EMR, Glue Jobs, Lambda, Kinesis Analytics, RedShift

Apache Hadoop on AWS

Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.

Read more about Hadoop here: https://aws.amazon.com/emr/details/hadoop/what-is-hadoop/

Amazon EMR

Amazon EMR is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark, on AWS to process and analyze vast amounts of data. By using these frameworks and related open-source projects, such as Apache Hive and Apache Pig, you can process data for analytics purposes and business intelligence workloads. Additionally, you can use Amazon EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.

Read more about Amazon EMR here:
https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-what-is-emr.html

Amazon EMR Serverless

Amazon EMR Serverless is a serverless option in Amazon EMR that makes it easy for data analysts and engineers to run open-source big data analytics frameworks without configuring, managing, and scaling clusters or servers. You get all the features and benefits of Amazon EMR without the need for experts to plan and manage clusters.

Read more about Amazon EMR Serverless here: https://aws.amazon.com/emr/serverless/

AWS Glue DataBrew

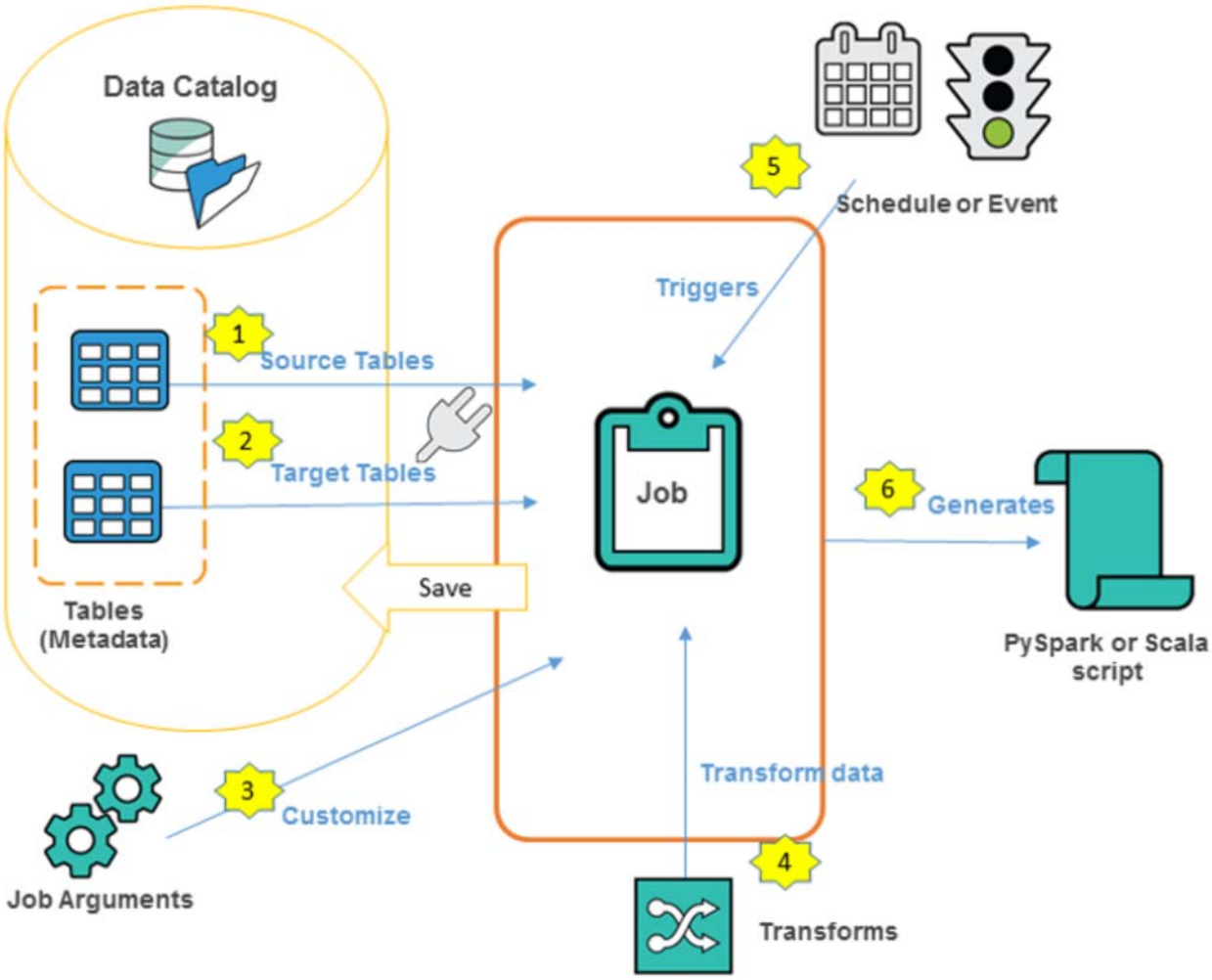
AWS Glue DataBrew is a new visual data preparation tool that makes it easy for data analysts and data scientists to clean and normalize data to prepare it for analytics and machine learning. You can choose from over 250 pre-built transformations to automate data preparation tasks, all without the need to write any code. You can automate filtering anomalies, converting data to standard formats, and correcting invalid values, and other tasks. After your data is ready, you can immediately use it for analytics and machine learning projects. You only pay for what you use - no

