**AWS Assignment**

1. Explain the difference between AWS Regions, Availability Zones, and Edge Locations. Why is this important for data analysis and latency-sensitive applications?

Answer:  
 **Region:** A geographical area (like us-east-1) with multiple Availability Zones.

* **Availability Zone (AZ):** A physically isolated data center in a region (like us-east-1a).
* **Edge Location:** Used by **CloudFront** to deliver low-latency content closer to users.

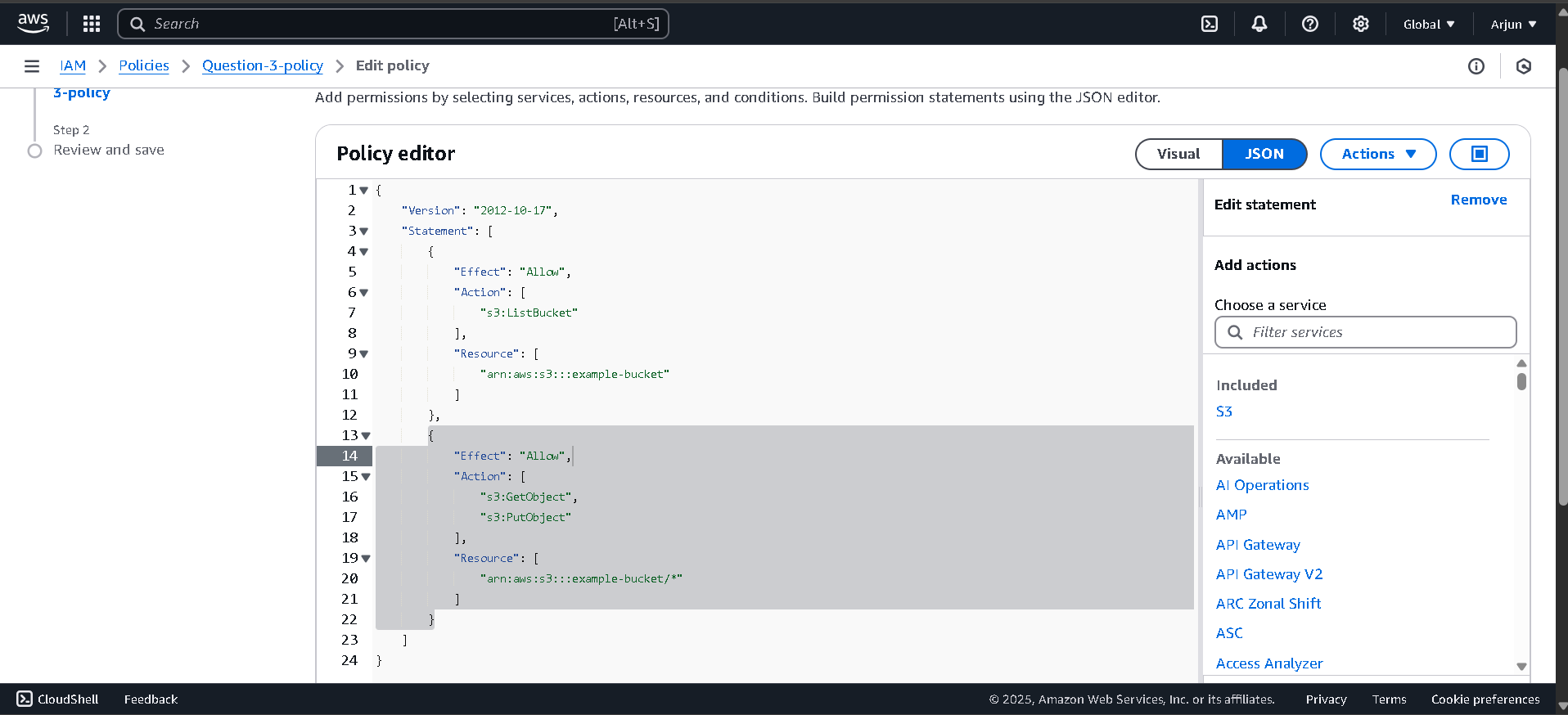
**Importance:**  
 Choosing correct regions and using edge locations reduces latency in data processing and analysis apps.

2. Using the AWS CLI, list all available AWS regions. Share the command used and the output.

Answer:  
**code:** aws ec2 describe-regions --query "Regions[\*].RegionName" --output tableScreenshot:  


3. Create a new IAM user with least privilege access to Amazon S3. Share your attached policies (JSON or screenshot)

Answer:  
 **AWS CLI Command to List All Regions**:



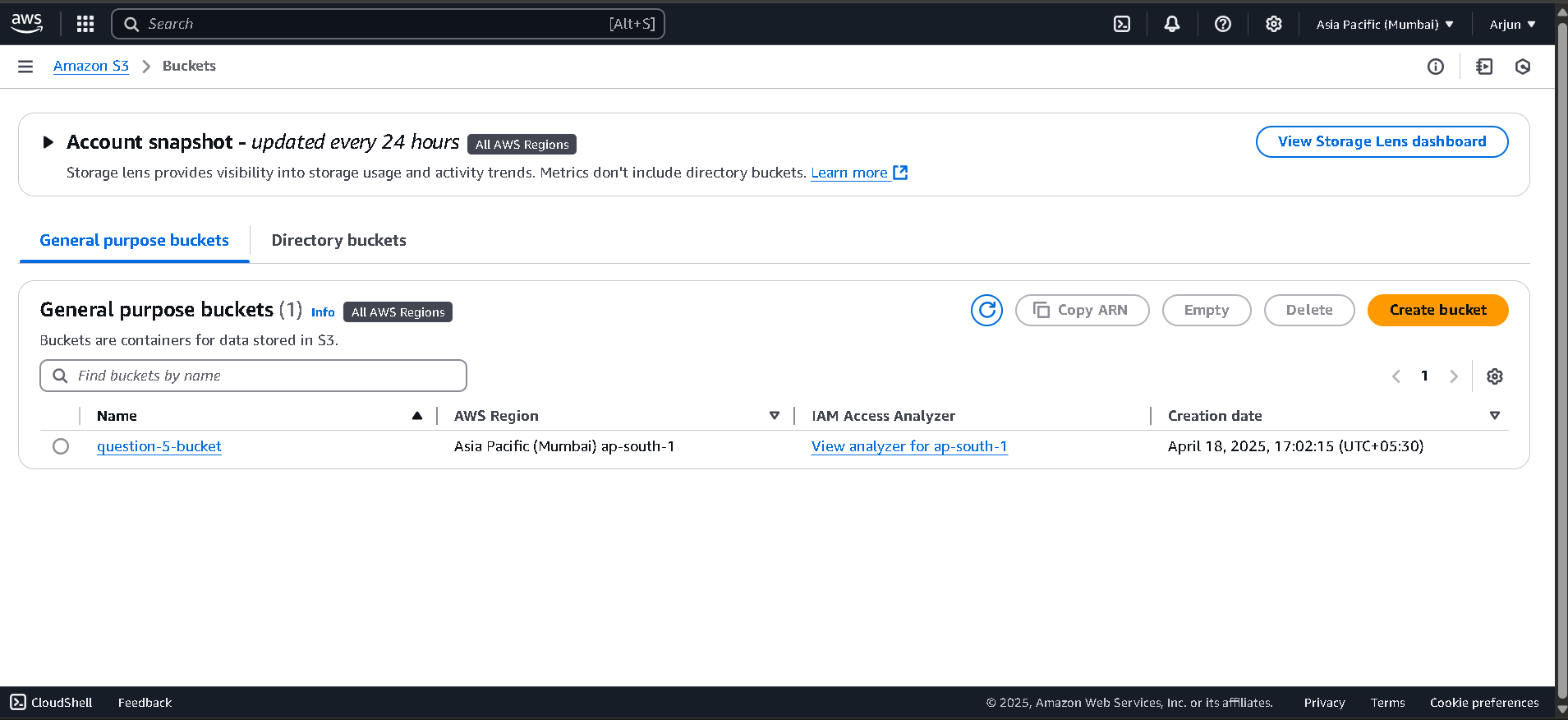
4. Compare different Amazon S3 storage (Standard, Intelligent-Tiering, Glacier). When should each be used in data analytics workflows?

