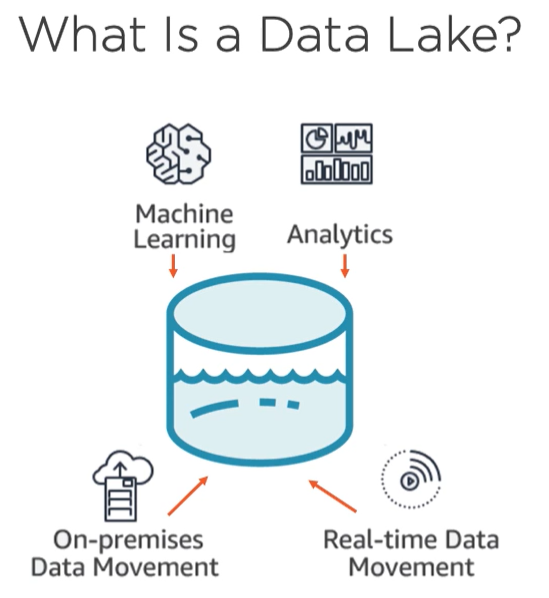
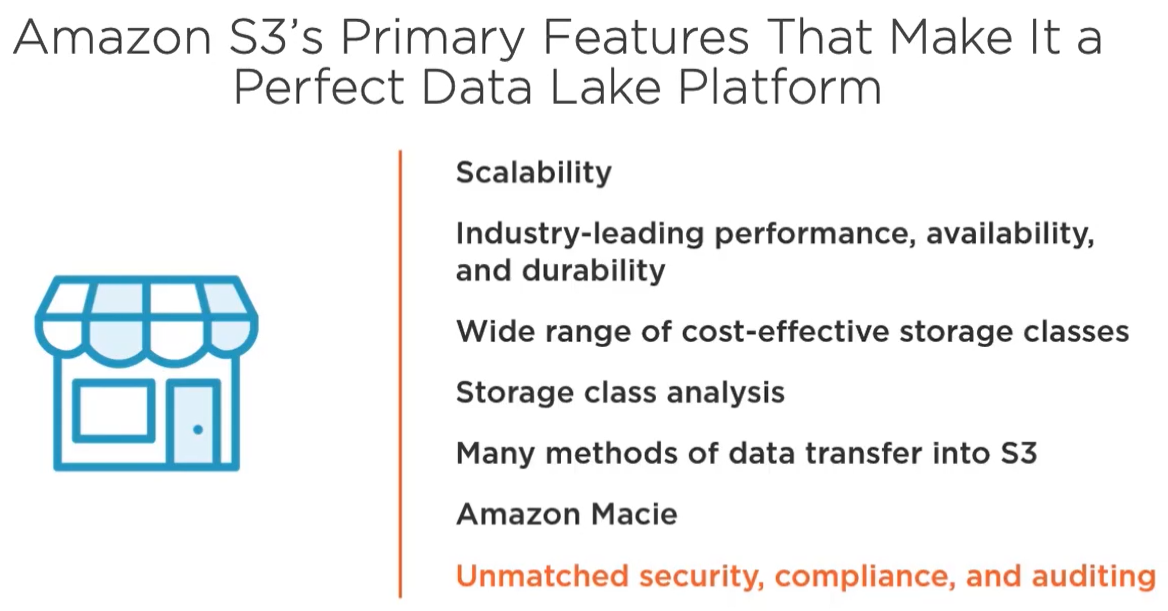
**Data Lake**



