Cloud Guru - 1 - Data Collection Quiz

Tips for Exam

Before You Begin

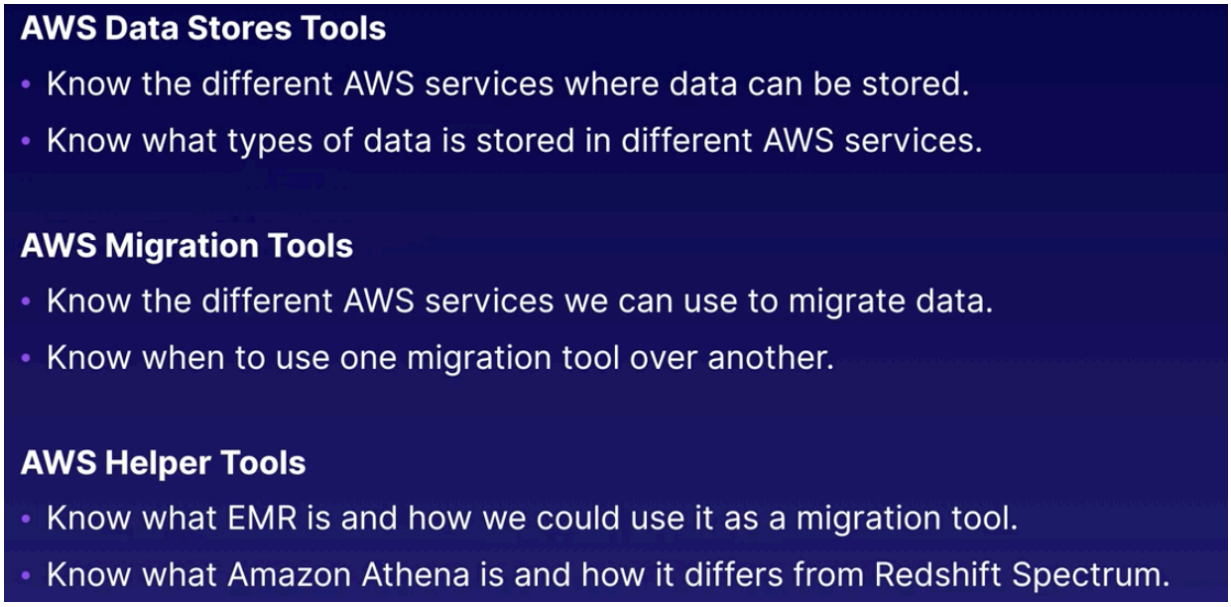
- Understand that before we gather input data we must formulate the problem we are trying to solve.
- Know how we can measure success and what your goals are.
- Determine if Machine Learning is even necessary.
- Understand what type of data is available to help solve problem.

Good Data

- Understand what makes up “good” data and why having good data is important.
- Understanding what “good” and “bad” data looks like.

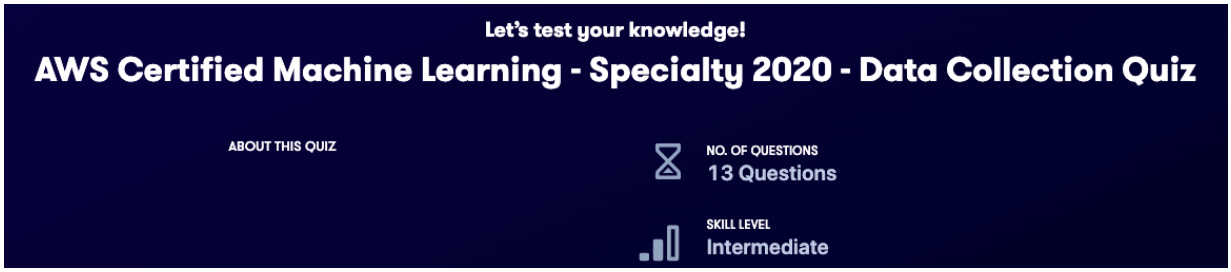
Data Terminology

- Know how to identify columns/attributes and rows/observations within a dataset.
- Know the difference in structured, semi-structured, and unstructured data.
- Know the different types of data repositories (databases, data warehouses, data lakes).
- Understand the differences between labeled data and unlabeled data.
- Be able to recognize categorical features and continuous features.
- Know terms like corpus, ground truth, time series data, and image data.



End goal is to get the data in S3

Quiz



QUESTION 1

You have been tasked with converting multiple JSON files within a S3 bucket to Apache Parquet format. Which AWS service can you use to achieve this with the LEAST amount of effort?

- Create an EMR cluster to run an Apache Spark job to process the data the Apache Parquet and output newly formatted files into S3.
- Create a Data Pipeline job that reads from your S3 bucket and sends the data the EMR. Create an Apache Spark job to process the data the Apache Parquet and output newly formatted files into S3.
- Create a Lambda function that reads all of the objects in the S3 bucket. Loop through each of the objects and convert from JSON to Apache Parquet. Once the conversion is complete, output newly formatted files into S3.
- ✓ Create an AWS Glue Job to convert the S3 objects from JSON to Apache Parquet, then output newly formatted files into S3.

Good work!

AWS Glue makes it super simple to transform data from one format to another. You can simply create a Job that takes in data defined within the Data Catalog and outputs in any of the following formats: avro csv, ion, grokLog, json, orc, parquet, glueparquet, xml

QUESTION 2

When you train your model in SageMaker, where does your training dataset come from?

- RedShift
- ✓ S3
- RDS
- DynamoDB

Good work!

Generally, we store our training data in S3 to use for training our model.

QUESTION 3

Your organization has given you several different sets of key-value pair JSON files that need to be used for a machine learning project within AWS. What type of data is this classified as and where is the best place to load this data into?

- Semi-structured data, stored in DynamoDB.
- Unstructured data, stored in S3.
- Structured data, stored in RDS.
- ✓ Semi-structured data, stored in S3.

Good work!

Key-value pair JSON data is considered Semi-structured data because it doesn't have a defined structure, but has some structural properties. If our data is going to be used for a machine learning project in AWS, we need to find a way to get that data into S3.

QUESTION 4

You are trying to set up a crawler within AWS Glue that crawls your input data in S3. For some reason after the crawler finishes executing, it cannot determine the schema from your data and no tables are created within your AWS Glue Data Catalog. What is the reason for these results?

- The crawler does not have correct IAM permissions to access the input data in the S3 bucket.
- The bucket path for the input data store in S3 is specified incorrectly.
- The checkbox for 'Do not create tables' was checked when setting up the crawler in AWS Glue.
- ✓ AWS Glue built-in classifiers could not find the input data format. You need to create a custom classifier.

Good work!

AWS Glue provides built-in classifiers for various formats, including JSON, CSV, web logs, and many database systems. If AWS Glue cannot determine the format of your input data, you will need to set up a custom classifier that helps AWS Glue crawler determine the schema of your input data.

QUESTION 5

In general within your dataset, what is the minimum number of observations you should have compared to the number of features?

- [x] 10 times as many observations as features.
- [ ] 10,000 times as many observations as features.
- [ ] 1000 times as many observations as features.
- [ ] 100 times as many observations as features.

Good work!

We need a large, robust, feature-rich dataset. In general, having AT LEAST 10 times as many observations as features is a good place to start. So for example, we have a dataset with the following features: id, date, full review, full review summary, and a binary safe/unsafe tag. Since id is just an identifier, we have 4 features (date, full review, full review summary, and a binary safe/unsafe tag). This means we need AT LEAST 40 rows/observations.

QUESTION 6

You have been tasked with collecting thousands of PDFs for building a large corpus dataset. The data within this dataset would be considered what type of data?

- Structured
- Semi-structured
- Relational
- ✓ Unstructured

Good work!

Since PDFs have no real structure to them, like key-value pairs or column names, they are considered unstructured data.

QUESTION 7

Which Amazon service allows you to build a high-quality training labeled dataset for your machine learning models? This includes human workers, vendor companies that you choose, or an internal, private workforce.

- S3
- Jupyter Notebooks
- Lambda
- ✓ SageMaker Ground Truth

Good work!

You could use Jupyter Notebooks or Lambda to help automate the labeling process, but SageMaker Ground Truth is specifically used for building high-quality training datasets.

QUESTION 8

You are a ML specialist within a large organization who helps job seekers find both technical and non-technical jobs. You've collected data from a data warehouse from an engineering company to determine which skills qualify job seekers for different positions. After reviewing the data you realise the data is biased. Why?

- [x] The data collected needs to be from the general population of job seekers, not just from a technical engineering company.
- [ ] The data collected only has a few attributes. Attributes like skills and job title are not included in the data.
- [ ] The data collected has missing values for different skills for job seekers.
- [ ] The data collected is only a few hundred observations making it bias to a small subset of job types.

Good work!

It's important to know what type of questions we are trying to solve. Since our organization helps both technical and non-technical job seekers, only gathering data from an engineering company is biased to those looking for technical jobs. We need to gather data from many different repositories, both technical and non-technical.

QUESTION 9

You are a ML specialist within a large organization who needs to run SQL queries and analytics on thousands of Apache logs files stored in S3. Which set of tools can help you achieve this with the LEAST amount of effort?

- Data Pipeline and RDS
- Redshift and Redshift Spectrum
- ✓ AWS Glue Data Catalog and Athena
- Data Pipeline and Athena

Good work!

Using Redshift/Redshift Spectrum and Data Pipeline/RDS could work, but require much more effort in setting up and provisioning resources. Using AWS Glue you can use a crawler to crawl the logs files in S3. This will create structured tables within your AWS Glue database. These tables can then be queried using Athena. This solution requires the least amount of effort.

QUESTION 10

An organization needs to store a mass amount of data in AWS. The data has a key-value access pattern, developers need to run complex SQL queries and transactions, and the data has a fixed schema. Which type of data store meets all of their needs?

- [x] RDS
- [ ] S3
- [ ] Athena
- [ ] DynamoDB

Good work!

Amazon RDS handles all these requirements. Transactional and SQL queries are the important terms here. Although RDS is not typically thought of as optimized for key-value based access, using a schema with a primary key can solve this. S3 has no fixed schema. Although Amazon DynamoDB provides key-value access and consistent reads, it does not support SQL based queries. Finally, Athena is used to query data on S3 so this is not a data store on AWS.

QUESTION 11

You are a ML specialist within a large organization who needs to run SQL queries and analytics on thousands of Apache logs files stored in S3. Your organization already uses Redshift as their data warehousing solution. Which tool can help you achieve this with the LEAST amount of effort?

- Athena
- Apache Hive
- S3 Analytics
- ✓ Redshift Spectrum

Good work!

Since the organization already uses Redshift as their data warehouse solution, Redshift spectrum would require less effort than using AWS Glue and Athena.

QUESTION 12

You are a ML specialist who is setting up a ML pipeline. The amount of data you have is massive and needs to be set up and managed on a distributed system to efficiently run processing and analytics on. You also plan to use tools like Apache Spark to process your data to get it ready for your ML pipeline. Which setup and services can most easily help you achieve this?

- Redshift out-performs Apache Spark and should be used instead.
- Self-managed cluster of EC2 instances with Apache Spark installed.
- Multi AZ RDS Read Replicas with Apache Spark installed.
- ✓ Elastic Map Reduce (EMR) with Apache Spark installed.

Good work!

Amazon's EMR allows you to set up a distributed Hadoop cluster to process, transform, and analyze large amounts of data. Apache Spark is a processing framework and programming model that helps you do machine learning, stream processing, or graph analytics using Amazon EMR clusters.

QUESTION 13

You are a ML specialist working with data that is stored in a distributed EMR cluster on AWS. Currently, your machine learning applications are compatible with the Apache Hive Metastore tables on EMR. You have been tasked with configuring Hive to use the AWS Glue Data Catalog as its metastore. Before you can do this you need to transfer the Apache Hive metastore tables into an AWS Glue Data Catalog. What are the steps you'll need to take to achieve this with the LEAST amount of effort?

Choose 2

- [x] Run a Hive script on EMR that reads from your Apache Hive Metastore, exports the data to an intermediate format in Amazon S3, and then imports that data into the AWS Glue Data Catalog.
- [ ] Create a second EMR cluster that runs an Apache Spark script to copy the Hive metastore tables from the original EMR cluster into AWS Glue.
- [x] Setup your Apache Hive application with JDBC driver connections, then create a crawler that crawlers the Apache Hive Metastore using the JDBC connection and creates an AWS Glue Data Catalog.
- [x] Create a Data Pipeline job that reads from your Apache Hive Metastore, exports the data to an intermediate format in Amazon S3, and then imports that data into the AWS Glue Data Catalog.
- [ ] Create DMS endpoints for both the input Apache Hive Metastore and the output data store S3 bucket, run a DMS migration to transfer the data, then create a crawler that creates an AWS Glue Data Catalog.

Sorry!

Correct Answer

The benefit of using Data Catalog (over Hive Metastore) is because it provides a unified metadata repository across a variety of data sources and data formats, integrating with Amazon EMR as well as Amazon RDS, Amazon Redshift, Redshift Spectrum, Athena, and any application compatible with the Apache Hive metastore. We can simply run a Hive script to query tables and output that data in CSV (or other formats) into S3. Once that data is on S3, we can crawl it to create a Data Catalog of the Hive Metastore or import the data directly from S3.