Quality Thought®

Bigdata on AWS

Mode of Training: Online Training

Name of Trainer: Mr. Sonali

Course Curriculum

Let's Draw

- Hadoop Architecture, Map-Reduce, HDFS, Yarn
- Spark (Python + Spark)
- Python
- Cloudwatch
- S3
- Lambda
- Glue
- Airflow
- Shell scripting
- DynamoDB
- Athena
- EMR
- Oozie
- Hive
- Hbase
- Spark Streaming
- Kafka
- Kinesis
- SNS
- SQS

==========

- **Hadoop Architecture, Map-Reduce, HDFS, Yarn**
- **Spark (Python + Spark)**
  - ✓ Introduction
  - ✓ Spark Fundamentals
  - ✓ Spark and it's Ecosystem
  - ✓ Spark vs Hadoop
  - ✓ RDD Fundamentals
  - ✓ Spark Transformations, Actions and Operations
  - ✓ Job, Stages and Task
  - ✓ RDD Creation
  - ✓ Spark SQL
  - ✓ Spark Dataframe basics
  - ✓ Reading files of different formats
  - ✓ Spark SQL Hive Integration
  - ✓ Sqoop on Spark
  - ✓ Twitter Streaming through Flume

- **Python**
  - ✓ Introduction of Why Python? Installing Python, Python 2 vs Python 3
  - ✓ Types in Python, Integers & Floats, String, Booleans, None Lists, Dictionary, Other Data Types
  - ✓ Statements in Python, If, Loops, Break & Continue, While
  - ✓ Exceptions in Python
  - ✓ Functions
  - ✓ File Management in Python

- ✓ Yield
- ✓ Lambda Functions
- ✓ Object Oriented Programming with Python o Classes o Methods o Constructors o Instance & Class Attributes o Inheritance & Polymorphism
- ✓ Python Tips & Tricks
- ✓ Strings & Collections
- ✓ Modularity
- ✓ Handling Exceptions

CloudWatch

- ✓ Monitoring and logging basics
- ✓ CloudWatch Metrics o Predefined o Custom Cloud Watch Metrics
- ✓ CloudWatch Alarms
- ✓ CloudWatch Billing Alarms
- ✓ CloudWatch Free tier Billing Alerts
- ✓ CloudWatch Logs
- ✓ CloudWatch Log Agent
- ✓ CloudWatch Log Insights
- ✓ CloudWatch Events
- ✓ CloudWatch and Simple Notification Services Integration

S3

- ✓ Architectural Overview
- ✓ Buckets
- ✓ Objects and Folders
- ✓ Storage Tiers
- ✓ Lifecycle Policies
- ✓ Versioning
- ✓ Locking
- ✓ Access to S3 Buckets
- ✓ Static Website Hosting
- ✓ S3 Cross Region Replications, CLI, CloudFormation

Lambda

- ✓ What is Lambda
- ✓ Serverless concepts
- ✓ Permissions
- ✓ Functions and tools
- ✓ Managing, Configuring & Invoking Functions
- ✓ Lambda Runtimes & Applications
- ✓ Working with Python

Glue

- ✓ What is AWS Glue?
- ✓ When do you use AWS Glue?
- ✓ AWS Glue Benefits
- ✓ AWS Glue Concepts
- ✓ AWS Glue Terminologies
- ✓ How does AWS Glue work?
- ✓ Importing CSV files from S3 into Redshift with AWS Glue

- **Airflow**
  - ✓ Introduction
  - ✓ What is Airflow
  - ✓ Use case and Why need Airflow
  - ✓ What is workflow
  - ✓ A typical workflow
  - ✓ A traditional ETL approach
  - ✓ Problems
  - ✓ Deep in Apache Airflow
  - ✓ Airflow DAG

- **Shell scripting**
  - ✓ Why and What of Shell Scripting
  - ✓ Shell Terminals
  - ✓ Creation & Execution of Shell Scripts
  - ✓ Variables & Variable Scopes
  - ✓ Conditions in Shell Scripts
  - ✓ Iterating with loops
  - ✓ Functions in Shell Scripts
  - ✓ Regular Expressions
  - ✓ Command Piping with grep
  - ✓ Stream Editor
    - Understanding basics of sed
    - Sed commands
  - ✓ AWK Fundamentals

- **dynamoDb**
  - ✓ NoSQL Overview
  - ✓ DynamoDB Concepts
  - ✓ Tables
  - ✓ Backups
  - ✓ Reserved Capacity
  - ✓ Indexes
  - ✓ Transactions
  - ✓ DAX, Dashboard, Clusters, Subnet groups, Parameter, Groups, Events

- **Athena**
  - ✓ Introduction
  - ✓ S3 -> Glue Crawler -> glue catalog -> Athena -> quick sight
  - ✓ Athena Cost Model
  - ✓ Creating Table
  - ✓ Athena Queries

Cluster

- **EMR**
  - ✓ What is EMR
  - ✓ Benefits
  - ✓ EMR Architecture
  - ✓ EMR Applications
  - ✓ Spark on EMR

- ✓ Hadoop on AWS using EMR Tutorial | | S3 | | Athena | | Glue | | QuickSight

- **Oozie**
  - ✓ What is Oozie
  - ✓ Why need Oozie
  - ✓ Running in Example
  - ✓ Oozie – A workflow engine