upfront commitment.

Read more about AWS Glue DataBrew here: https://aws.amazon.com/glue/features/databrew/

AWS Glue Jobs

A job is the business logic that performs the extract, transform, and load (ETL) work in AWS Glue. When you start a job, AWS Glue runs a script that extracts data from sources, transforms the data, and loads it into targets. You can create jobs in the ETL section of the AWS Glue console.



Read more about authoring AWS Glue jobs here:
https://docs.aws.amazon.com/glue/latest/dg/author-job.html

AWS Lambda

AWS Lambda is a compute service that lets you run code without provisioning or managing servers. AWS Lambda runs your code only when needed and scales automatically, from a few requests per day to thousands per second. You pay only for the compute time you consume - there

is no charge when your code is not running. With AWS Lambda, you can run code for virtually any type of application or backend service - all with zero administration. AWS Lambda runs your code on a high-availability compute infrastructure and performs all of the administration of the compute resources, including server and operating system maintenance, capacity provisioning and automatic scaling, code monitoring and logging.

When using AWS Lambda, you are responsible only for your code. AWS Lambda manages the compute fleet that offers a balance of memory, CPU, network, and other resources. This can be helpful when processing incoming data for your data lake being hosted on AWS.

Read more about AWS Lambda here:
https://docs.aws.amazon.com/lambda/latest/dg/welcome.html

Amazon Athena

Amazon Athena is an interactive query service that makes it easy to analyze data directly in Amazon Simple Storage Service (Amazon S3) using standard SQL. With a few actions in the AWS Management Console, you can point Athena at your data stored in Amazon S3 and begin using standard SQL to run ad-hoc queries and get results in seconds.

Read more about Athena here: https://docs.aws.amazon.com/athena/latest/ug/what-is.html

Amazon RedShift

Amazon Redshift makes it simple and cost effective to run high performance queries on petabytes of structured data so that you can build powerful reports and dashboards using your existing business intelligence tools.

Read more about Amazon RedShift here: https://aws.amazon.com/redshift/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc

Amazon Kinesis Data Analytics

With Amazon Kinesis Data Analytics for SQL Applications, you can process and analyze streaming data using standard SQL. The service enables you to quickly author and run powerful SQL code against streaming sources to perform time series analytics, feed real-time dashboards, and create real-time metrics.

To get started with Kinesis Data Analytics, you create a Kinesis data analytics application that continuously reads and processes streaming data. The service supports ingesting data from Amazon Kinesis Data Streams and Amazon Kinesis Data Firehose streaming sources. Then, you author your SQL code using the interactive editor and test it with live streaming data. You can also configure destinations where you want Kinesis Data Analytics to send the results. Kinesis Data

Analytics supports Amazon Kinesis Data Firehose (Amazon S3, Amazon Redshift, Amazon Elasticsearch Service, and Splunk), AWS Lambda, and Amazon Kinesis Data Streams as destinations.

Read more about Amazon Kinesis Data Analytics here: https://docs.aws.amazon.com/kinesisanalytics/latest/dev/what-is.html

Amazon OpenSearch Service

Please note that Amazon Elasticsearch Service (Amazon ES) has a successor called **Amazon OpenSearch Service.** If in the videos you hear the instructors referring to Amazon ElasticSearch Service, please know they are talking about the service Amazon OpenSearch Service.

Amazon OpenSearch Service is an open source, distributed search and analytics suite derived from Elasticsearch. Elasticsearch is a popular open-source search and analytics engine for use cases such as log analytics, real-time application monitoring, and clickstream analysis. With OpenSearch Service, you get access to the latest versions of OpenSearch Service, support for 19 versions of Elasticsearch (1.5 to 7.10 versions),

Read more about Amazon OpenSearch Service here: https://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/what-is-amazon-elasticsearch-service.https://docs.aws.amazon.com/opensearch-service/latest/developerguide/gsg.html