mission cloud

Data And Analytics

Published July 05, 2022

Amazon EMR Serverless vs. AWS Glue

AWS provides several tools for big data extract, transform, and load (ETL) processes and analytics, including Amazon EMR Serverless and AWS Glue. While these two tools are similar, there are some key differences between them. Amazon designed Amazon EMR Serverless for analytics and AWS Glue in tandem with other tools for ETL processes.

Amazon Elastic MapReduce (EMR) Serverless is a new deployment option, still in preview in Amazon EMR. It simplifies the operation and cost optimization of running open-source frameworks like Apache Spark, Hive, and Presto-built applications at a petabyte scale.

to choose one over the other, and their cost benefits.

Amazon EMR Serverless

Before Amazon EMR Serverless, users primarily leveraged Amazon EMR on Elastic Compute Cloud (EC2), Elastic Kubernetes Service (EKS), and Outpost. These options necessitated planning, configuration, management, and scaling of clusters.

Aside from errors that could arise with such deployments or mandates to conform to business corporate security policies, there were often lopsided finance and efficiency costs for over or under-provisioning when data sizes changed. These inefficiencies meant that an operator must be highly skilled to process jobs. Because clusters were complex to set up, data engineers ran them longer than necessary, which led to higher spending.

In response, Amazon released EMR Serverless in November 2021. EMR Serverless offers a serverless runtime environment that precludes direct intervention with cluster configuration, management, and scaling. There's no need to configure, manage, or scale clusters because Amazon handles them. Additionally, you don't need to manage virtual machines (VMs) or install and maintain runtime software. You can start, stop, and delete apps on-demand, easing operations and reduces labor and financial costs.

EMR Serverless provides petabyte analytics processing using popular big data open-source software like Apache Spark, Apache Hive, and Presto. Especially for Spark and Hive, you can enjoy processing speeds at Amazon-optimized runtimes up to twice as fast as the open-source edition.

EMR Serverless also provides plenty of flexibility for executing jobs. At your discretion, you may pre-initialize a driver and set of executors. You can specify the maximum number of workers to which your application can scale. You can choose from one, two, or four virtual central processing units (vCPU) for each worker and from 2 to 30 GB per worker

You can expect staggered updates to EMR Serverless. Updates include shared access to Jupyter Notebooks using EMR Studio, which makes it easier to query and analyze your data interactively using Spark UI, Tez UI, MXNet, and TensorFlow after the ETL process.

If you are interested in data pipelines, you can run any pipelines created by AWS Step Functions, AWS Managed Workflows for Apache Airflow (Amazon MWAA), and SageMaker (for machine learning). There are alsoEMR notebooks, serverless notebooks that you can use to run queries and code. Unlike traditional notebooks, the contents of an EMR notebook are equations, queries, models, code, and narrative text run on a client. You execute commands using a kernel on the EMR cluster.

EMR Serverless Cost Benefits

EMR Serverless offers further cost savings because there are no upfront costs and it supports sharing. Supported sharing allows multi-tenants with different identities and access management (IAM) roles to use the same application. Additionally, you only pay for the aggregate virtual CPU (vCPU), memory, and storage computing resources that you use.

Ultimately, EMR Serverless lets you use the big data processing and analysis tooling you're already familiar with in a fully-managed environment.

Amazon Glue

Big data in its raw form usually requires ETL in a data warehouse for processing, analytics, and machine learning. AWS Glue is an easy-to-use serverless ETL tool with well-working individual parts.

For more complex ETL scenarios that can benefit from automation, AWS Glue includes reusable blueprints that accept parameters that can create workloads on-demand or bya schedule.

Perhaps, AWS Glue's niftiest feature is its ETL engine can generate Python or Scala code. Not all open-source tooling can do this. Sometimes have to write code just to connect parts of your data pipeline and ensure seamless operation.

Under the AWS Glue umbrella is AWS Glue DataBrew, which can be used for cleaning and normalizing data with a no-code visual interface. The FindMatches feature uses machine learning to find duplicates or imperfect matches of records and AWS Glue Elastic Views combines and replicates data in multiple data stores. Together, these tools help automate the ETL process so you can spend more time analyzing your data.

AWS Glue Cost Benefits

As you can expect, each tool you use incurs charges. AWS Glue charges an hourly rate. You pay separately for crawlers and ETL jobs by the second. For the AWS Glue Data Catalog, you pay a monthly fee for storing and accessing the metadata. However, the first million objects and accesses are free.