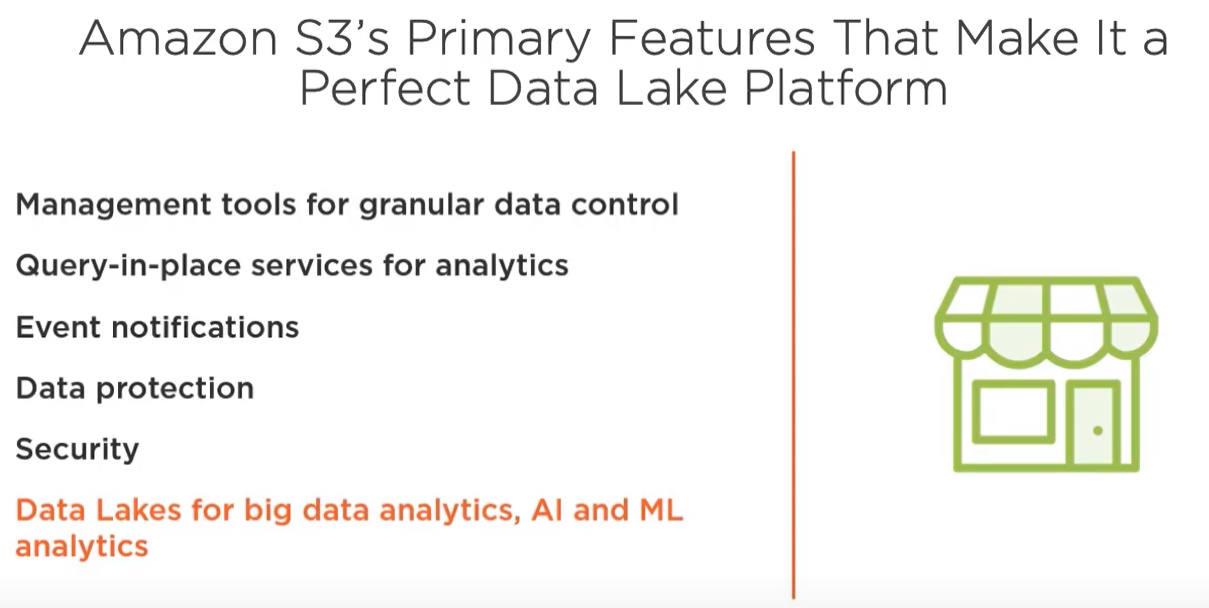
* A Data Lake is a central repository that allows you to store all your structured and unstructured data any scale, you can store your data as is without having to first structure the data.

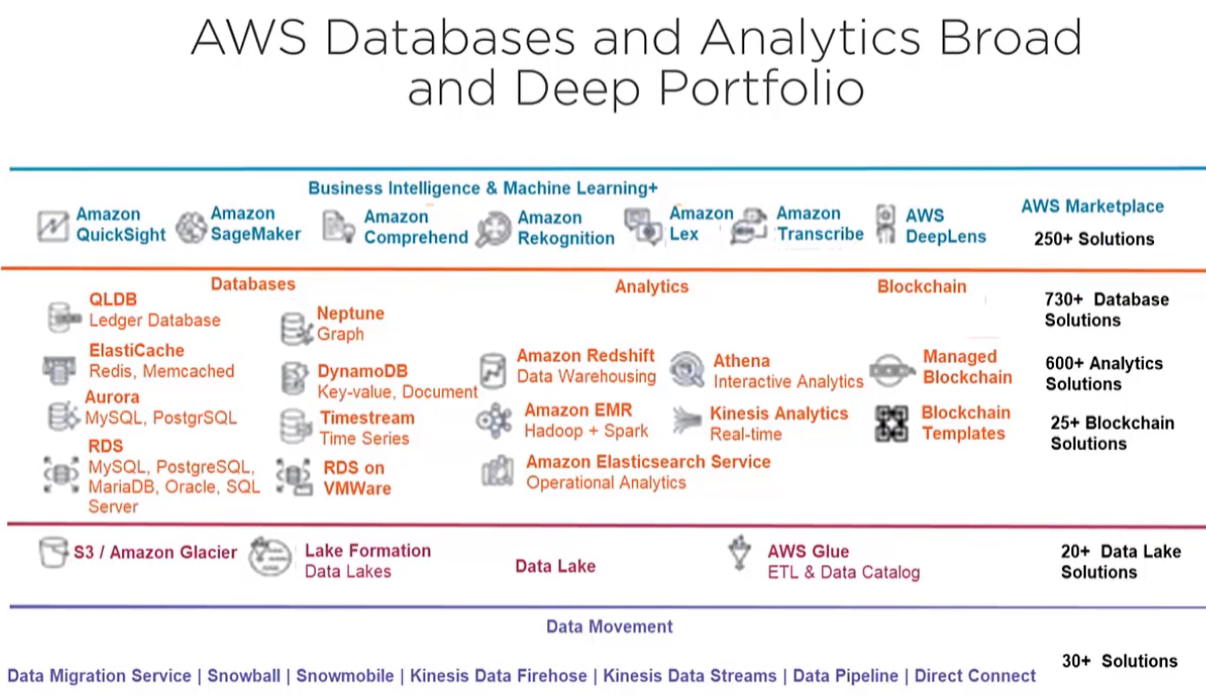
**S3 as Data Lake**

* When you use Amazon S3 as the data lake storage platform, you have a complete portfolio of data exploration, reporting, analytics, machine learning, artificial intelligence, and visualization tools to use on all your data.
* Amazon S3 is used as the primary storage platform for a data lake and it is perfect data lake storage platform.
* S3 is an optimal foundation for a data lake because of its virtually-unlimited scalability.
* S3 is designed for the 11-nines of data durability because it automatically creates and stores copies of all S3 objects across multiple systems.
* This means your data is available when needed and protected against failures, errors, and threats.
* You can take advantage of the range of S3 storage classes, moving to a less-expensive S3 class when certain criteria is met. S3 lifecycle policies can be enabled so your data will automatically transfer to a different storage class without any changes to your application. You can use Amazon S3 storage class analysis to discover data that should move to a lower-cost storage class based on access patterns. These last two benefits contribute to substantial cost savings. There are many ways to transfer data into S3.



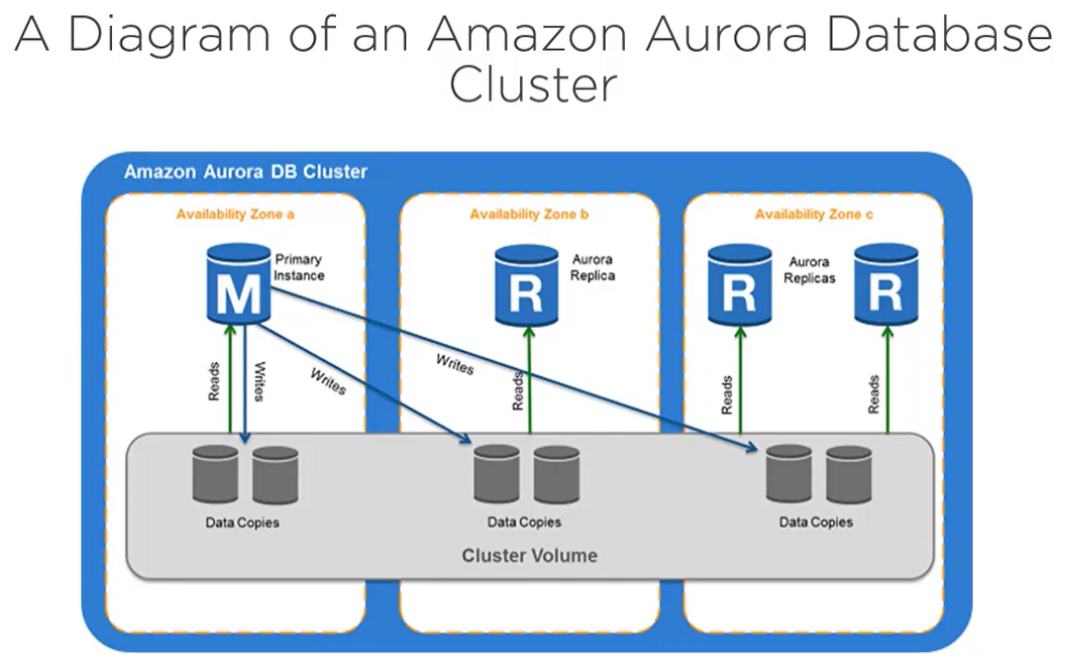
* **Amazon Macie** is a security service that uses machine learning to automatically discover, classify, and protect sensitive data stored in S3. It recognizes sensitive data like **personally-identifiable information, or PII, and intellectual property**. It's a fully-managed service that continually monitors data access activity for anomalies and generates detailed alerts when it detects risk of unauthorized access or inadvertent data leaks.

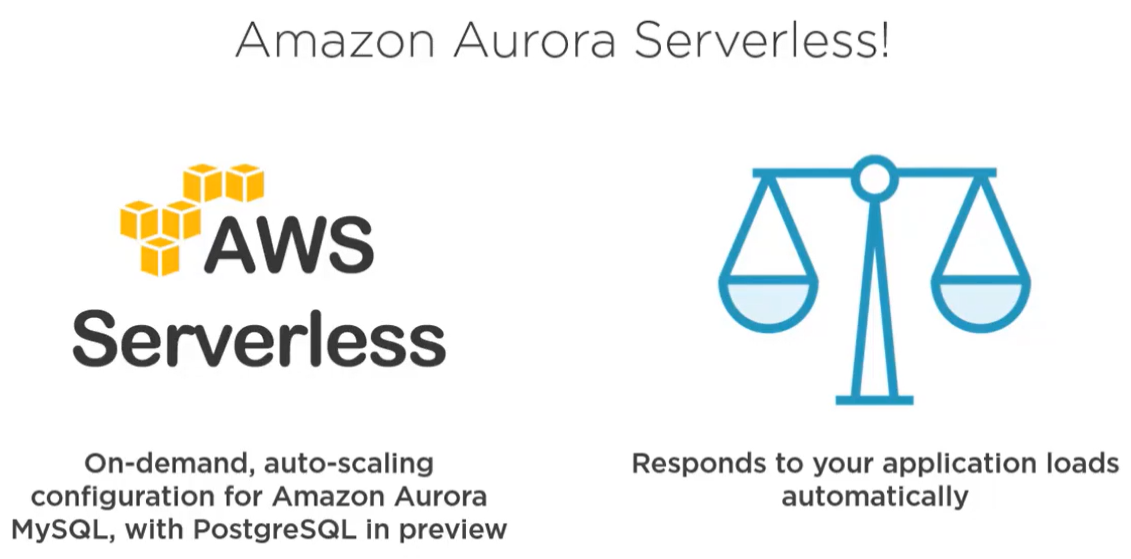




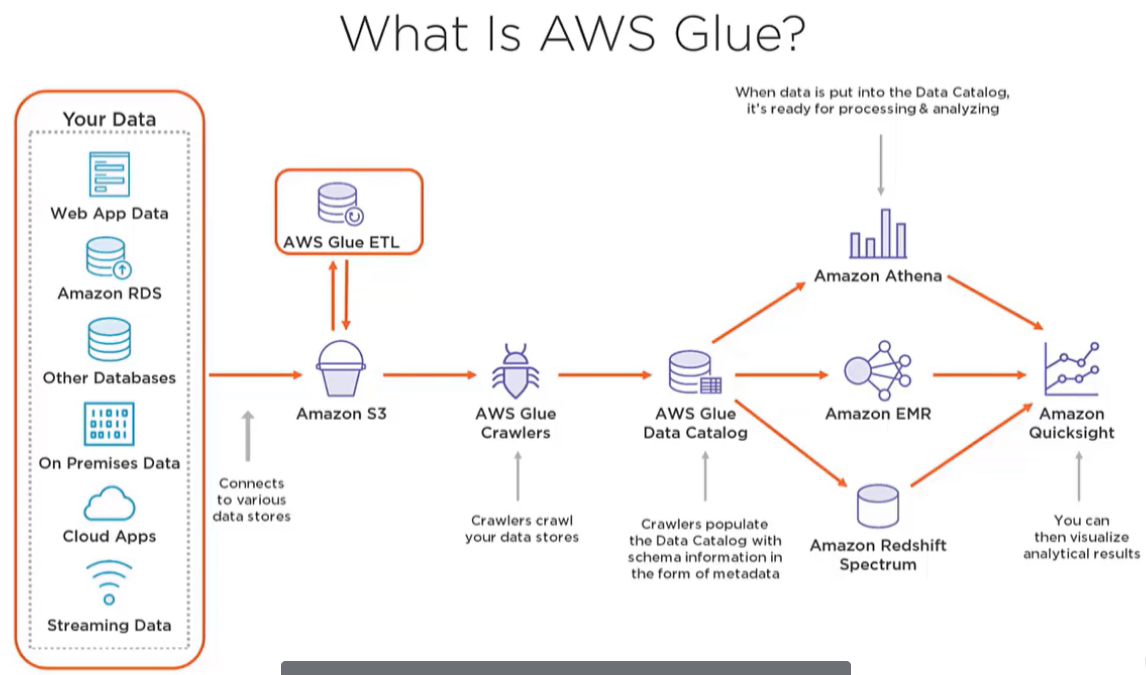
**Amazon Aurora: The Relational Database of Choice for Analytics**

* Amazon Aurora is a MySQL or PostgreSQL compatible relational database engine.

