Amazon Athena

is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL.

Athena is a serverless product so with Amazon Athena AWS  manages all the infrastructure for you, and you pay only for the queries that you run.

With Amazon Athena AWS makes it extremely easy to use – simply point to your data in Amazon S3, define the schema, and start querying using standard SQL.

Amazon Athena uses Presto with full standard SQL support and works with a variety of standard data formats, including CSV, JSON, ORC, Apache Parquet and Avro.

While Amazon Athena is ideal for quick, ad-hoc querying and integrates with Amazon QuickSight for easy visualization, it can also handle complex analysis, including large joins, window functions, and arrays.

Amazon Athena uses a managed Data Catalog to store information and schemas about the databases and tables that you create for your data stored in Amazon S3.

With Amazon Athena, you don’t have to worry about managing or tuning clusters to get fast performance.

Amazon Athena is optimized for fast performance with Amazon S3.

Amazon Athena automatically executes queries in parallel, so that you get query results in seconds, even on large datasets.

Most results are delivered within seconds.

With Amazon Athena, there’s no need for complex ETL jobs to prepare data for analysis.

This makes it easy for anyone with SQL skills to quickly analyze large-scale datasets.

Amazon Athena is out-of-the-box integrated with AWS Glue Data Catalog, allowing you to create a unified metadata repository across various services, crawl data sources to discover schemas and populate your Catalog with new and modified table and partition definitions, and maintain schema versioning.

You can also use Glue’s fully managed ETL capabilities to transform data or convert it into columnar formats to optimize cost and improve performance.

**Amazon Athena Features**

You can query Amazon Athena using Standard SQL

Amazon Athena uses Presto, an open source, distributed SQL query engine optimized for low latency, interactive data analysis.

Athena supports a wide variety of data formats such as CSV, JSON, ORC, Avro, or Parquet.

With Athena’s federated data source connectors, you can query additional data stores and join the data with data stored in Amazon S3.

You can access Athena and run queries from the Athena console, API, CLI, AWS SDK,

On Amazon Athena AWS supports the integration with business intelligence and SQL development applications through Athena’s JDBC and ODBC drivers.

With Amazon Athena AWS use the highly redundant Global infrastructure to maintain high availability.

Whenever a particular facility is unavailable, Amazon Athena routes queries to another facility based on the available compute resources.

In Amazon Athena AWS uses Amazon S3 as an origin, making your data highly available and durable.

In addition to providing durable infrastructure for keeping important data, Amazon S3 is designed for 99.999999999% durability on a per object basis.

Amazon Athena integrates directly with Identity and Access Management and you can leverage the use of bucket policies within S3 also.

By granting IAM policies to IAM users, you can easily control S3 buckets for your users. By controlling access to data in S3, you can restrict users from querying it using Athena.

Additionally, Athena allows you to query encrypted data stored in Amazon S3 and write encrypted results back to your bucket.

Athena provides connectors for enterprise data sources including Amazon DynamoDB, Amazon Redshift, Amazon OpenSearch, MySQL, PostgreSQL, Redis, and other popular third-party data stores.

**Amazon Athena Limitations**

In terms of optimization capabilities, AWS Athena doesn’t offer much in the way of capabilities.

Although Amazon Athena is a highly scalable and reliable service, it is hosted in a muti-tenant environment.  
  
This multi-tenancy approach might trigger throttling from time to time, leading to potential issues.

Amazon Athena doesn’t include any kind of Data Manipulation interface for inserting, deleting, and updating data.

**I**f you intend to run your SQL queries efficiently, you might want to partition the data sets stored in Amazon S3. The number of partitions that you manage to create will substantially affect the speed and performance.

Indexing is not natively supported within Amazon Athena, which is built upon AWS.

**Use Cases**

Query services like Amazon Athena, data warehouses like Amazon Redshift, and sophisticated data processing frameworks like Amazon EMR, all address different needs and use cases.

Amazon Redshift provides the fastest query performance for enterprise reporting and business intelligence workloads, particularly those involving extremely complex SQL with multiple joins and sub-queries.

Amazon EMR makes it simple and cost effective to run highly distributed processing frameworks such as Hadoop, Spark, and Presto when compared to on-premises deployments. Amazon EMR is flexible – you can run custom applications and code, and define specific compute, memory, storage, and application parameters to optimize your analytic requirements.

Amazon Athena provides the easiest way to run ad-hoc queries for data in S3 without the need to setup or manage any servers.

With Athena, you can analyze unstructured, semistructured, and structured data stored in Amazon S3. Examples include CSV, JSON, or columnar data formats such as Apache Parquet and Apache ORC.

In Athena, ANSI SQL queries can be run ad-hoc without aggregating or loading the data.

Data visualization is made simple with Athena’s integration with Amazon QuickSight.