- **Hive**
  - ✓ What is the requirement of Hive?
  - ✓ What is Hive?
  - ✓ Hive Advantages
  - ✓ Where not to use Hive?
  - ✓ Hive Features
  - ✓ MapReduce vs Hive
  - ✓ Hive Architecture
  - ✓ Partitions in Hive
  - ✓ Why we needed Hive?
  - ✓ What is Hive?
  - ✓ Features of hive
  - ✓ Hive Architecture
  - ✓ Hive Components
  - ✓ Install Hive
  - ✓ Hive Datatypes
  - ✓ Hive Operators
  - ✓ Hive Data Models

- **Redshift**

  - ✓ Traditional Data Warehouse
  - ✓ Amazon Redshift – A to Z
  - ✓ Demo on Amazon Redshift
  - ✓ Importing CSV files from S3 into Redshift with AWS Glue

- **Hbase**
  - ✓ about HBase and the functionalities it can perform.
  - ✓ Why we needed HBase?
  - ✓ What is HBase?
  - ✓ Difference between HBase and HDFS
  - ✓ HBase Storage
  - ✓ Features of HBase
  - ✓ HBase Architecture
  - ✓ NoSQL databases
  - ✓ What is Hbase?
  - ✓ Where to use HBase?
  - ✓ Where not to Use HBase?
  - ✓ The Advent of HBase
  - ✓ Hbase Architecture
  - ✓ What is HBase?
  - ✓ HBase Use Case
  - ✓ Applications of HBase

- ✓ HBase vs RDBMS
- ✓ HBase Storage
- ✓ HBase Architectural Components

- **Cassandra**
  - ✓ What apache cassandra
  - ✓ How it works internally
  - ✓ Detailed architecture
  - ✓ Cql - cassandra query language
  - ✓ Cassandra operations
  - ✓ Cassandra tools
  - ✓ Data modeling in cassandra and many detailed topics like - cassandra data partitions
  - ✓ Consistent hashing, ring and tokens in cassandra
  - ✓ Data replication
  - ✓ Data versioning
  - ✓ Data repairing
  - ✓ Tunable consistency
  - ✓ Failover detection
  - ✓ Gossip protocol
  - ✓ Commit logs
  - ✓ Memtables
  - ✓ Sstables
  - ✓ Storage in cassandra
  - ✓ Functions and operations of cql
  - ✓ Data indexing
  - ✓ Materialized views
  - ✓ Security and roles in cassandra
  - ✓ Json support
  - ✓ Primary and clustered keys in data modeling of cassandra
  - ✓ Snitch
  - ✓ Bootstrap
  - ✓ Read repair
  - ✓ Compression
  - ✓ Data backup
  - ✓ Bulk loading of data
  - ✓ Cassandra metrics
  - ✓ Tools like cqlsh
  - ✓ SStable tool
  - ✓ Cassandra stress and all internal concepts
  - ✓ Understand Cassandra and NoSQL domain.
  - ✓ Create Cassandra cluster for different kinds of applications.
  - ✓ Understand Apache Cassandra Architecture
  - ✓ Design and model Applications for Cassandra.
  - ✓ Port existing application from RDBMS to Cassandra.
  - ✓ Learn to use Cassandra with various programming languages.

Streaming

- **Spark streaming**
  - ✓ Introduction
  - ✓ Introduction to Real-Time Analytics

- ✓ Batch & Real-Time Systems
- ✓ Input Output Connectors
- ✓ Twitter Streaming in Real-Time
- ✓ Features in Spark Streaming to Prevent Data Loss
- ✓ basics of Spark Streaming
- ✓ The various data sources used in streaming and the features of Spark Streaming

You will also come across concepts like caching, checkpointing, and accumulators. Finally, you will see a real-time example of Spark Streaming and do a demo to count the occurrence of words in a file.

- **Kafka**
  - ✓ Need of Messaging System
  - ✓ What is Kafka?
  - ✓ Kafka Features
  - ✓ Kafka Components
  - ✓ Kafka architecture
  - ✓ Installing Kafka
  - ✓ Working with Single Node Single Broker Cluster
  - ✓ What is apache kafka
  - ✓ Architecture of apache kafka
  - ✓ kafka topics & partitions, publisher/subscriber workflow, various cli tools in kafka
  - ✓ How to configure a single node and how to configure multi node cluster setup.

- **Kinesis**
  - ✓ What is AWS Kinesis?
  - ✓ Advantages
  - ✓ Capabilities
  - ✓ Use Cases
  - ✓ Kinesis vs SQS
  - ✓ How it works?
  - ✓ AWS - Kinesis Data Stream - Hands On!

- **SNS**
  - ✓ Introduction
  - ✓ SNS Topics
  - ✓ SNS Subscriptions and SNS Subscription protocols
  - ✓ SNS Push notifications

- **SQS**
  - ✓ What is Amazon SQS
  - ✓ Creating Queues
  - ✓ Adding Permissions to Queue
  - ✓ Sending, Receiving and Deleting a Message
  - ✓ Subscribing a Queue
  - ✓ Purging a Queue
  - ✓ Configuring Queues