Any provisioned development endpoint for the interactive development of your ETL code incurs an hourly rate billed by the second. It can be expensive for long-running sessions. As for Glue DataBrew, AWS bills separately for sessions and jobs by the minute. So, while AWS Glue may be inexpensive, using another suite of tools for extended periods can raise costs. Also, different groups in your company can use AWS Glue to collaborate on data integration tasks, including extraction, cleaning, normalization, combining, loading, and running scalable ETL workflows. The capacity for collaboration reduces the time it takes to analyze your data and take it to market.

and processing tasks just like an EC2 and a relational database service (RDS) instance can run databases. The key difference is Amazon's recommended use for each — AWS Glue for ETL and AWS EMR Serverless for data processing and analytics.

When to Use Which

If you want to use well-known data processing and analysis tools that aren't necessarily AWS-specific, Amazon EMR is a fast, cost-effective solution. The Apache Spark framework, for example, is most in-demand in all major big data processing and analytics industries, and its codebase is actively contributed to and maintained. With Amazon EMR, you can run mainstream open-source big data processing software such as Apache Spark and Hive at AWS-optimized runtime speeds and scaled costs.

However, consider AWS Glue if you prefer a simpler and more cohesive data pipeline. AWS Glue connects to tools for data ingestion, discovery, query, visualization, and loading tasks.

Many companies that ingest streaming data use AWS Glue. One success story is the BMW Group. The BMW Group saves on utility and costs because AWS Glue supports sources like Amazon Kinesis and Apache Kafka, facilitates the cleaning and transformation of data streams in-flight, and continuously loads the results into AWS S3, Data Lake, and other data stores

If your data is raw, requires ETL, and relies on a bevy of tools for extended periods, AWS Glue is your best bet. It's widely used and has ample documentation. Keep in mind that AWS Glue is more expensive than EMR Serverless for similar compute resources.

EMR Serverless is a good choice if your data is already in the preferred format, as petabyte-scale processing will be batched and super-fast. Opting for this tooling provides simple, cost-efficient, multi-tenanted operations with

Conclusion

Amazon EMR and AWS Glue are great tools with overlapping capabilities. Both can do a great job with ETL and processing. However, the best one for you depends on your specific circumstances.

Not sure how to start? Get in touch with Mission for a free consultation.

FAQ

How do the costs of using Amazon EMR Serverless compare to AWS Glue for typical data processing workloads?

When evaluating the cost implications of utilizing Amazon EMR Serverless versus AWS Glue, it's essential to consider the unique pricing models of each service. Amazon EMR Serverless charges based on the compute and memory resources consumed by the data processing tasks, offering a pay-as-you-go model that aligns with the serverless paradigm. This means costs are directly tied to the scale and complexity of the jobs being run. On the other hand, AWS Glue's pricing is determined by the runtime of the data processing jobs and the number of Data Processing Units (DPUs) used, which measure the computational power allocated to each job. Given these differing pricing structures, the most cost-effective solution will vary depending on the specific nature of the workload, such as the volume of data, the complexity of the processing tasks, and the frequency at which jobs are run.

Can Amazon EMR Serverless and AWS Glue be integrated to leverage the strengths of both services in a single data processing pipeline?

The integration of Amazon EMR Serverless and AWS Glue within a unified data processing pipeline can offer a comprehensive solution leveraging both services' strengths. While AWS Glue excels in data integration and ETL processes, Amazon EMR Serverless provides a robust environment for big data analytics and processing using efficient processing.

What are the specific security features and compliance certifications available for Amazon EMR Serverless and AWS Glue?

Security and compliance are paramount considerations in the cloud, and both Amazon EMR Serverless and AWS Glue offer robust security features designed to protect sensitive data and meet regulatory requirements. These include encryption at rest and in transit, fine-grained access controls, and integration with AWS Identity and Access Management (IAM) for secure authentication and authorization. Additionally, both services adhere to AWS's compliance programs, covering standards such as HIPAA for healthcare data, PCI DSS for payment information, and GDPR for data protection in the European Union. However, the specific compliance certifications and security mechanisms can vary slightly between services, so businesses should review the detailed security documentation and best practices for each service to ensure they meet their particular security and compliance needs.

Go Back   Share this post   in f 𝕏

Author Spotlight: