Brock Tubre

TRAINING ARCHITECT

John Hanna

TRAINING ARCHITECT

AWS Certified Data Analytics - Specialty Study Guide

Course Introduction

Welcome!

In late 2019, AWS announced that they would be renaming the AWS Big Data Specialty certification to the AWS Data Analytics Specialty certification. There are some carryover topics with these changes, but many of the topics are brand new and more in-depth.

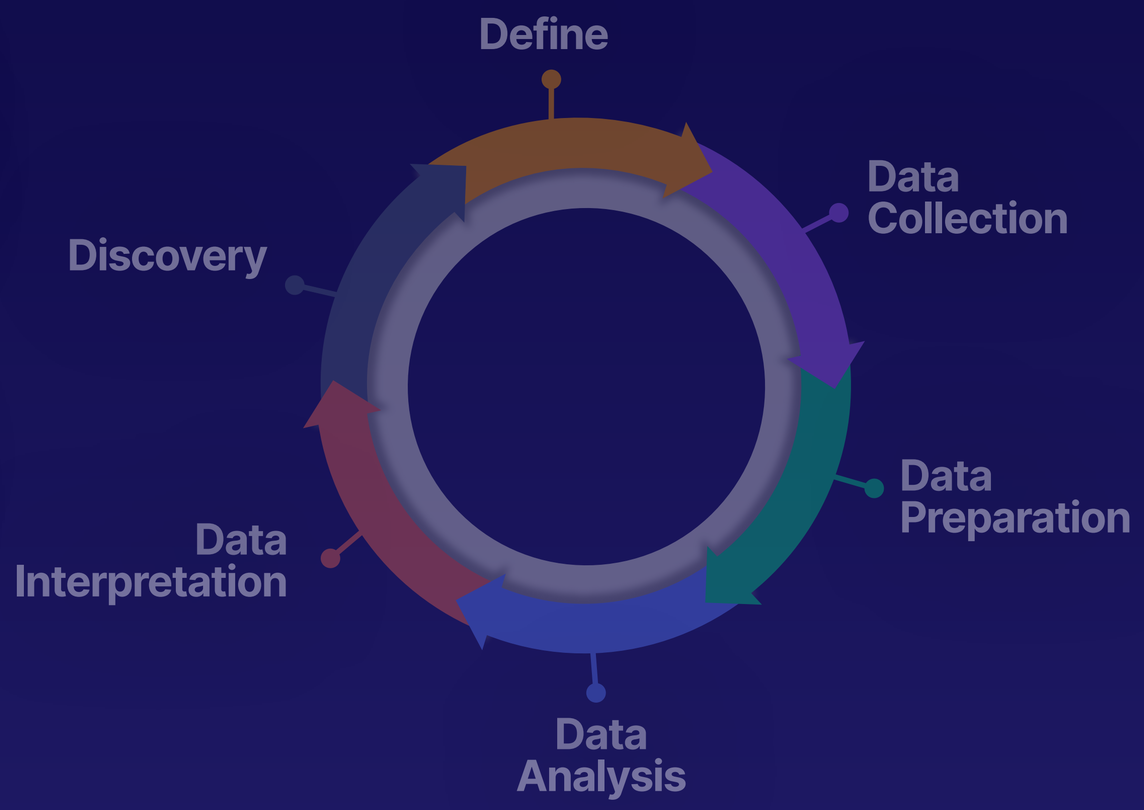
This course is designed to teach you the new services and tools that you can use in AWS to build data analytics applications, as well as build and manage the lifecycle of collecting, storing, processing, and visualizing your data. Just because the name of the certification is changing doesn't mean big data is going away. In fact, data is growing for organizations at an alarming rate, and that means there is a growing need for engineers who know how to gather insight from that data.

That is where the AWS Data Analytics Specialty certification fits into the picture. This certification validates a deep understanding of AWS data analytics services and understands how they integrate with each other. Once you have passed the exam, you should be able to explain how AWS data analytics services fit in the data life cycle and answer important questions your organization's data is providing.