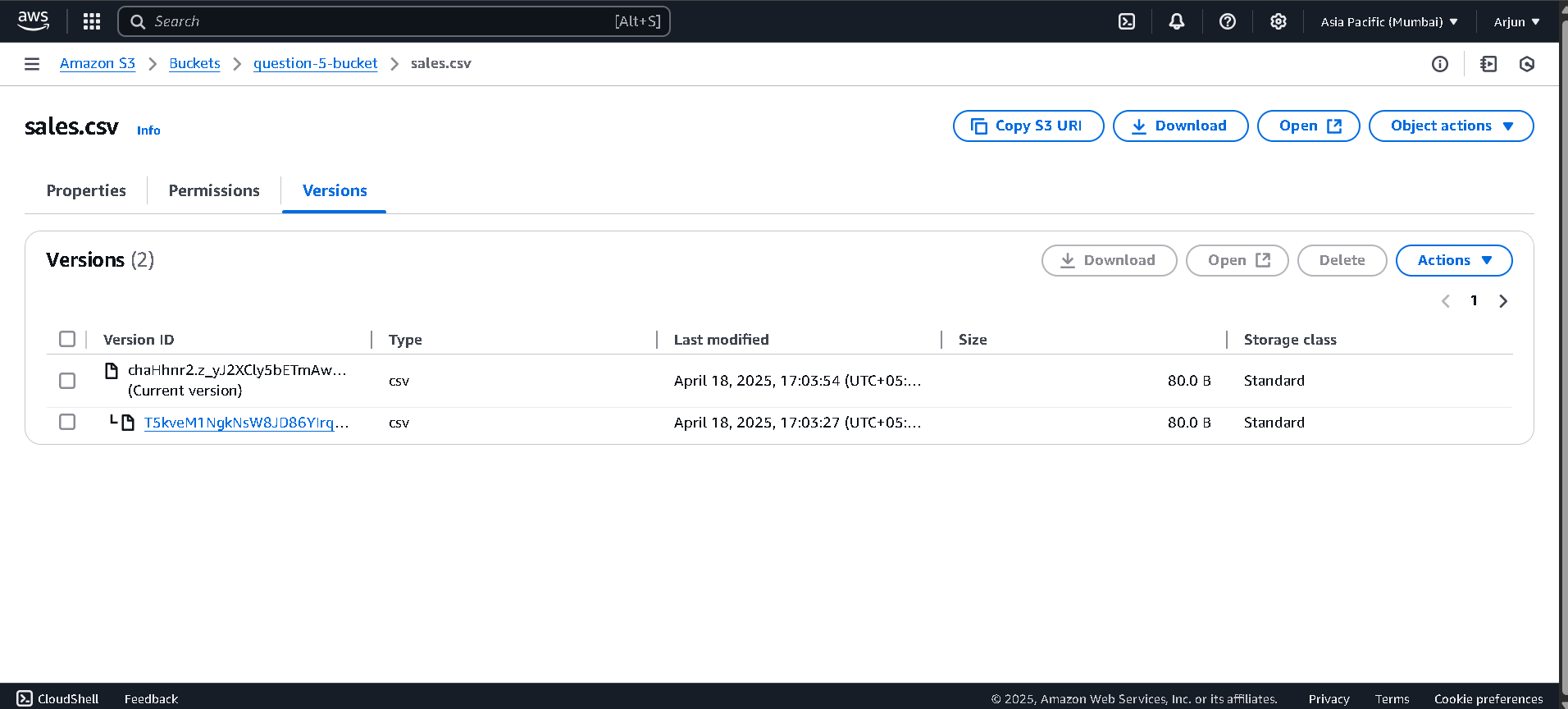
Answer:

| **Storage Class** | **Use Case in Data Analytics** |
| --- | --- |
| Standard | Frequently accessed datasets |
| Intelligent-Tiering | Unknown or changing access patterns |
| Glacier | Archive old log/data (not used often) |

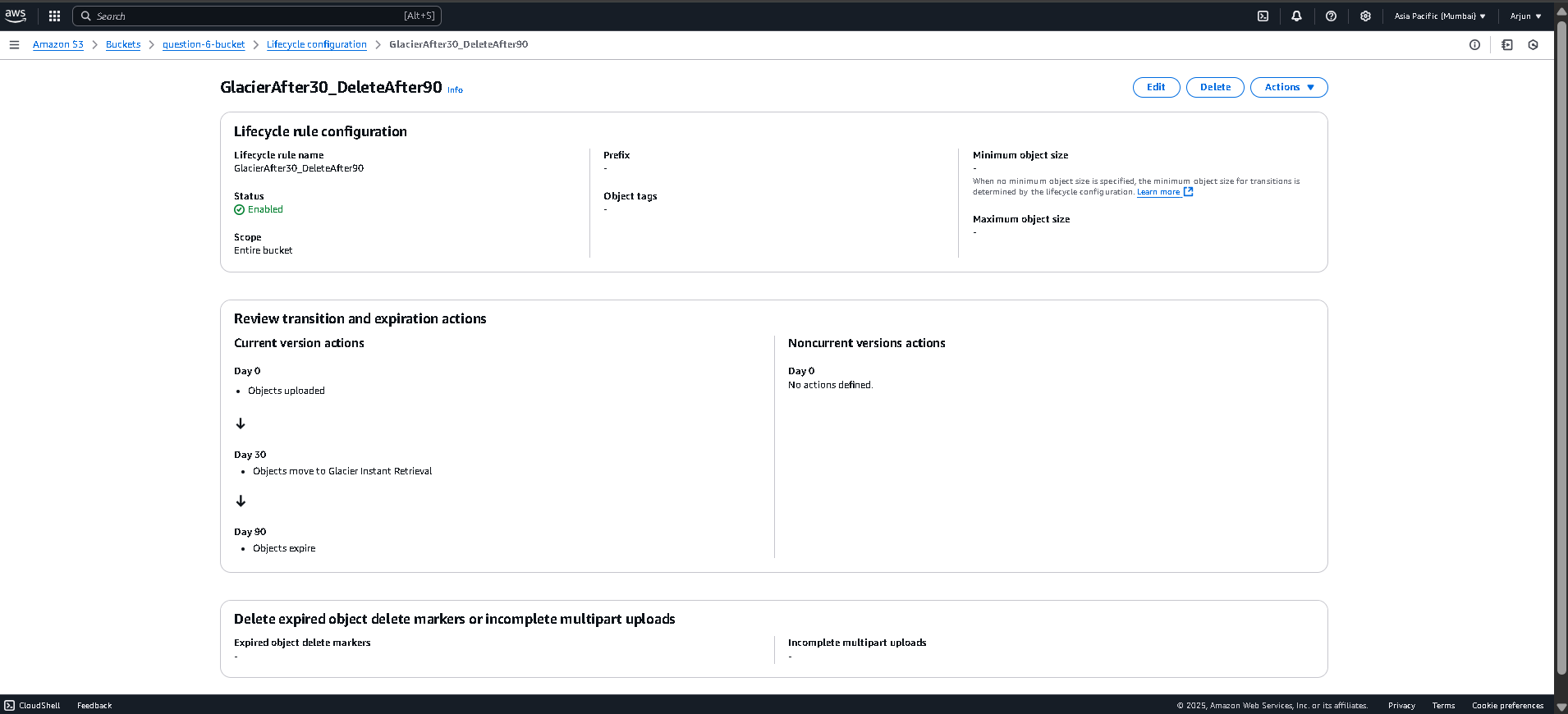
5. Create an S3 bucket and upload a sample dataset (CSV or JSON). Enable versioning and show at least two versions of one file

Answer:  
Attached bucket overview and two versions for the question-5-bucket:





6. Write and apply a lifecycle policy to move files to Glacier after 30 days and delete them after 90. Share the policy JSON or Screenshot

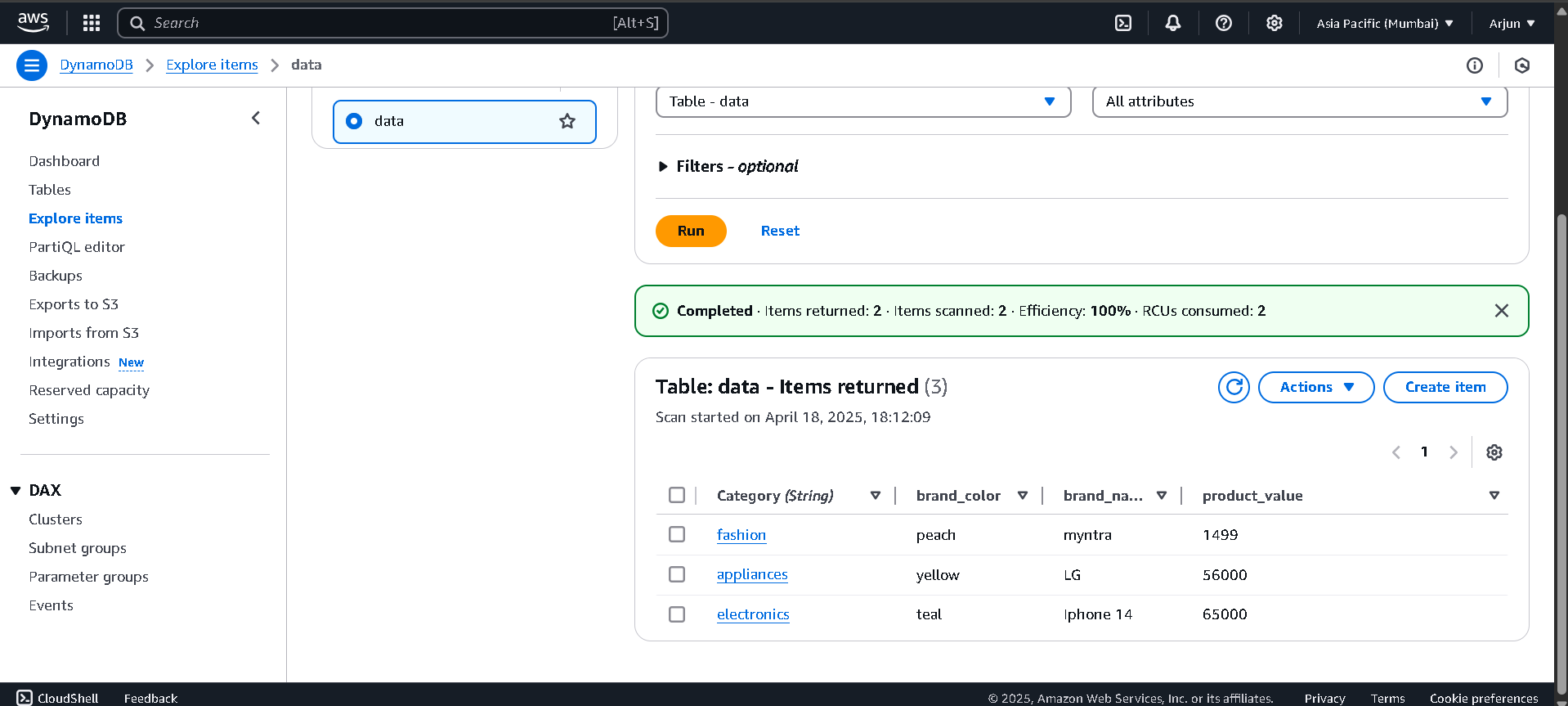
Answer:  
 

