Amazon Athena Concepts & Interview Q&A; — Illustrated Guide

A Light■Theme Illustrated PDF covering architecture, hands■on SQL examples, best practices, and 25+ interview questions for data engineers.

■ Amazon Athena Architecture & Concepts

Amazon Athena is a serverless interactive query service that allows SQL querying directly on data stored in Amazon S3. It uses schema■on■read and integrates tightly with AWS Glue Data Catalog for metadata.

• Serverless Query Engine — No clusters or infrastructure to manage.

• Schema-on-Read — Define metadata for existing S3 data without data movement.

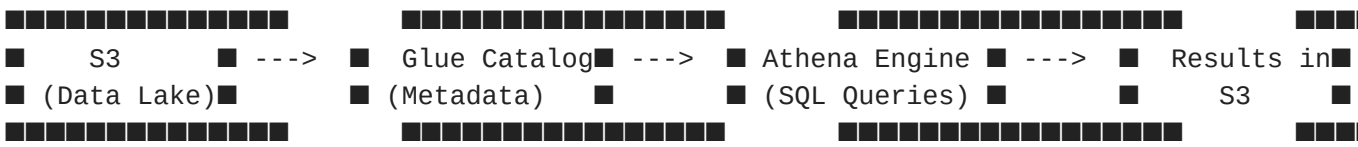
• Glue Data Catalog — Acts as the metadata store for Athena databases and tables.

• Federated Queries — Run SQL queries across S3 and external sources (RDS, DynamoDB, etc.).

• Result Storage — Each query stores results in an S3 output bucket.

• Integration — Works seamlessly with QuickSight for visualization and Redshift Spectrum for hybrid querying.

■■ Simplified Athena Query Flow



■ Hands-On SQL Examples in Athena

Define databases, tables, and queries directly on S3 data. Use Glue Catalog to store schema metadata.

```
-- Create a Database
CREATE DATABASE IF NOT EXISTS analytics_db;

-- Create a Table over S3 data
CREATE EXTERNAL TABLE IF NOT EXISTS analytics_db.sales_data (
    order_id        string,
    customer_id     string,
    region          string,
    amount_usd      double,
    order_date      date
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'
WITH SERDEPROPERTIES ('serialization.format' = ',')
LOCATION 's3://my-bucket/raw/sales/'
TBLPROPERTIES ('has_encrypted_data'='false');
```

■ Optimize performance by using columnar formats and partitions:

```
-- Partitioned Parquet Table
CREATE EXTERNAL TABLE analytics_db.sales_partitioned (
    order_id string,
    customer_id string,
    amount_usd double
)
PARTITIONED BY (region string, order_date date)
STORED AS PARQUET
LOCATION 's3://my-bucket/processed/sales/';
```

■ Query Optimization Tips

```
-- Example of optimized query using projection & partition filter
SELECT region, SUM(amount_usd) AS total_sales
FROM analytics_db.sales_partitioned
WHERE order_date BETWEEN DATE '2025-01-01' AND DATE '2025-01-31'
GROUP BY region;

-- Create a Workgroup for cost control
CREATE WORKGROUP finance_reporting
WITH configuration = ('enforce_workgroup_configuration'='true');
```

■ Amazon Athena Interview Q&A;

Q: What is Amazon Athena?

A: A serverless interactive query service that runs SQL directly on S3 data.

Q: How does Athena integrate with AWS Glue?

A: Athena uses the Glue Data Catalog as its metastore to manage database and table metadata.

Q: What data formats are supported?

A: CSV, JSON, Parquet, ORC, Avro; Parquet and ORC are preferred for performance and cost.

Q: How does Athena pricing work?

A: You pay for the amount of data scanned by each query; optimize by compression, partitioning, and projection.

Q: What are federated queries?

A: Queries that can access data stored outside S3 using connectors like RDS, DynamoDB, and CloudWatch logs.

Q: How do you improve query performance?

A: Use Parquet/ORC, partition data by key columns, and avoid SELECT *; compress data to reduce scanned volume.

Q: Explain schema-on-read.

A: Data remains in S3; schema is applied at query time, allowing flexible structure evolution.

Q: How does Athena differ from Redshift Spectrum?

A: Athena is serverless and ad-hoc; Redshift Spectrum extends Redshift SQL to S3 data with tighter integration.

Q: How do you secure Athena?

A: Use IAM policies, KMS for encryption, workgroups for query limits, and restrict S3 bucket access.

Q: What is an Athena Workgroup?

A: A logical group for query execution control, cost limits, and isolation among teams.

Q: What are query result locations used for?

A: Athena stores output results and metadata in a specified S3 bucket for later access.

Q: How does partition projection help?

A: Allows Athena to infer partitions without scanning the metastore, improving performance for large datasets.

Q: Can you write to S3 from Athena?

A: Athena is read-only; however, CTAS (CREATE TABLE AS SELECT) can materialize query results into S3.

Q: Explain CTAS in Athena.

A: CREATE TABLE AS SELECT stores the output of a query in a new table in S3.

Q: What are limitations of Athena?

A: Limited support for updates/deletes, slower for large joins, dependent on S3 file optimization.

Q: When should you use Athena?

A: For ad-hoc analytics, quick insights, or exploratory queries directly on data lakes.

Q: How to integrate Athena with QuickSight?

A: Connect QuickSight to Athena using the Data Catalog; results refresh automatically.

Q: What is Athena Federation?

A: Using data source connectors to run SQL on RDS, DynamoDB, or custom JDBC sources via Lambda.

Q: What is schema evolution?

A: Ability to add or modify fields in Glue Catalog without reloading data, supported in Athena.

Q: Difference between Athena and Presto?

A: Athena is AWS’s managed Presto service with automatic scaling and integrated security.

■ Athena vs Redshift vs Spectrum

Service	Type	Use Case	Infrastructure
Athena	Serverless SQL	Ad-hoc queries on S3	Fully managed
Redshift	Data Warehouse	Persistent analytics workloads	Managed cluster
Spectrum	Redshift Extension	Hybrid querying between Redshift & S3	Depends on Redshift

■ Athena Best Practices

• Use Glue Catalog as a centralized schema registry.

• Store data in Parquet/ORC with compression to minimize scan cost.

• Partition data on high-cardinality columns (like date, region).

• Use Workgroups for budget control and query segregation.

• Use CTAS for creating reusable aggregated tables.

• Enable encryption for both input and output data.

• Avoid SELECT * — query only required columns.