Overview of Data Analytics in AWS — Glue, Athena and DataLake

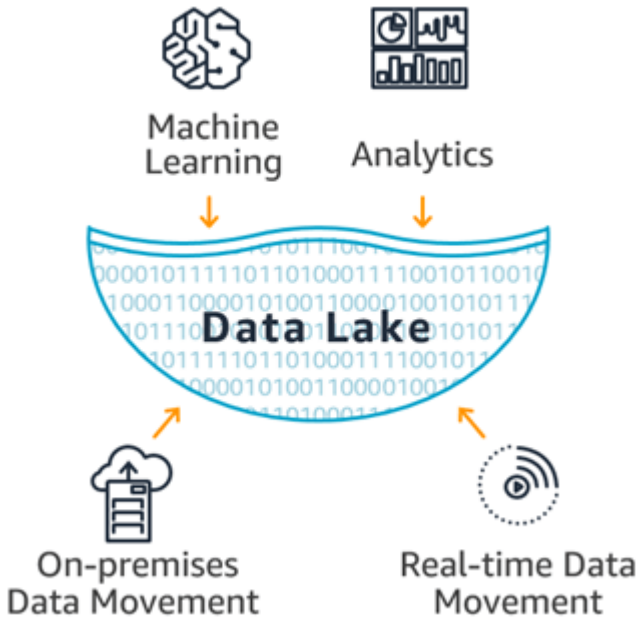
Craig Godden-Payne
Feb 7 · 6 min read ★

I hope with this post to discuss the current state of analytics in AWS Cloud.

What is a data lake?

A data lake could be considered as a centralised repository of information, except data can be structured or unstructured. The data is usually stored in S3, and the data is used for data exploration, reporting, analytics, machine learning and artifical intelligence. A datalake makes the data available to more users across more lines of business, to be analysed for insights.

Since the data is heterogeneous, (data with high variability of data types and formats) it needs to be transformed before it can be analysed



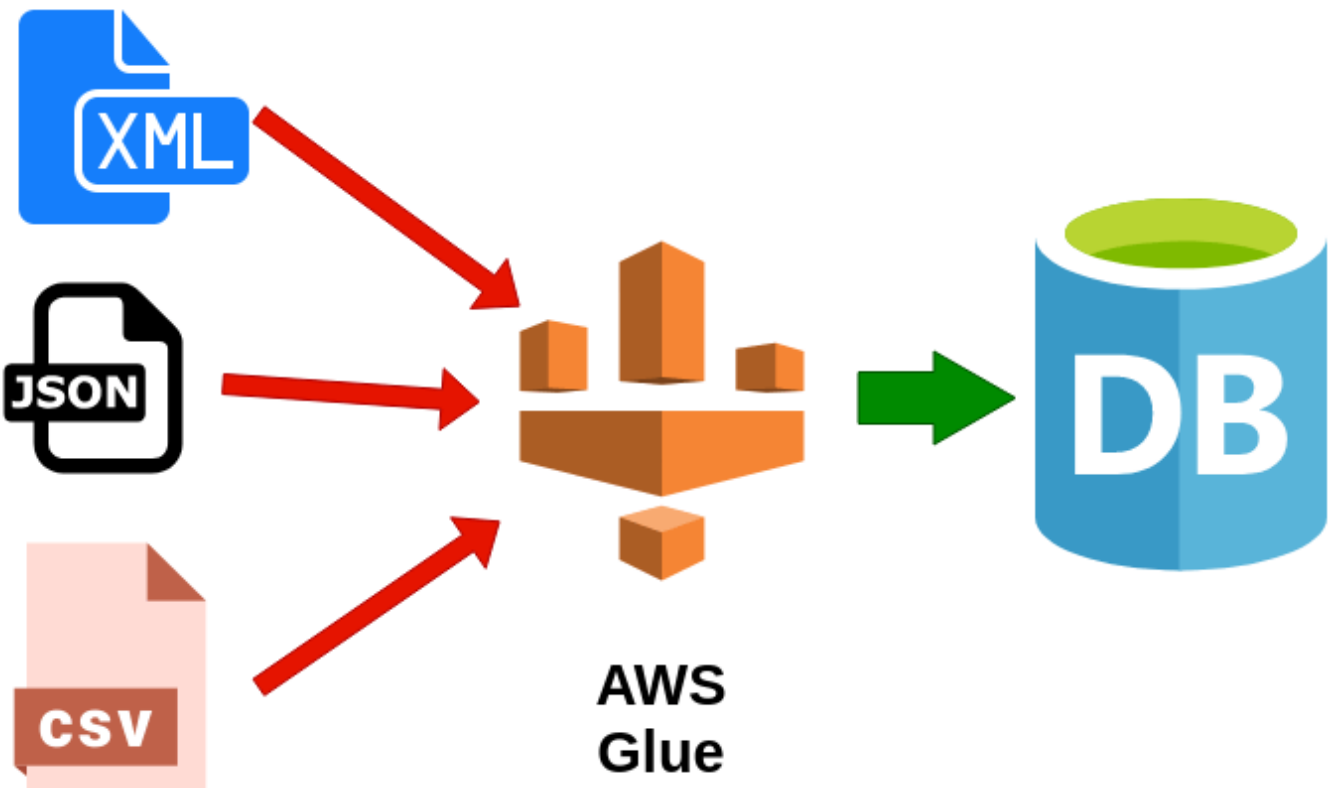
Why is S3 the primary storage?

- Scalable (almost unlimited scalability)
- Performance
- Availability
- Durability
- Cost effective (especially when moving to cheaper tiers automatically)

- Easy to transfer
- Data insights
- Security is very good
- Compliance (has been audited by many authorities and proved compliant)
- Auditing
- Pluggable into event notifications, such as SQS, SNS or Lambda
- Versioning (if needed)
- Cross region replication
- Flexible access control mechanisms (bucket policies, temporary links etc.)
- Access logs

Other benefits directly related to this post is that S3 plugs into to numerous analytics and machine learning services provided by AWS.

What is AWS Glue and how does it work?



ETL is defined as a process that extracts the data from different source systems, transforms the data (like applying calculations, concatenations, etc.) and then loads the data into the Data Warehouse system. ETL full-form is Extract, Transform and Load.

The way AWS Glue works, is that you:

- Point a Glue crawler at a data source.
- Crawler will create data catalogue with enough information to recreate the dataset.
- Glue jobs are then used to perform the ETL (jobs can be run on demand or using triggers).

Once AWS Glue has catalogued the data, it is ready to be used for analytics. You can use tools like AWS Athena to analyse and process data, or you can view visualise analytical results within quicksight.

AWS Glue Data Catalogue

AWS Glue solves the problem of analysing heterogeneous data types, it provides one central location for all your company data, including data from on premises, which solves the problem of data silos in different locations. It uses Glue crawlers to gather schemas and statistics about the data, and populates the Glue Data Catalogue with the mgathered metadata.

Glue automatically generates the ETL code to:

- Extract the data from the source
- Transform the data to match the target schema
- Load the data source into the target

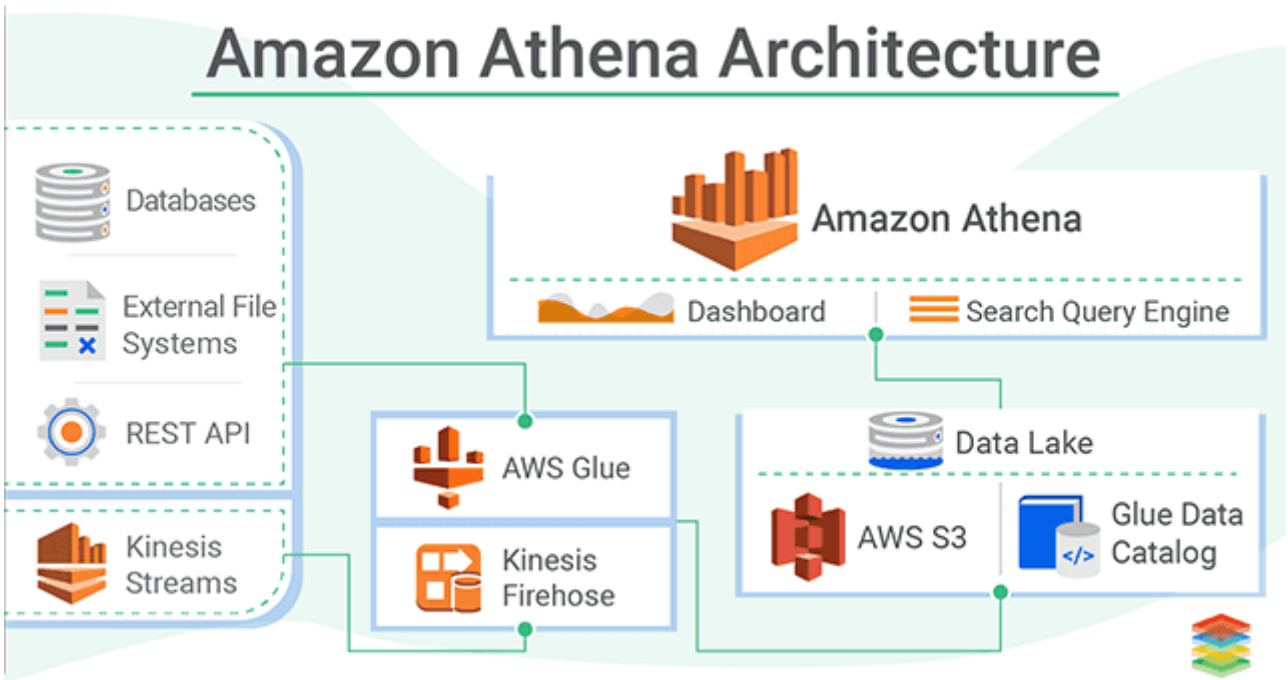
There are other types of transformations you can apply during this stage. The Glue jobs can be run on demand, on a schedule or based on events

What is AWS Athena?