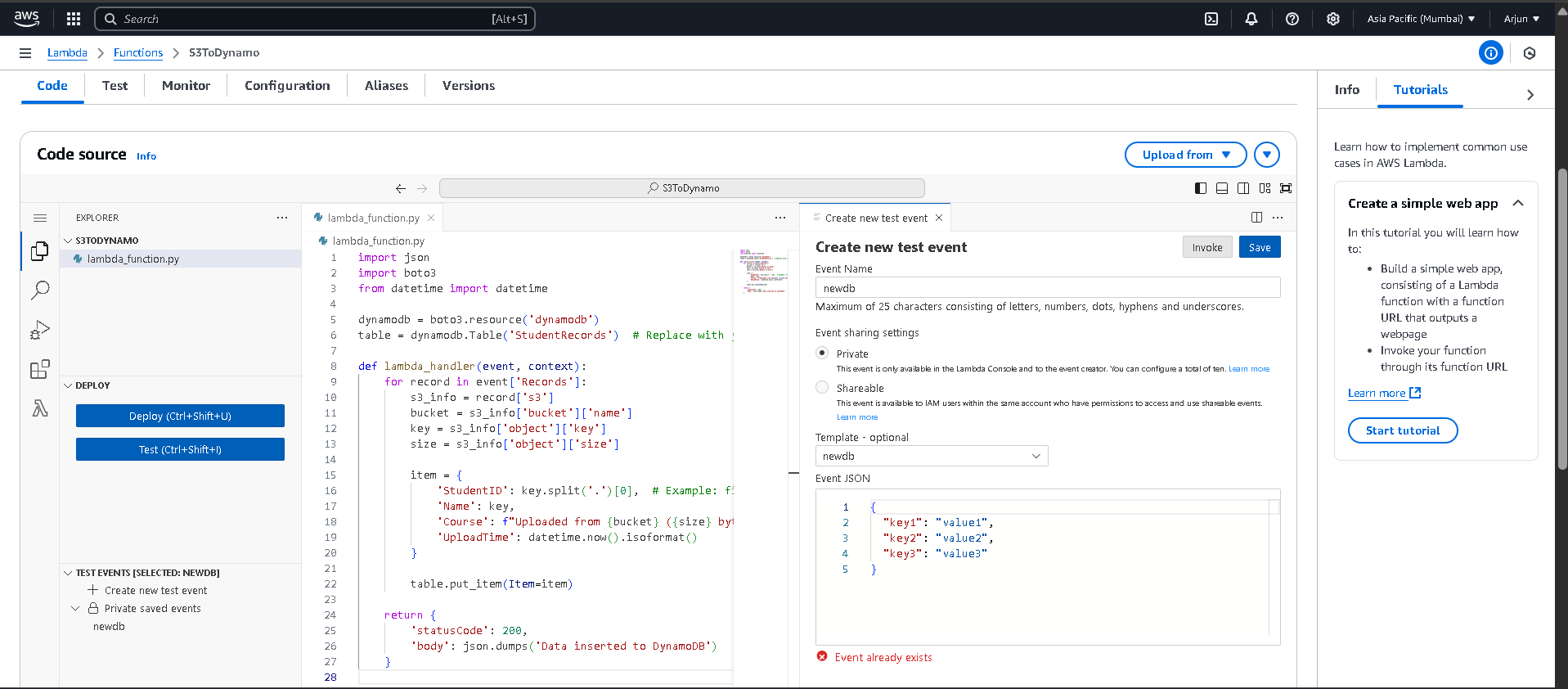
7. Compare RDS, DynamoDB, and Redshift for use in different stages of a data pipeline. Give one use case for each.

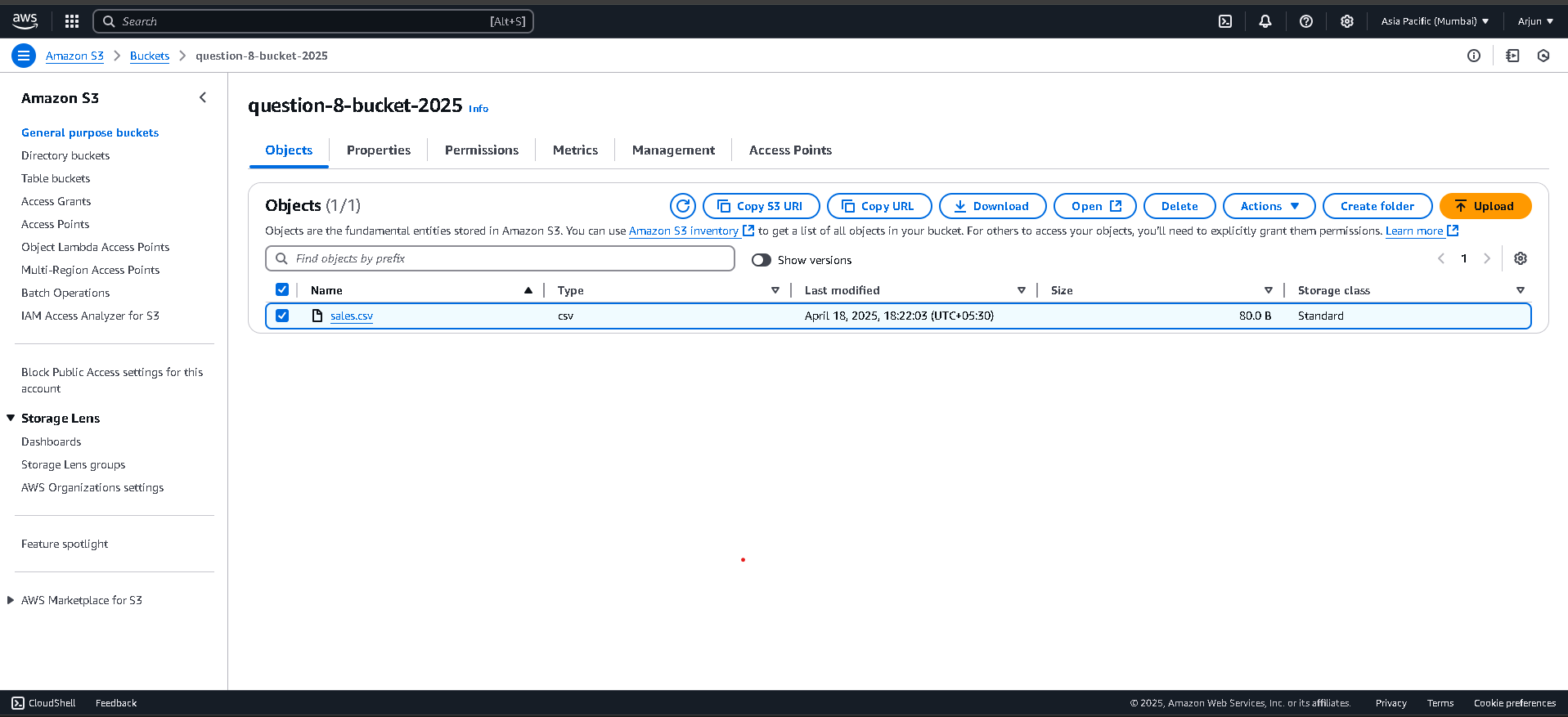
Answer:

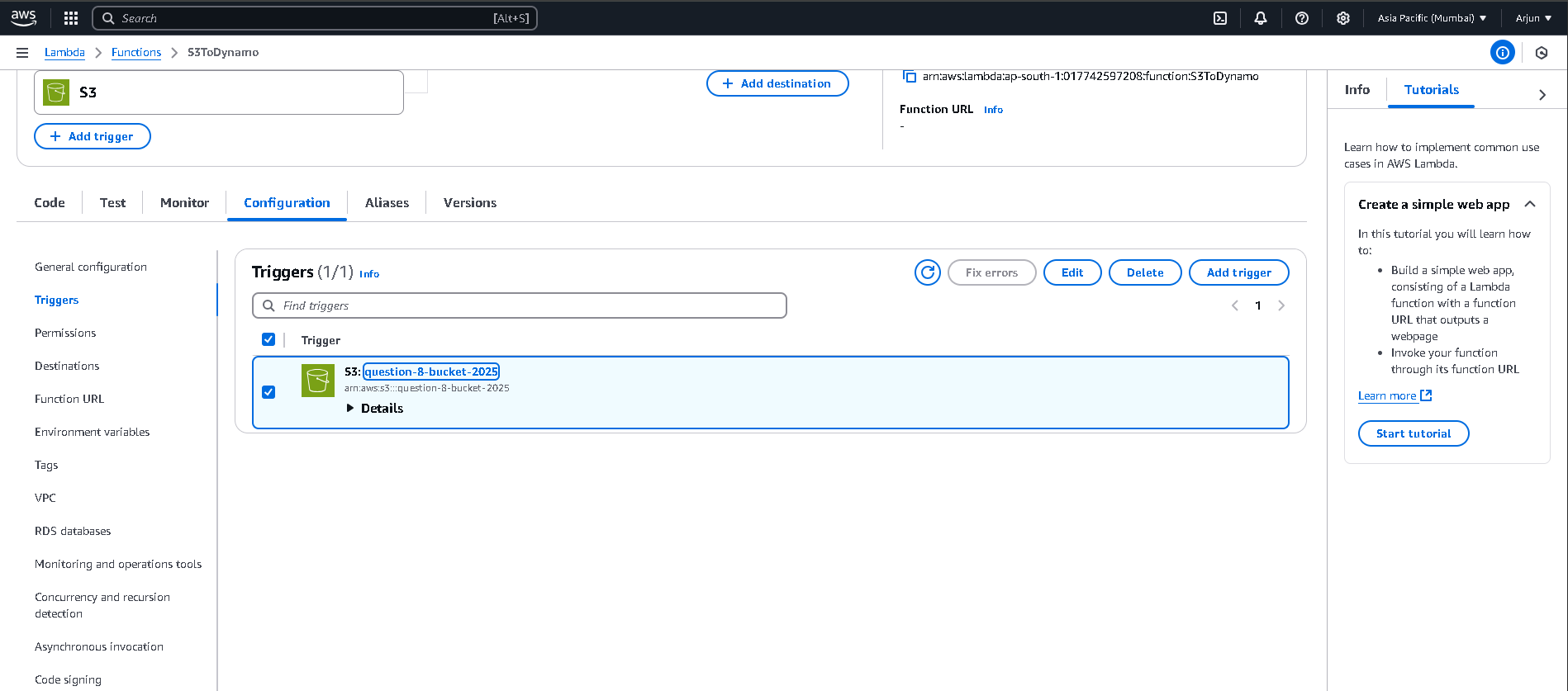
| **Service** | **Use Case** |
| --- | --- |
| RDS | Transactional DB (orders, users) |
| DynamoDB | Real-time user data (IoT, apps) |
| Redshift | Analytics/BI on large data sets |

8. Create a DynamoDB table and insert 3 records manually. Then write a Lambda function that adds records when triggered by S3 uploads.

Answer:  






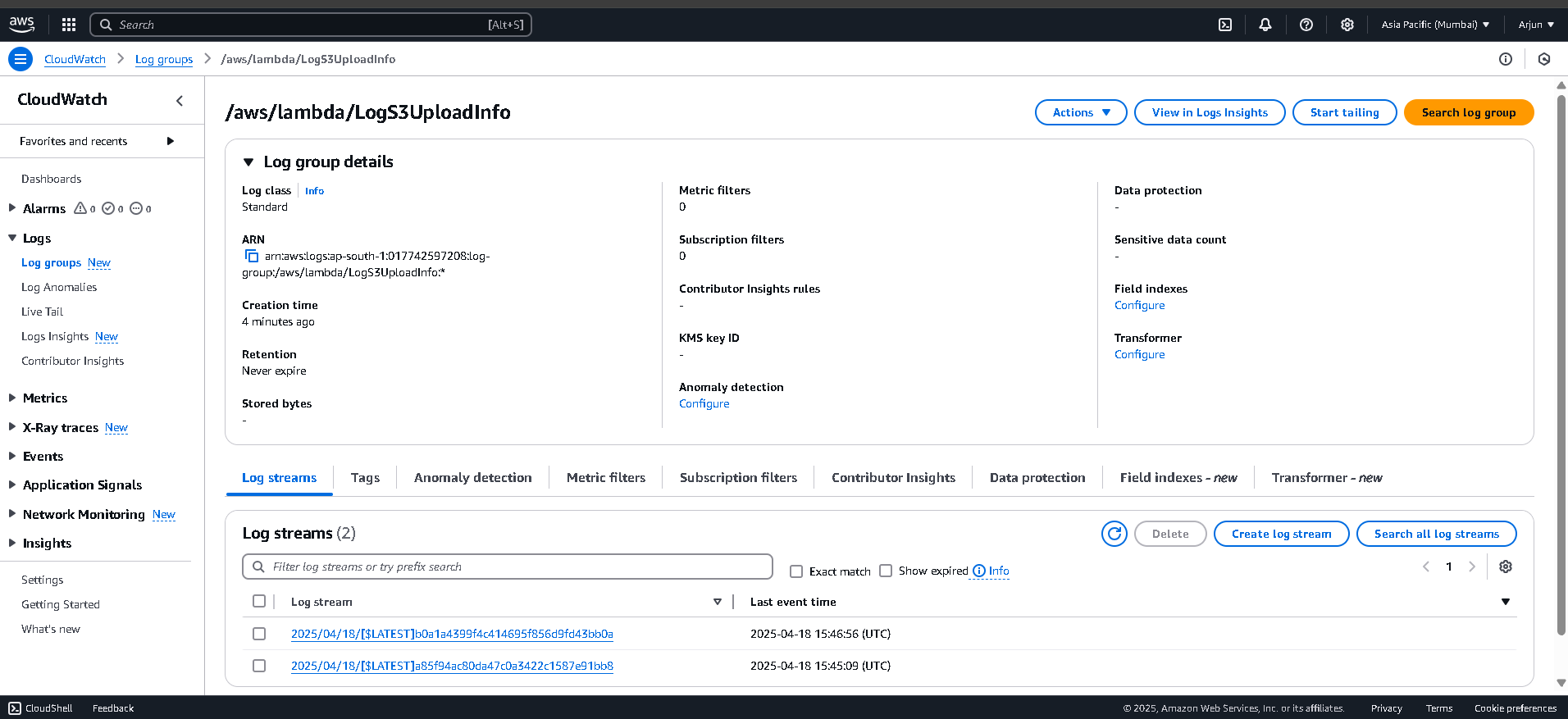


9. What is serverless computing? Discuss pros and cons of using AWS Lambda for data pipelines

Answer:

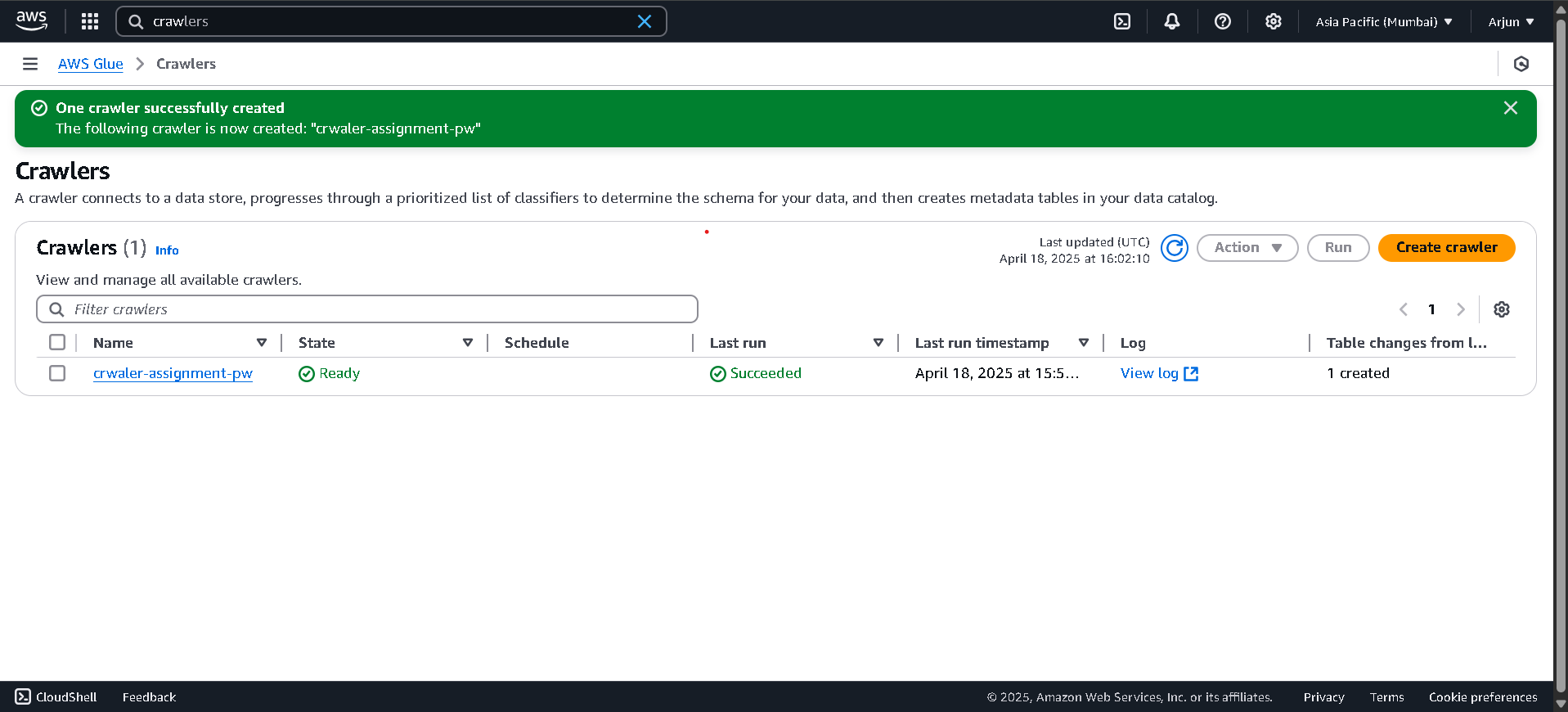
* Serverless computing**:** No need to manage servers; auto-scale and cost-efficient
* **Lambda Pros:** Scales automatically, good for event-driven tasks
* **Cons:** Limited execution time (15 min), debugging can be tough