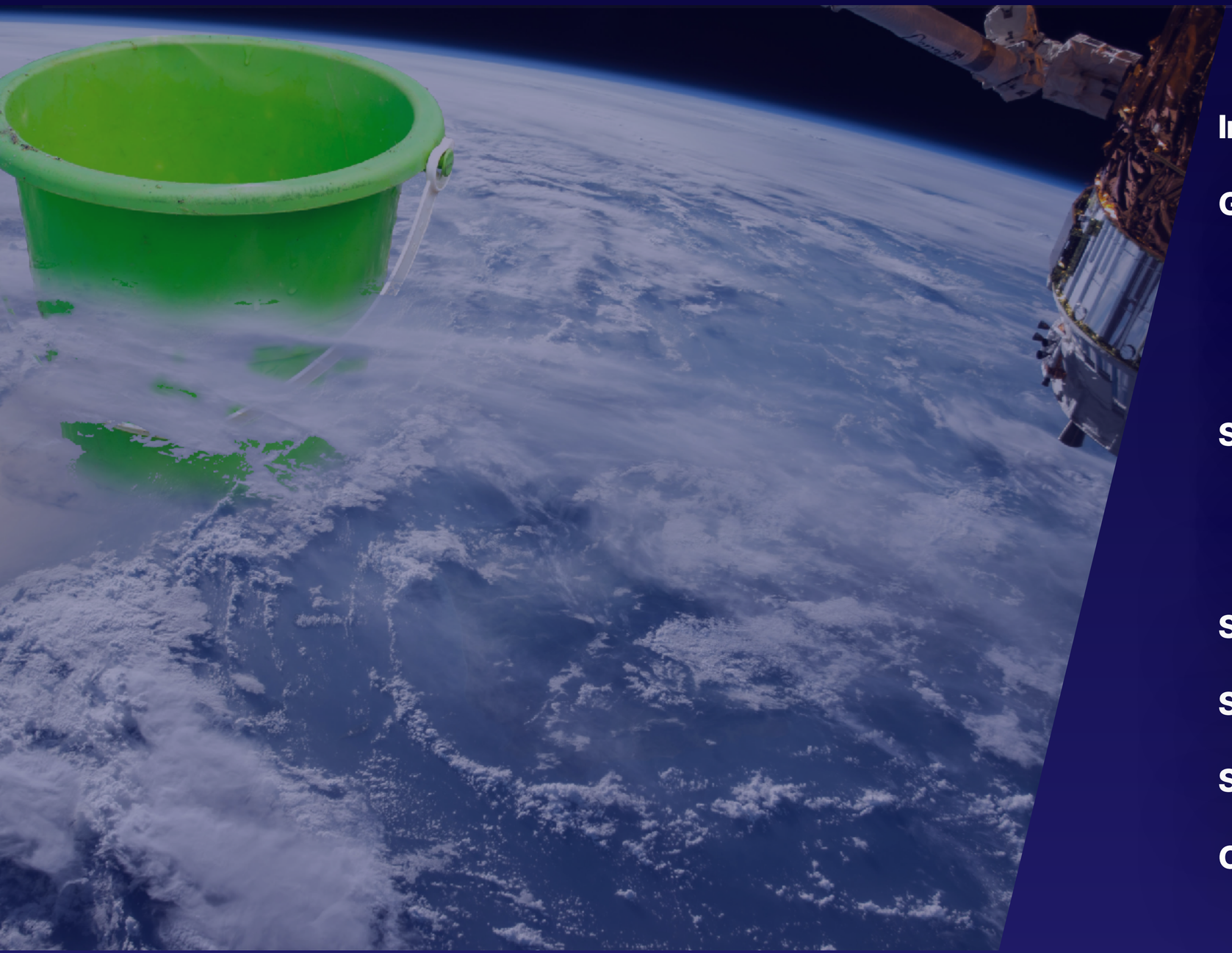
This study guide provides links to supplemental material for topics covered in this course. We strongly recommend spending some time exploring these links to further your knowledge and gain deeper insight into AWS.

Resources

Course GitHub Repo



Simple Storage Service



Resources

Introduction to S3

S3 Documentation

Getting Data into S3

- Lesson Files
- S3 Getting Started
- AWS CLI S3
- AWS CLI S3API
- AWS S3 Boto3 (Python SDK)

S3 Multipart Upload

- Example Script
- Boto3: Create a Multipart Upload
- Boto3: Perform an Upload
- Boto3: Complete a Multipart Upload

S3 Storage Classes

Storage Classes Documentation

S3 Lifecycle Policies

Object Lifecycle Management

S3 Security and Encryption

Amazon S3 Security

Checkpoint S3

S3 FAQs

Databases in Amazon Web Services

Resources

Relational Database Service (RDS)

Amazon RDS Resources

Neptune

- Neptune Documentation
- Tinkerpop Gremlin
- SPARQL

DocumentDB

DocumentDB Documentation

Serverless Options

- S3 Select
- Athena Documentation
- DynamoDB Documentation
- Aurora Serverless Documentation

Checkpoint

- RDS FAQs
- Aurora FAQs
- Neptune FAQs
- DocumentDB FAQs
- Athena FAQs
- DynamoDB FAQs

Collecting Streaming Data

Resources

The Kinesis Family

Amazon Kinesis

Kinesis Data Streams (Part 1)

- Whitepaper: Streaming Data Solutions on AWS with Amazon Kinesis
- Building a Streaming Platform with Kinesis Data Stream

Kinesis Data Streams (Part 2)

- CloudFormation Template and Python Script
- Developing Producers Using the Amazon Kinesis Producer Library
- Developing Producers Using the Amazon Kinesis Data Streams API with the AWS SDK for Java
- Writing to Amazon Kinesis Data Streams Using Kinesis Agent
- Developing KCL Consumers

Kinesis Data Firehose (Part 1)

What is Amazon Kinesis Data Firehose

Kinesis Data Firehose (Part 2)

Configuring Settings for Kinesis Data Firehose

Kinesis Video Streams

What is Kinesis Video Streams

Kinesis Data Analytics (Part 1)

Amazon Kinesis Data Analytics for SQL Applications: How It Works

Collecting Streaming Data (Cont.)

Resources

Kinesis Data Analytics (Part 2)

- Create Your Starter Amazon Kinesis Data Analytics Application
- SQL Query Demo - Download

Amazon Managed Service for Kafka (MSK)

- Apache Kafka
- Getting Started Using Amazon MSK

Checkpoint

- Kinesis Data Streams FAQs
- Kinesis Data Firehose FAQs
- Kinesis Data Analytics FAQs
- Kinesis Video Streams FAQs
- MSK FAQs
- AWS re:Invent 2018: High Performance Data Streaming with Amazon Kinesis: Best Practices

Data Collection and Getting Data into AWS

Resources

Direct Connection , Snowball, Snowball Edge, Snowmobile

- What is AWS Direct Connect?
- AWS Snow Family
- Extending Data Centers to Cloud: Connectivity Options and Best Practices
- Data Migration and Edge Computing with the AWS Snow Family

Database Migration Service

- Big Data Analytics Architectural Patterns and Best Practices
- Getting Started with AWS Database Migration Service

Data Pipeline

- What is AWS Data Pipeline?
- Data Pipeline Concepts

Lambda, API Gateway, and CloudFront (Part 1)

- Big Data, Analytics, and Machine Learning on AWS Lambda
- Moving to Event Driven Architectures

Lambda, API Gateway, and CloudFront (Part 2)

- Lambda Limits
- Using AWS Lambda with Amazon Kinesis

Comparing Our Options

Get Your Data to AWS: How to Choose and Use Data Migration Services

Data Collection and Getting Data into AWS (Cont.)

Resources

Checkpoint

- Snow Family FAQs Direct Connect FAQs
- Lambda FAQs
- CloudFront FAQs
- API Gateway FAQs
- Cloud Data Migration in AWS

Amazon Elastic Map Reduce (EMR)

Resources

Apache Hadoop and EMR Software Collection

- Apache Hadoop
- Amazon Elastic Map Reduce Overview
- EMR Storage Options

EMR Architecture

Overview of Amazon EMR Architecture

EMR Operations - Transient vs Long-Running

- Configure Cluster to Auto-Terminate (Transient)
- Controlling Cluster Termination

EMR Operations - Choosing an Instance Type

Supported EMR Instance Types

EMR Operations - Choosing the Right Number of Instances

Cluster Configuration Guidelines and Best Practices

EMR Operations - On-Demand and Spot Instances

Instance Purchase Options

EMR Operations - Monitoring and Resizing Clusters

- Monitor EMR Metrics with CloudWatch
- Manually Resizing Your EMR Cluster
- Using Automatic Scaling with a Custom Policy for Instance Groups
- Auto-Scaling EMR Cluster - Blog

Amazon Elastic Map Reduce (EMR) (Cont.)