Athena can be described as a serverless interactive query service, for allowing instantaneous querying of datasets within S3. It can query unstructured, semi structured and structured datasets, and will scale automatically depending on the size of the operation.

The query language used is Presto, which is an open source, distributed sql query engine. It can query data in any format, such as CSV, JSON or columular data such as Parquet. It is extremely fast, and executes queries in parrallel, and is optimised for fast performance with Amazon S3.

Athena integrates with AWS Glue.



How the integration works

- AWS Glue crawler will crawl S3 bucket (raw dataset)
- AWS Glue crawler writes metadata into Data Catalogue
- When you run a query in Athena over the raw datasets in S3
- Schema definitions are used from Data Catalogue to optimise the queries

What is AWS's current Portfolio for Data products

Data movement

- Data migration services — AWS Database Migration Service helps you migrate your databases to AWS with virtually no downtime.
- Snowball — Snowball is a petabyte-scale data transport solution that uses devices designed to be secure to transfer large amounts of data into and out of the AWS Cloud.

- Snowmobile — AWS Snowmobile is an Exabyte-scale data transfer service used to move extremely large amounts of data to AWS.
- Kinesis Firehose — Prepare and load real-time data streams into data stores and analytics tools.
- Kinesis Stream — Amazon Kinesis Data Streams (KDS) is a massively scalable and durable real-time data streaming service.
- Data Pipeline — AWS Data Pipeline is a web service that helps you reliably process and move data between different AWS compute and storage services, as well as on-premises data sources, at specified intervals.
- Direct Connect — AWS Direct Connect is a cloud service solution that makes it easy to establish a dedicated network connection from your premises to AWS.

Databases

- QLDB — Quantum Ledger Database, which provides a transparent, immutable, and cryptographically verifiable transaction log owned by a central trusted authority.
- Elasticache — Redis compatible in-memory data store built for the cloud. Power real-time applications with sub-millisecond latency.
- Aurora — MySQL and PostgreSQL-compatible relational database built for the cloud. Performance and availability of commercial-grade databases at 1/10th the cost.
- RDS — Managed relational databases in the cloud, to reduce the amount of management involved in running your own cluster.
- Neptune — Neptune is a fast, reliable, fully managed graph database service that makes it easy to build and run applications that work with highly connected datasets.
- DynamoDB — DynamoDB is a key-value and document database that delivers single-digit millisecond performance at any scale.
- Timestream — Amazon Timestream is a fast, scalable, fully managed time series database service for IoT and operational applications that makes it easy to store and analyze trillions of events per day at 1/10th the cost of relational databases.

Analytics

- Redshift — Redshift powers mission critical analytical workloads for Fortune 500 companies, startups, and everything in between.
- EMR — Easily Run and Scale Apache Spark, Hadoop, HBase, Presto, Hive, and other Big Data Frameworks.
- Elasticsearch — mazon Elasticsearch Service is a fully managed service that makes it easy for you to deploy, secure, and operate Elasticsearch at scale with zero down time.
- Athena — Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL.
- Kinesis — Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information.

BI and Machine Learning

- Quicksight — Amazon QuickSight is a fast, cloud-powered business intelligence service that makes it easy to deliver insights to everyone in your organization.
- Sage Maker — Amazon SageMaker provides every developer and data scientist with the ability to build, train, and deploy machine learning models quickly.
- Comprehend — Amazon Comprehend is a machine learning powered service that makes it easy to find insights and relationships in text.
- Reckognition — Amazon Rekognition makes it easy to add image and video analysis to your applications. You just provide an image or video to the Rekognition API, and the service can identify the objects, people, text, scenes, and activities, as well as detect any inappropriate content.
- Lex — Amazon Lex is a service for building conversational interfaces into any application using voice and text.
- Transcribe — Amazon Transcribe makes it easy for developers to add speech-to-text capability to their applications. Audio data is virtually impossible for computers to search and analyze.
- DeepLens — AWS DeepLens helps put machine learning in the hands of developers, literally, with a fully programmable video camera, tutorials, code, and pre-trained models designed to expand deep learning skills.

It is also worth mentioning about AWS Macie — AWS security service, which uses machine learning to automatically discover, classify and protected sensitive data stored in S3. It recognises sensitive data, such as personally identifiable or intellectual property

Written on November 21, 2019.

Originally published on: https://craig.goddenpayne.co.uk/building-a-modern-analytics-platform/

Data Analytics AWS Glue Athena Datalake