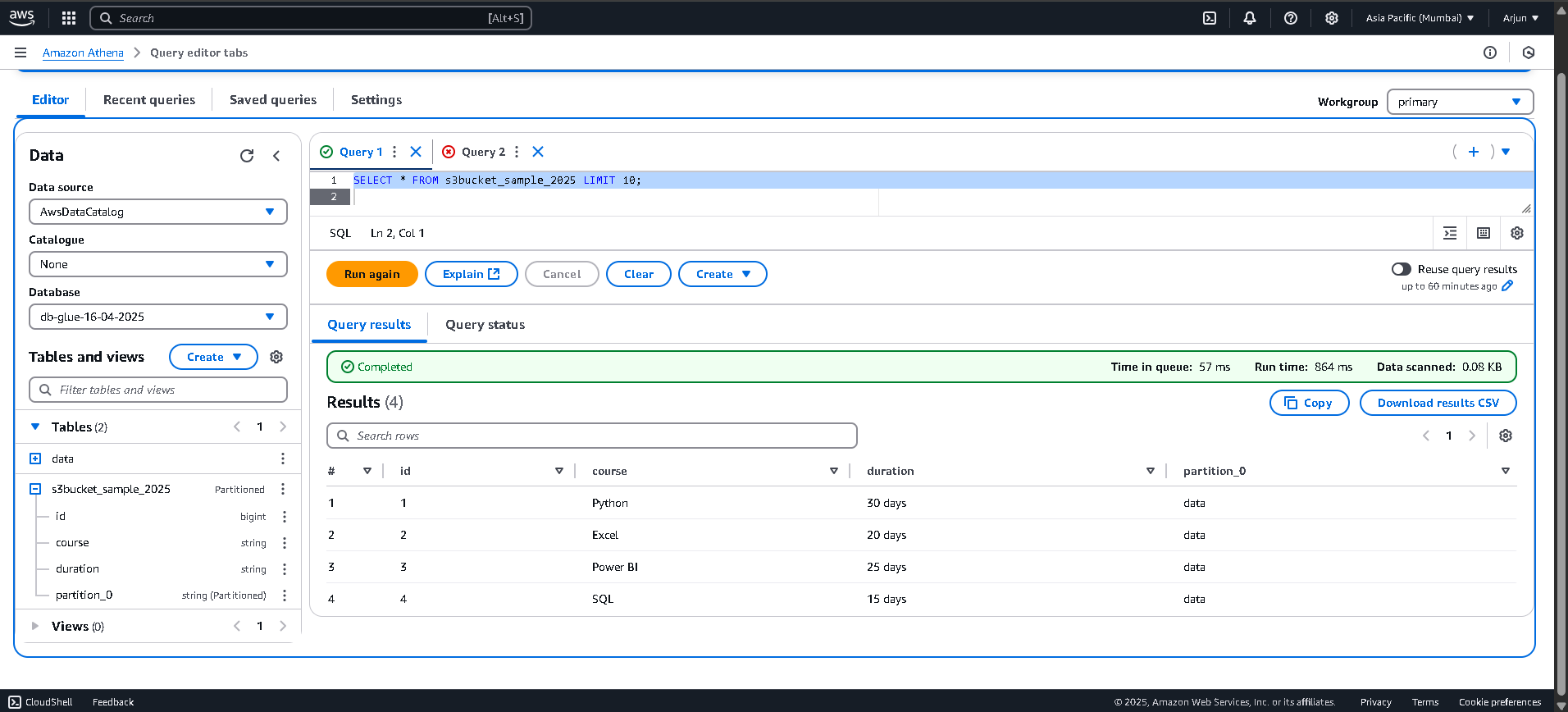
10. Create a Lambda function triggered by S3 uploads that logs file name, size, and timestamp to CloudWatch. Share code and a log screenshot

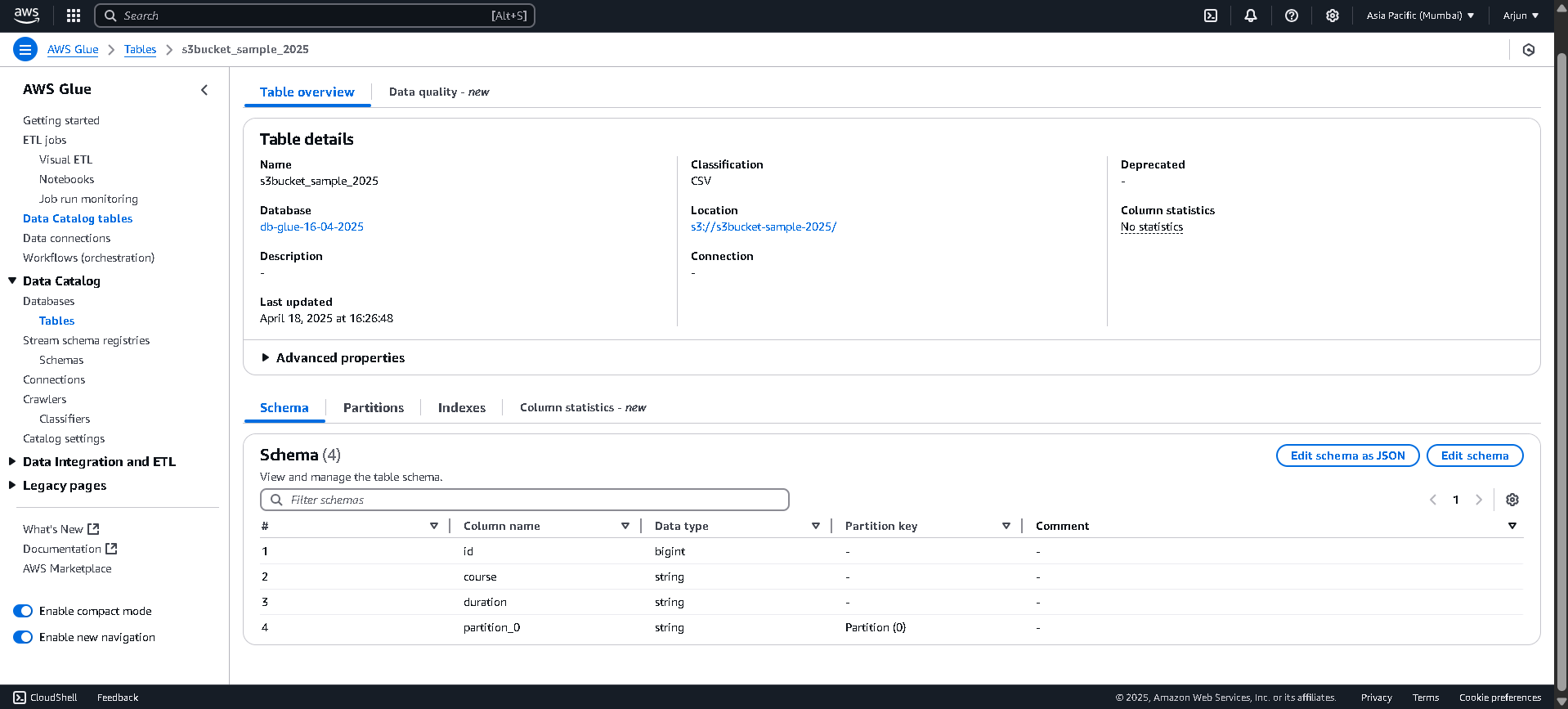
Answer:  


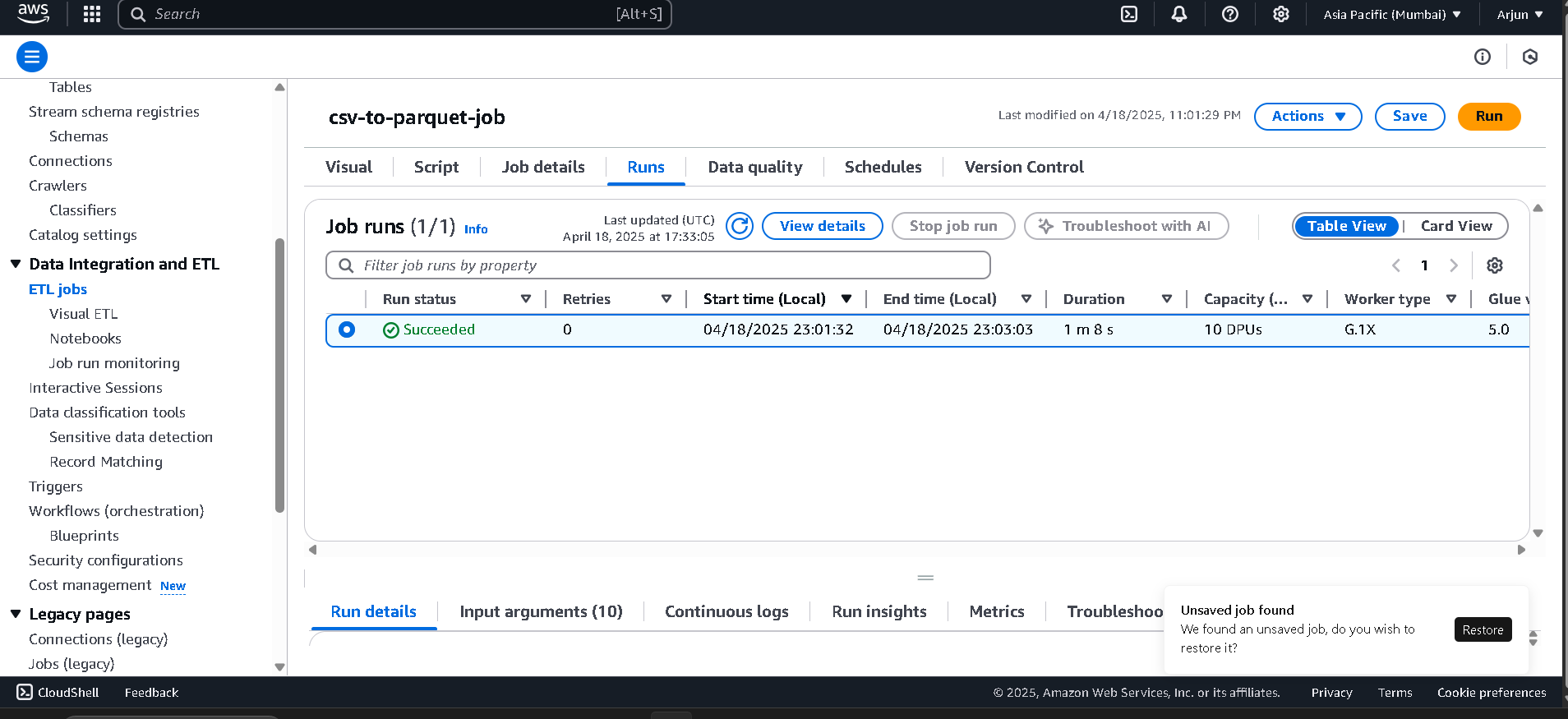
11. Use AWS Glue to crawl your S3 dataset, create a Data Catalog table, and run a Glue job to convert CSV data to parquet. Share job code and output location

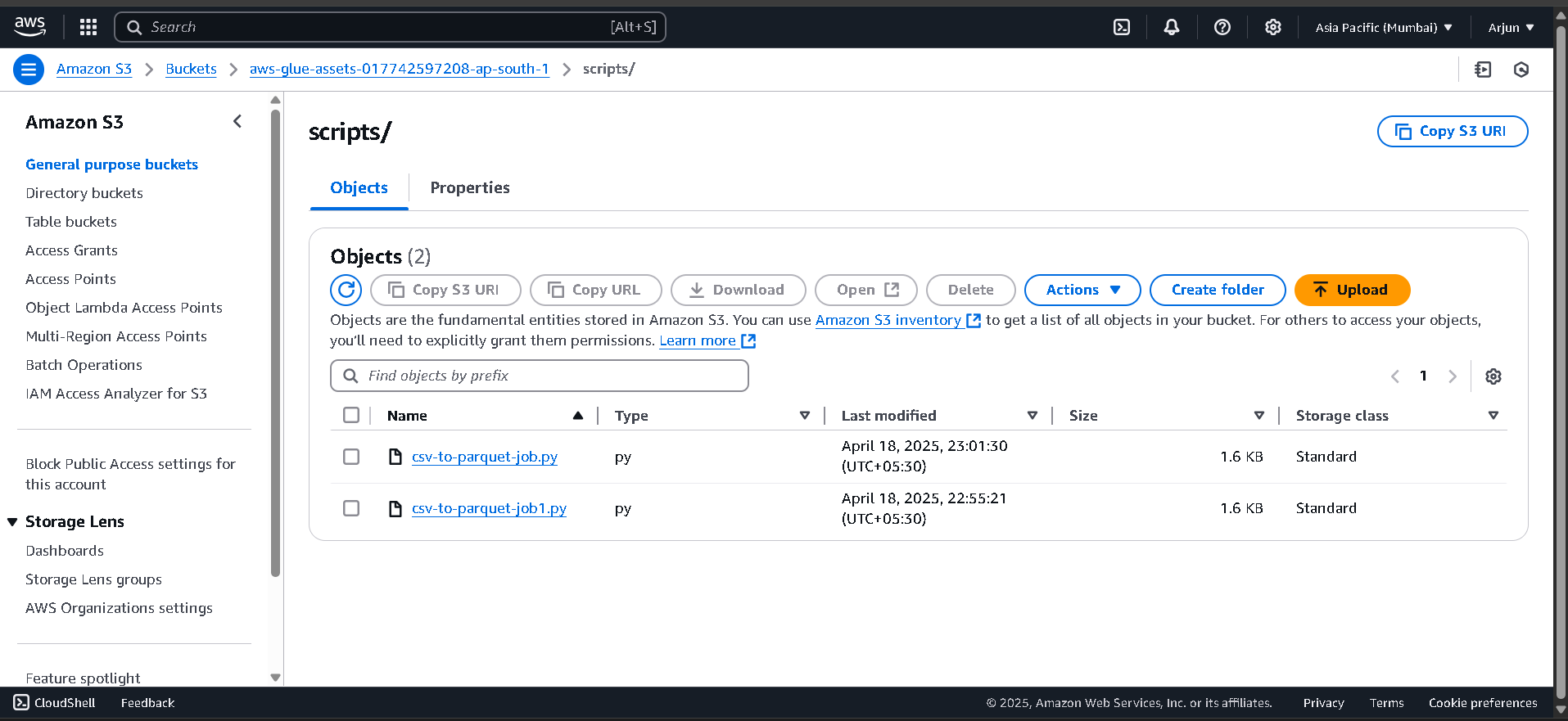
Answer:  
I uploaded a sample CSV to S3 and used AWS Glue Crawler to create a Data Catalog table. Then I created a Glue Job to convert the CSV into Parquet format and stored the result in another S3 folder. Parquet helps in better performance and is more cost-efficient for large-scale analytics.

Screenshot:  










12. Explain the difference between Kinesis Data Streams, Kinesis Firehose, and Kinesis Data Analytics. Provide a real-world example of how each would be used

Answer:

| **Service** | **Purpose** | **Example** |
| --- | --- | --- |
| Kinesis Data Streams | Real-time data ingestion | IoT data from sensors |
| Kinesis Firehose | Stream to destinations | Log data to S3 or Redshift |
| Kinesis Data Analytics | Real-time SQL processing | Fraud detection in transactions |

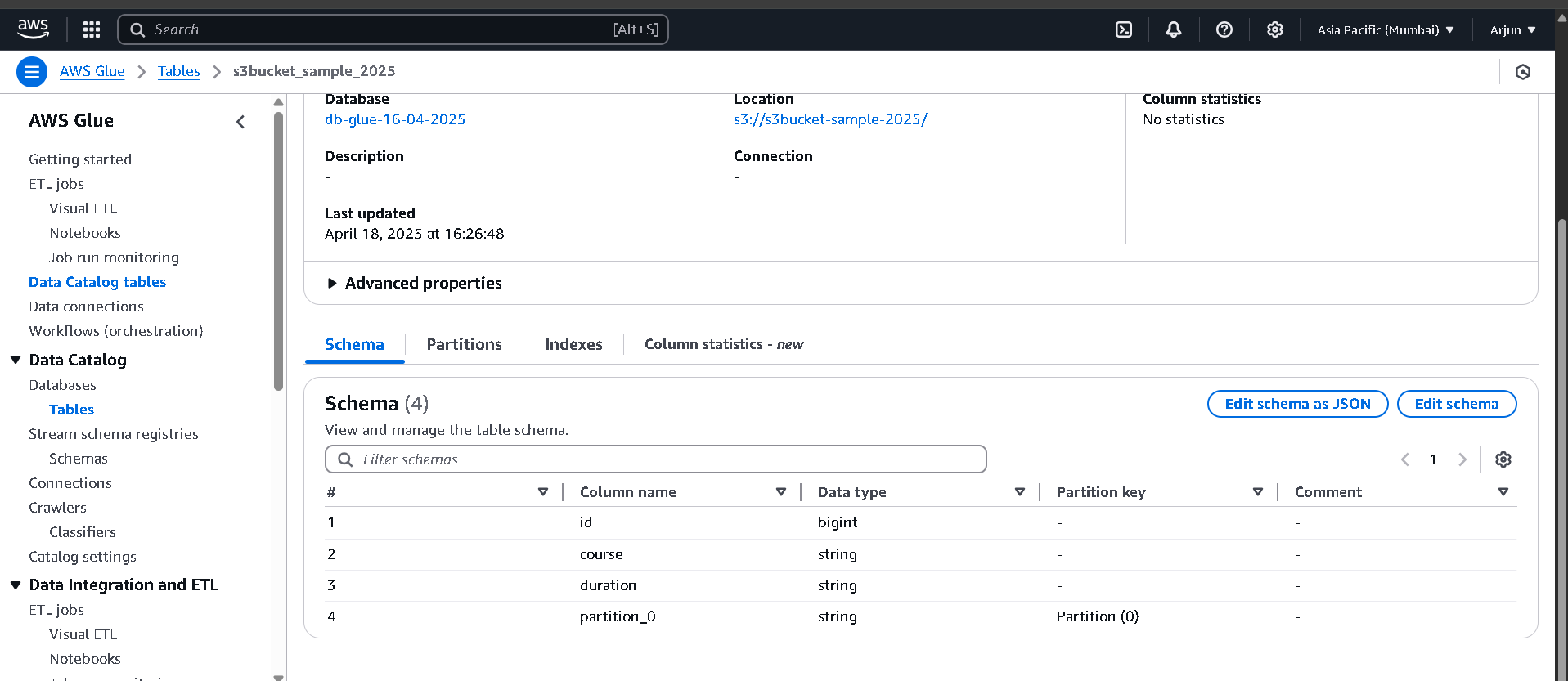
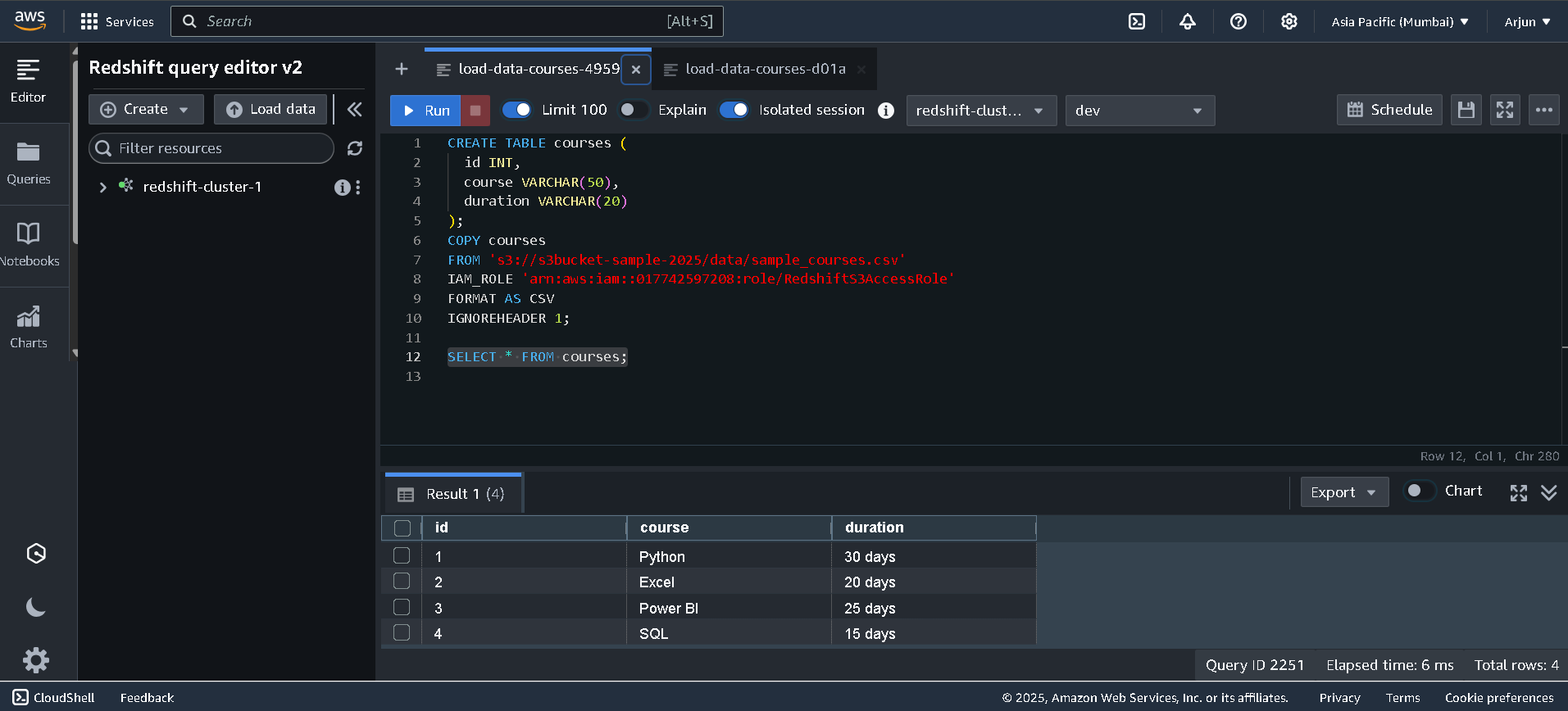
13. What is columnar storage and how does it benefit Redshift performance for analytics workloads

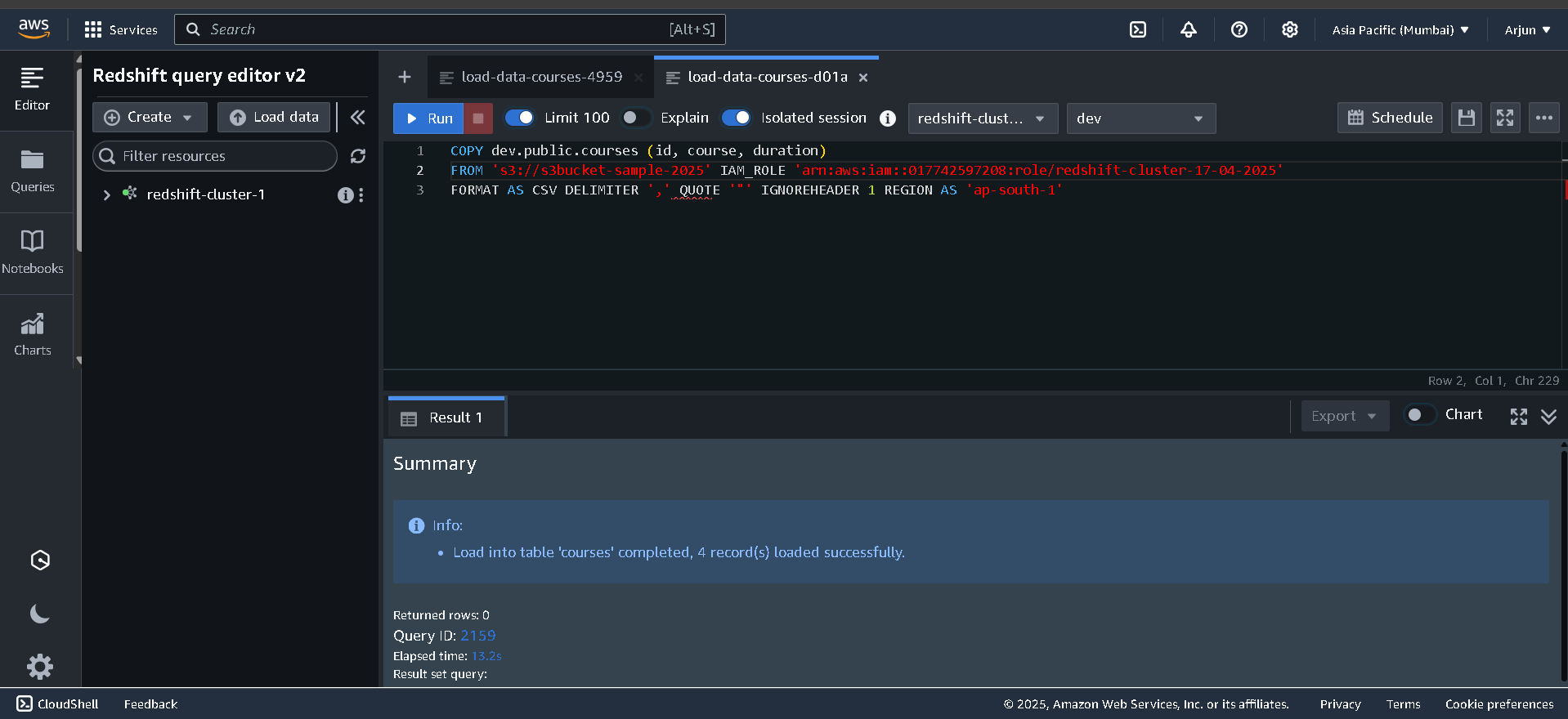
Answer:  
**Columnar Storage in Redshift**

* Stores data by column instead of row
* Faster for analytic queries that read specific columns
* Reduces I/O and improves compression

14. Load a CSV file from S3 into Redshift using the COPY command. Share table schema, command used, and sample output from a query

Answer:





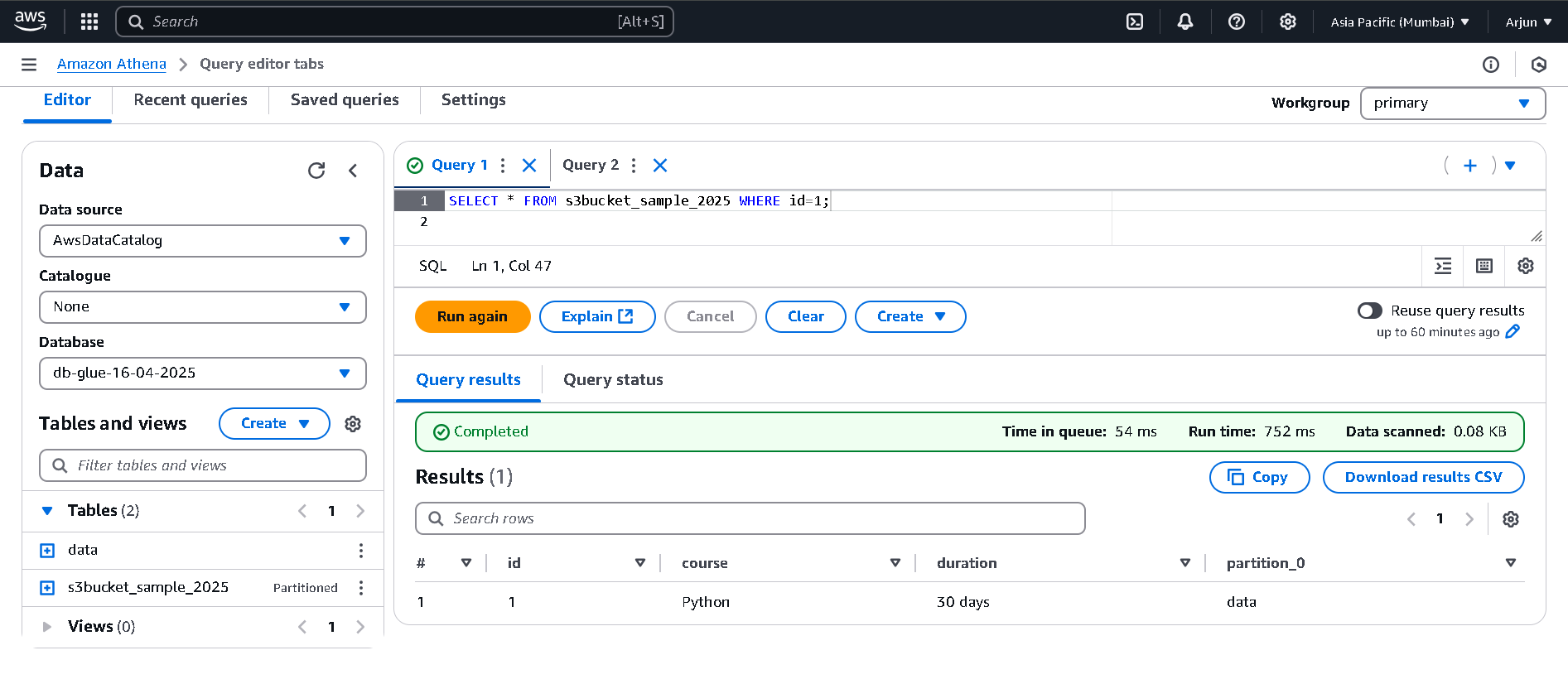
15. What is the role of the AWS Glue Data Catalog in Athena? How does schema-on-read work?  
Answer:

* **Glue Catalog** stores metadata (table names, columns, data types) for S3 data.
* Athena uses this catalog to query S3 files without loading or transforming them.
* Acts like a virtual schema layer for data in S3.

**Schema-on-Read**:

* Data is kept raw in S3.
* Schema is applied at **query time**, not during data load.
* Enables flexible and fast querying without changing original files.

16. Create an Athena table from S3 data using Glue Catalog. Run a query and share the SQL + result screenshot

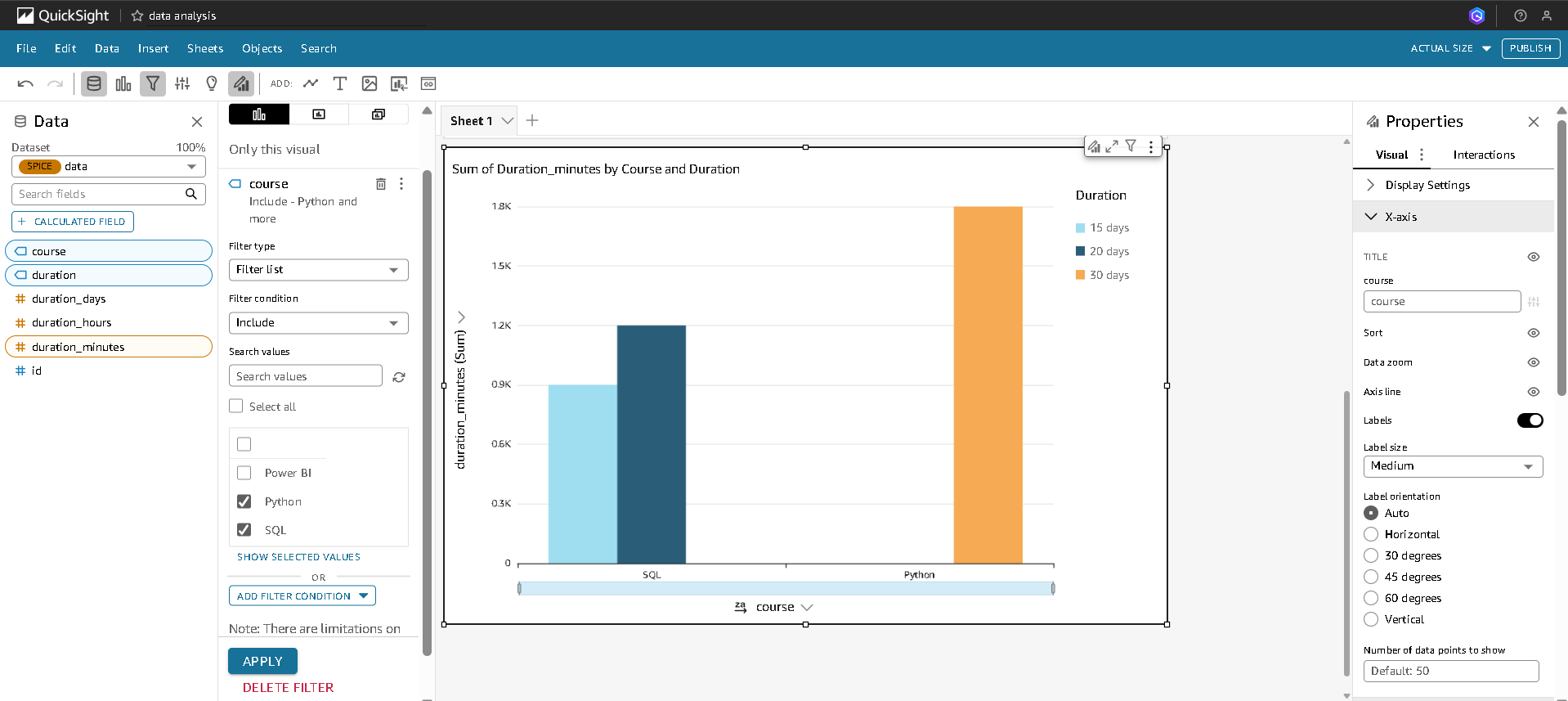
Answer:  


17. Describe how Amazon QuickSight supports business intelligence in a serverless data architecture. What are SPICE and embedded dashboards?

Answer:  
Amazon QuickSight is a **serverless BI tool** that helps visualize and analyze data without managing servers. It connects to data sources like S3, Athena, Redshift, etc., and scales automatically.

* **SPICE (Super-fast, Parallel, In-memory Calculation Engine):**  
  QuickSight's in-memory engine that speeds up queries and reduces load on data sources.
* **Embedded Dashboards:**  
  Allow integration of interactive dashboards into websites or applications for real-time analytics access by end-users.

18. Connect Quicksight to Athena or Redshift and build a dashboard with at least one calculated field and one filter. Show me a screenshot of your finished dashboard.

Answer:  


19. Explain how AWS CloudWatch and Cloud trail differ. IN a data analytics pipeline, what role does each play in monitoring, auditing, and troubleshooting?

Answer:

| **Feature** | **AWS CloudWatch** | **AWS CloudTrail** |
| --- | --- | --- |
| Purpose | Real-time monitoring & alerting | Auditing and security analysis |
| Data Captured | Logs, metrics, alarms | API call history (who did what, when, where) |
| Alert Support | Yes (with alarms and dashboards) | No (used for review, not alerts) |
| Use Case | Monitor Glue job failures, S3 usage | Track who triggered Glue jobs or accessed S3 |

**Role in Data Analytics Pipeline:**

* **AWS CloudWatch** helps monitor services like:
  + **Glue**: Tracks job success/failure, logs errors
  + **S3**: Monitors read/write activity
  + **QuickSight**: Observes dashboard performance
  + Enables alerts if something fails (e.g., Glue job fails)
* **AWS CloudTrail** tracks:
  + Who uploaded/deleted files in **S3**
  + Who started/stopped **Glue** jobs or crawlers
  + Who accessed/modified **QuickSight** dashboards

20. Describe a complete end-to-end data analytics pipeline using AWS services. Include services for data ingestion, storage, transformation, querying, and visualization. (Example: S3 → Lambda → Glue → QuickSight)

Explain why you would choose each service for the stage it’s used in

Answer:  
**Pipeline:**  
**S3 → Lambda → Glue → Athena → QuickSight**

* **S3 (Ingestion):**  
  Stores raw data from various sources. Scalable, durable, and cost-effective.
* **Lambda (Trigger):**  
  Automatically triggers on new S3 uploads. Used for light data validation or processing.
* **Glue (Transformation):**  
  Cleans and transforms data. Creates a data catalog for easy access.
* **Athena (Querying):**  
  Serverless SQL engine to query data directly from S3 using Glue Catalog.
* **QuickSight (Visualization):**  
  Creates dashboards and visual reports from Athena queries.

**Why to chose each service for the stage’s setup?**  
It’s serverless, cost-efficient, scalable, and integrates well for end-to-end analytics.

*I have completed all the answers as per the assignment. Wherever screenshots were required, I have attached them. Please check accordingly.*

|  |  |
| --- | --- |
| **Name** | **Arjun Kumar Jha** |
| **Course** | **Data Analytics (Batch nov)** |
| **Email** | [**kumararjunjha@gmail.com**](mailto:kumararjunjha@gmail.com) |
| **Ph no.** | **7762919820** |