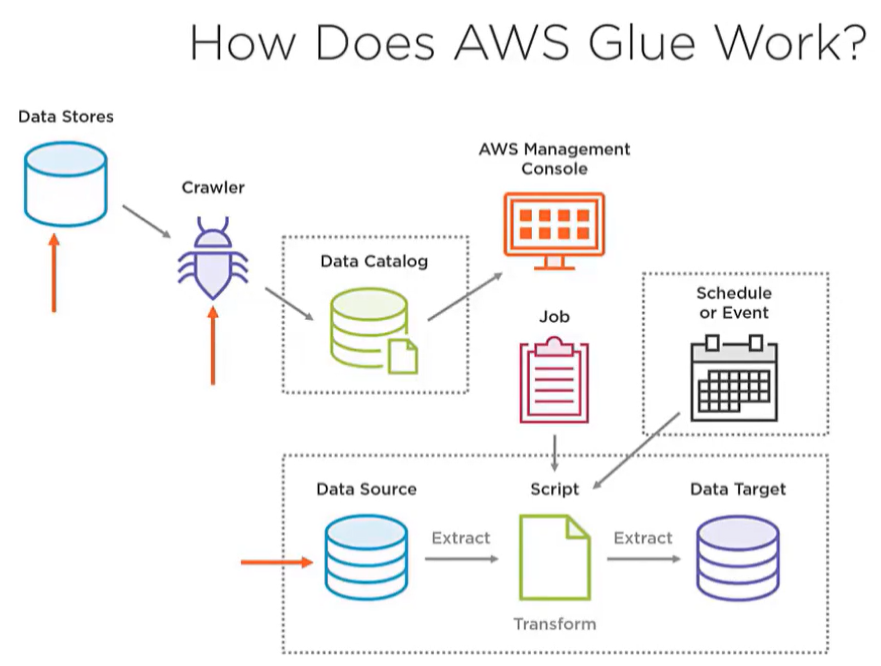




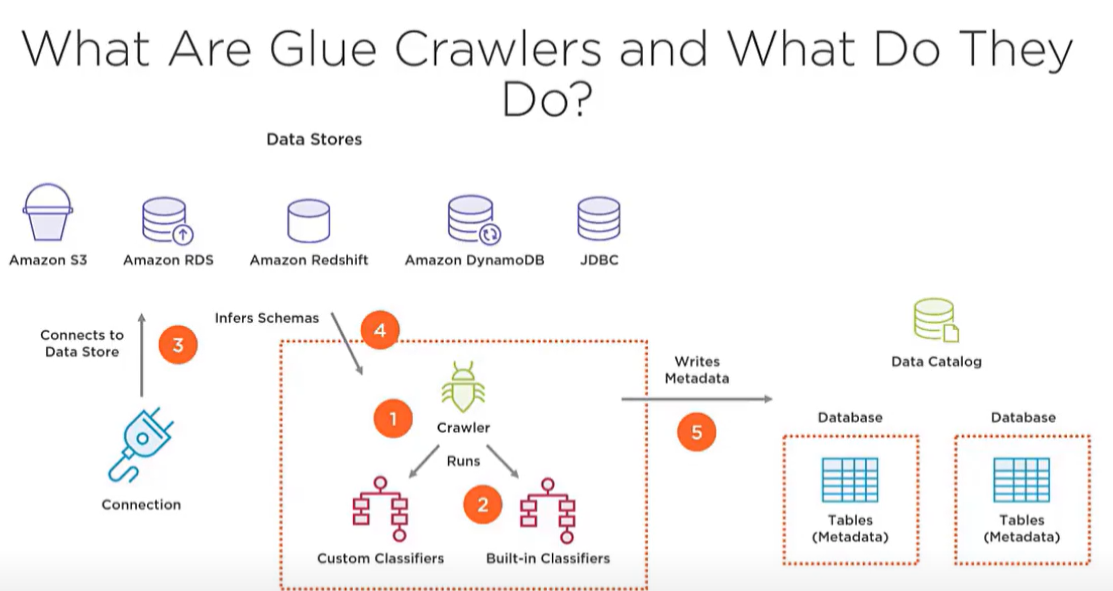
**AWS Glue**



* The emergence of unstructured and semi-structured data produced new problems of performing ETL on non-relational data that's necessary to do prior to analysis. **Glue is a fully-managed ETL service that solves that business problem.**
* You can run Glue ETL at any time in this workflow, which in this diagram is after data arrives in S3, but it can be performed at any point in this workflow. Glue crawlers crawl all your data stores, determines the schema, then populates the Glue Data Catalog with that schema and other metadata necessary to reconstruct the database when it's ready to be used for analytics.
* Most often, after you have the results of your data analysis, it's common to visualize the information. In this diagram, the visualization is done with Amazon Quicksight, but you can use any visualization tool you like.

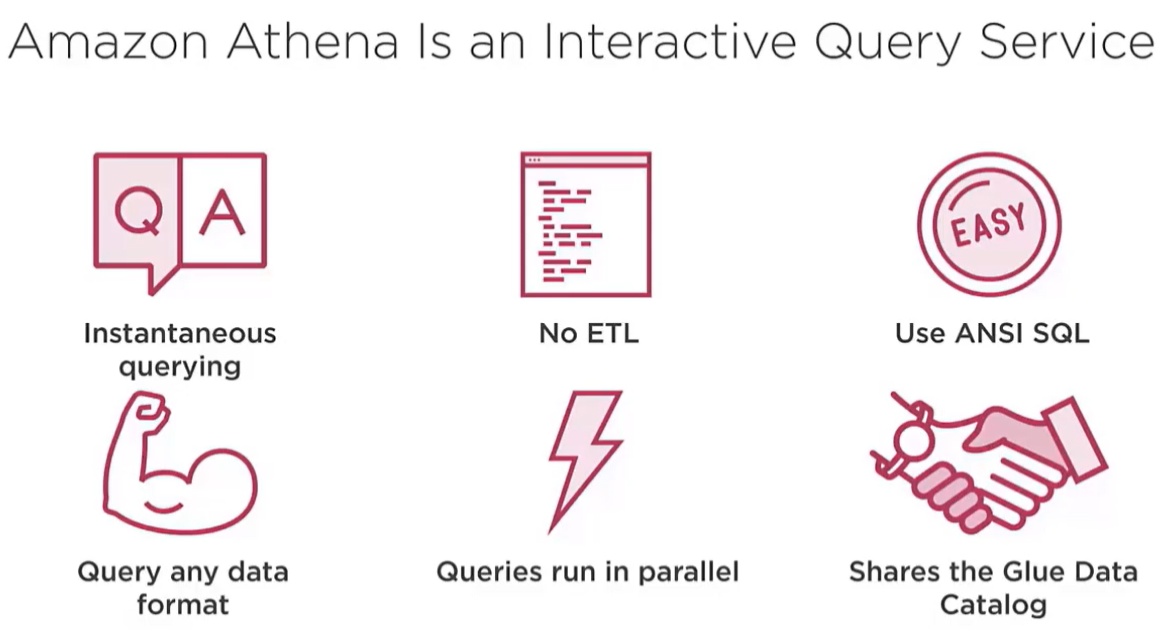
****

* **Let's now look at how Glue works.** To crawl the data source, you simply point a Glue crawler at the source data, and the crawlers populate your Glue Data Catalog with enough metadata and statistics to reconstruct the source data when you request that data to be used in analytics.
* Glue jobs perform the ETL on the raw data in an Amazon S3 data lake. During the ETL process, many different types of transformations can be done. Once the ETL is complete, the Glue job puts the transform data into wherever you specify, which is called a target data source.
* You can run jobs on demand, or you can set up triggers to automatically start a job whenever certain criteria are met. Triggers can be time based or event based. When your job runs, an automated script written in either Python or Scala extracts the data from your data source, transforms it, and loads the transformed data into the data target. AWS Glue solves the business problems that come with the need for analyzing heterogenous data types.