Resources

EMR File Storage and Compression

- Types of Input Amazon EMR Can Accept
- What Formats can Amazon EMR Return?
- Writing to an S3 Bucket You Do Not Own from EMR
- How to Get Data Into Amazon EMR

Checkpoint

- Customizing Cluster and Application Configuration with Earlier AMI Versions of Amazon EMR FAQs
- Deep Dive into EMR re:Invent
- EC2 Pricing
- S3 Pricing
- EMR Best Practices

Using Redshift

Resources

Introduction

Redshift Documentation

Redshift Architecture

Data Warehouse System Architecture

Redshift Table Design

- Data Types
- Compression Encoding
- Sort Keys
- Distribution Styles
- Constraints

Redshift Spectrum

Redshift Spectrum Documentation

Redshift Maintenance and Operations

Resources

Launching a Redshift Cluster

- Getting Started with Amazon Redshift
- AWS CLI Redshift create-cluster
- Boto3 Redshift create_cluster

Resizing a Redshift Cluster

- Managing Redshift Clusters
- AWS CLI resize-cluster
- AWS CLI describe-node-configuration-options
- Boto3 Redshift resize_cluster
- Boto3 Redshift describe_node_configuration_options

Utilizing Vacuum and Deep Copy

- Redshift VACUUM
- Redshift Performing Deep Copy

Backup and Restore

- Redshift Snapshots
- Redshift COPY
- Redshift UNLOAD

Monitoring

Monitoring Redshift

Checkpoint

Redshift FAQs

AWS Glue, Athena, and QuickSight

Resources

Glue Data Catalog (Part 1)

- What is AWS Glue?
- AWS Glue Concepts
- AWS Glue Components

Glue Data Catalog (Part 2)

- Populating the AWS Glue Data Catalog
- Defining Crawlers
- Converting Semi-Structured Schemas to Relational Schemas

Glue Jobs (Part 1)

- Authoring Jobs
- Editing Scripts in AWS Glue
- AWS Glue Triggers

Glue Jobs Demo (Part 2)

- Adding Jobs in AWS Glue
- Built-in Transformations

Glue Jobs (Part 3)

- Monitoring for DPU Capacity Planning
- Format Options for ETL Inputs and Outputs in AWS Glue
- Connect to and run ETL jobs across multiple VPCs using a dedicated AWS Glue VPC

AWS Glue, Athena, and QuickSight (Cont.)

Resources

Job Bookmarks

Tracking Processed Data Using Job Bookmarks

Getting Started with Athena

- What is Athena?
- AWS re:Invent 2019: Deep Dive into Amazon Athena
- Using Federated Queries
- Athena Service Integrations
- Using CTAS and INSERT INTO for ETL and Data Analysis

When to Use Athena

- S3 Select
- Selecting Objects with S3 Select and Glacier Select
- Glacier Select
- When Should I Use Athena?

QuickSight Visualizations and Dashboards

- Gapminder World Data Site
- What is Amazon QuickSight?
- How QuickSight Works
- Supported Data Sources
- Working With Visualizations
- Exploring Dashboards in QuickSight

AWS Glue, Athena, and QuickSight (Cont.)

Resources

QuickSight Security and Authentication

- Security in QuickSight
- QuickSight Endpoints in AWS
- Identity and Access Management in QuickSight

Checkpoint

- AWS Glue FAQs
- AWS Glue Developer Guide - MUST READ
- Athena FAQs
- QuickSight FAQs

Elasticsearch

Resources

Using Elasticsearch

- Amazon Elasticsearch Service Documentation
- Elasticsearch API Reference
- Elasticsearch Query Data Search Language (DSL)
- Elasticsearch SQL Access

Visualizing Elasticsearch Data

Kibana Visualization Examples

Checkpoint

Amazon Elasticsearch Service FAQs

Security Services

Resources

IAM

AWS Identity and Access Management FAQs

KMS

AWS Key Management Service FAQs

Secrets Manager

AWS Secrets Manager FAQs

VPC Network Security Features

Amazon VPC FAQs

Wrap Up

Resources

LinkedIn

- Brock's LinkedIn
- John's LinkedIn

Twitter

- Brock's Twitter
- John's Twitter

ACG Forum

AWS Certified Data Analytics - Specialty Forums

Community Slack

linuxacademycommunity.slack.com/