, `month`, and `day` helps Athena optimize queries, especially when querying by date.  
 - \*\*Parquet Files\*\*: Ensure your logs are in \*\*Parquet\*\* format for better performance.  
 - \*\*Cost Optimization\*\*: Athena charges based on the amount of data scanned, so partitioning and using columnar formats (like Parquet) reduces costs.