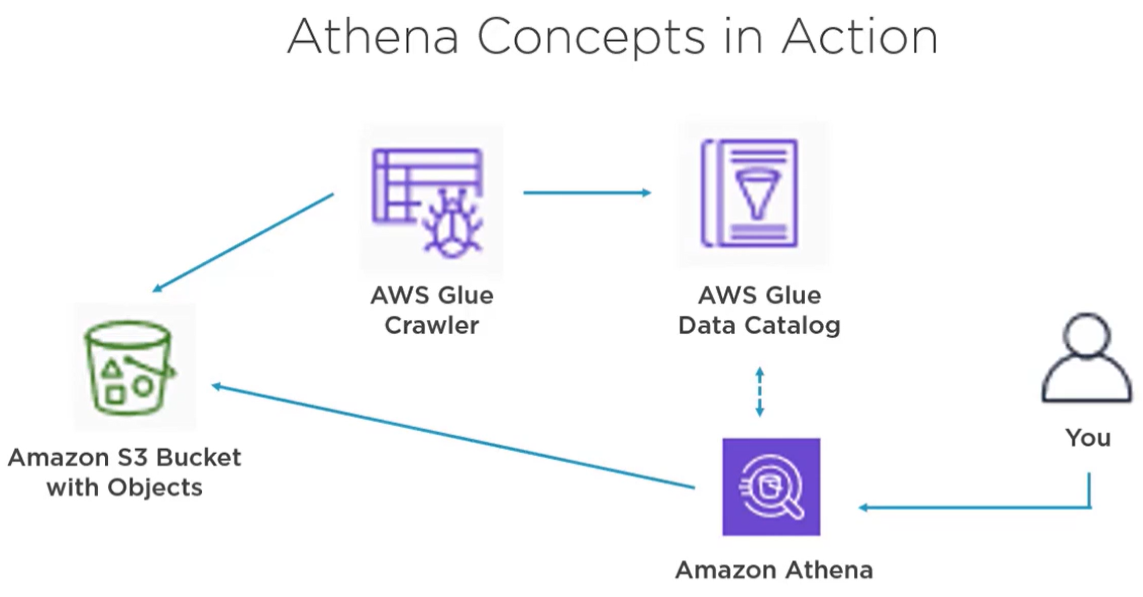


* The **AWS Glue Data Catalog** is a central repository to store structural and operational metadata for all your data assets, even from on premises.
* The **AWS Glue crawlers** connect to data stores to extract the schema and other statistics about the data. Then it populates the Glue Data Catalog with this metadata.
* Glue automatically generates the ETL code necessary to extract data from the source, transform the data to match the target schema, and load it into the target by pointing to your data source and target, making it easy to prepare and load data for analytics.

**Amazon Athena**



* **Amazon Athena** is an interactive query service that makes it easy to analyze data directly in Amazon S3. With Athena, **there's no need for complex ETL jobs to prepare your data for analysis.** This makes it easy for anyone with SQL skills to quickly analyze large-scale datasets.
* Amazon Athena uses Presto, an open-source distributed SQL query engine optimized for low-latency ad hoc analysis of data. You can use Athena to run ad hoc queries using ANSI SQL without the need to aggregate or load the data into Athena, with full support for large joins, window functions, and arrays.
* Athena helps you analyze unstructured, semi-structured, and structured data stored in Amazon S3. Examples include SVC, JSON, or columnar data formats such as Apache Parquet and Apache ORC. Athena is very fast.



* **Amazon Athena integrates out of the box with AWS Glue.** With the Glue Data Catalog, you'll be able to create a unified metadata repository across various services, crawl data sources to discover and populate your Data 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 query performance and reduce costs.
* This conceptual diagram illustrates Athena's concepts in action. An AWS Glue crawler crawls the raw dataset in an S3 raw zone bucket. The crawler writes metadata into the AWS Glue Data Catalog. You query those raw datasets in Athena. Athena uses schema definition to read raw datasets from S3 and then returns the results. **Athena solves the business problems associated with querying heterogenous data without the need to first perform ETL.**
* Athena is an interactive query service. This makes it easy to analyze data directly in S3 using standard SQL.
* Athena can process unstructured, semi-structured, and structured datasets.
* Athena is serverless, there's no infrastructure to manage whatsoever.
* Athena scales automatically, giving fast results even with large datasets and complex queries. Athena executes queries in parallel, so results are returned in seconds.
* Athena can query encrypted data in S3 and write encrypted data back to another S3 bucket.