Using Athena, you will be able to generate reports or to explore your data with the help of business intelligence tools or SQL clients that are connected to a JDBC or ODBC driver to generate reports or to explore data.

With AWS Glue Data Catalog, AWHena integrates with your data to create a persistent metadata store in Amazon S3, which will allow you to access your data from anywhere.

With AWS Glue, you can create tables and query data in Athena using a central metadata store that is accessible throughout your Amazon Web Services account.

In order to troubleshoot a performance issue on your website, Athena is useful for running a quick query on web logs. A table for your data can be defined quickly with Athena, and you can start querying using the standard SQL language.

You should use Amazon Athena if you want to interact with your Cost and Usage Report stored in S3 to gain extremely specific information on how your AWS bill is being calculated. This can be done natively within the console.

The table below shows the primary use case and situations for using a few AWS query and analytics services:

|  |  |  |
| --- | --- | --- |
| **AWS Service** | **Primary Use Case** | **When to use** |
| Amazon Athena | Query | Run interactive queries against data directly in Amazon S3 without worrying about formatting data or managing infrastructure. Can use with other services such as Amazon RedShift |
| Amazon RedShift | Data Warehouse | Pull data from many sources, format and organize it, store it, and support complex, high speed queries that produce business reports. |
| Amazon EMR | Data Processing | Highly distributed processing frameworks such as Hadoop, Spark, and Presto. Run a wide variety of scale-out data processing tasks for applications such as machine learning, graph analytics, data transformation, streaming data. |
| AWS Glue | ETL Service | Transform and move data to various destinations. Used to prepare and load data for analytics. Data source can be S3, RedShift or another database. Glue Data Catalog can be queried by Athena, EMR and RedShift Spectrum |

**Best Practices**

[**Best practices**](https://aws.amazon.com/blogs/big-data/top-10-performance-tuning-tips-for-amazon-athena/) for performance with Athena:

* **Partition your data** – Partition the table into parts and keeps the related data together based on column values such as date, country, region, etc. Athena supports Hive partitioning.
* **Bucket your data** – Partition your data is to bucket the data within a single partition.
* **Use Compression** – AWS recommend using either Apache Parquet or Apache ORC.
* **Optimize file sizes** – Queries run more efficiently when reading data can be parallelized and when blocks of data can be read sequentially.
* **Optimize columnar data store generation** – Apache Parquet and Apache ORC are popular columnar data stores.
* **Optimize ORDER BY** – The ORDER BY clause returns the results of a query in sort order.
* **Optimize GROUP BY** – The GROUP BY operator distributes rows based on the GROUP BY columns to worker nodes, which hold the GROUP BY values in memory.
* **Use approximate functions** – For exploring large datasets, a common use case is to find the count of distinct values for a certain column using COUNT(DISTINCT column).
* **Only include the columns that you need** – When running your queries, limit the final SELECT statement to only the columns that you need instead of selecting all columns.

**Pricing**

With Amazon Athena, you pay only for the queries that you run.

You are charged based on the amount of data scanned by each query.

You can get significant cost savings and performance gains by compressing, partitioning, or converting your data to a columnar format, because each of those operations reduces the amount of data that Athena needs to scan to execute a query.

**Amazon Athena Frequently Asked Questions (FAQs)**

Q: What can I do with Amazon Athena?

With Amazon Athena, you can analyze data stored in Amazon S3. Athena supports interactive analytics using ANSI SQL without aggregation or loading data. Data sets can be unstructured, semistructured, or structured using Amazon Athena. You can use Amazon Athena with Amazon QuickSight to visualize CSV, JSON, Avro or columnar data formats like Apache Parquet and Apache ORC.

Q: How do I get started with Amazon Athena?

Using Amazon Athena is as simple as logging into the AWS Management Console for Athena and writing your DDL statements or using a create table wizard. A built-in query editor lets you query data.

Q: What is the difference between Amazon Athena, Amazon EMR, and Amazon Redshift?

The Amazon Redshift platform delivers the fastest query performance for enterprise reporting and business intelligence workloads, especially for SQL that has multiple joins.

Comparatively to on-premises deployments, Amazon EMR makes highly distributed processing frameworks like Hadoop, Spark, and Presto simple and cost-effective.

Amazon EMR allows you to run custom applications and code, and define specific compute, memory, storage, and application parameters.

Without setting up or managing any servers, Amazon Athena offers interactive queries for S3 data.

Q: When should you use a full featured enterprise data warehouse, like Amazon Redshift vs. a query service like Amazon Athena?

‘Q: How do I create tables and schemas for my data on Amazon S3?

<https://digitalcloud.training/amazon-ec2-auto-scaling/>

<https://digitalcloud.training/amazon-rds/>

<https://digitalcloud.training/aws-certificate-manager/>