M

aster Big Data with PySpark and AWS

Bigdata Software installation (In Windows & mac & Ubuntu)

- Intellij
- Pycharm
- Anaconda
- Hadoop 2.7.2
- Spark 2.4.8
- Kafka 2.4.0
- Git
- Sbt
- Sql-workbench
- Java 8
- Scala 2.11.12
- Putty
- WinSCP

Introduction to Bigdata and Hadoop

- What is Bigdata?
- What is Hadoop
- What is Spark
- What is Nosql databases
- Difference between Hadoop, Spark
- Common Bigdata problems
- Hadoop Ecosystem

Python basics

- python introduction
- math operation
- Basic datatypes
- variables
- Lists
- tuples& string
- dictionaries &sets
- NumPy &Arrays
- pandas dataframe

- read &write data
- if else for loop, while
- functions
- cleaning data
- try catch
- dates

SQL topics

- Select * from table
- group by table
- join table
- self-join
- having clause
- functions
- Date functions
- Window functions
- different type joins
- with case operators
- DDL commands (create, alter, drop, rename, truncate)
- DML commands (Insert update delete merge)

AWS Introduction

EC2:

- Create Windows/mac/Linux servers
- Create a sample website
- Autoscaling
- image

Athena

- What is serverless computing?
- Athena process json, csv data
- Recommended approaches

S3:

- store data,
- Client mode submit s3 commands.
- Get data from various sources and store
- S3 bucket Policies

IAM (Identity and Access Management)

- Users
- Groups
- Roles
- Custome policies

Redshift:

- Load data from S3 process data
- Sortkey, Distkey power
- Redshift architecture
- Get data from various sources

Glue:

- How to process csv, json data using Glue
- Get Athena data using glue
- Crawler, Job execute Pyspark and Scala spark
- Glue architecture/internals
- Advanced concepts & best practice

RDS:

- Create different databases
- create sample tables and process
- best practice/low cost
- Practice oracle MySQL using rds.

EMR:

- Practice Py-spark, hive,
- Create EMR (Elastic Map Reduce) cluster and process
- EMR vs ec2
- Hive internals sample programs
- Sqoop import data from RDS store in s3

Hadoop Ecosystem

HDFS:

- What is HDFS?
- Hadoop architecture
- How HDFS replicate data
- Limitations in Hadoop.
- Namenode Importance
- Datanode responsibilities

ode
- High Availability
- Hdfs commands Handson
- Hadoop 1.x Vs 2.x Vs 3.x

Yarn:

- Secondary name node
  - Daemons in Yarn
  - Node manager
  - Application master
  - Resource Manager
  - Yarn Commands
  - How Yarn allocates resources
  - Container
  - How spark /Mapreduce running in Yarn

Hive basics: (90% hands-on)

- Hive architecture
- Sql Vs HQL
- How to process CSV data
- How to process Json data
- Serdes
- Partition
- Bucketing
- Orc vs Parquet importance
- Limitation in Hive

Sqoop (90% hands-on)

- Sqoop architecture
- Import data from Oracle
- Import data from MySQL
- Import data from MsSql data
- Shell script importance in Sqoop
- Import data to Hive
- Compression techniques (parquet, sequence, Avro)
- Best practice

Oozie: (90% hands-on)

- Oozie architecture
- Workflow importance in oozie
- Job.properties importance in oozie
- Coordinator importance in oozie
- Multiple actions in workflow
- How to automate Sqoop & Hive applications using Oozie

Nosql Database Introduction

- What is NOSQL?
- Cap Theorem
- Cassandra

- Cassasndra Architecture
- Cassandra installation in EMR
- Keyspace & tables
- Cassandra Limitation

- Hbase
  - Hbase Architecture
  - Hbase commands
  - Hbase limitations
- Phoenix
- Phoenix Architecture
- Process different type data

Apache Spark Training (98% handson)

Spark Core

- Why Spark why not Hadoop?
- HDFS/Yarn importance in Spark
- Spark architecture
- Different types of APIs
  - RDD (Resilient Distributed Dataset)
  - Dataframe
  - Dataset
- Where using Spark?
- Why spark faster than MapReduce?
- Why /How spark process in Memory?
- Why MapReduce Slow?

RDD Internals:

- RDD Properties
  - Immutability
  - Laziness
  - Fault tolerance
- SparkContext, SqlContext, SparkSession Internals
- Create RDD different ways
- Transformations
- Action
- Commonly used transformations & Actions
- Narrow transformations
- Wide transformations
- Debugging transformations
- Spark web UI

RDD Handson (Where to use, how to use) (90% handson)

(Both Pyspark Scala Spark)

- Map
- FlatMap

- Filter
- Distinct
- ReduceByKey Vs GroupByKey
- SortBy
- Other Transformations & Actions
- Spark-submit
- Minimum 20 RDD use case programs

Spark SQL

Dataframe:

- Convert RDD to Dataframe
- Python Dataframe
- Spark dataframe Introduction
- Dataframe reader
- Dataframe Vs dataset
- Process different type data
  - CSV
  - Json (complex)
  - XML
  - Avro
  - Orc
  - Text data
  - Parquet
  - Spark vs Hive
  - Spark process Hive data
- **Process Different Database data**
  - Oracle
  - MySQL
  - MySQL data analysis
  - Sqoop Vs Spark
  - Data-migration Project
  - ETL project Vs Spark project
- **Process different NoSQL Database data**
  - Spark integrate with HBase and Phoenix
  - Spark Cassandra Integration
  - Spark MongoDb integration

PySpark Advanced Concepts:

- Dataset Api importance
- Spark Memory management
- Resource optimization
- Spark submit num-executors, --executor-cores, --executor-memory importance
- Spark debugging using client mode and web UI.
- How to automate spark using Oozie

- Get data from S3 and process using Databricks
- How to automate spark using Airflow

Spark Streaming

- Spark Streaming Introduction
- Spark streaming introduction
- Micro-batch processing Vs Stream processing
- spark D-stream Api internal
- Get Live data Process using spark
- Realtime Use case
- Spark get Twitter data
- Structure streaming introduction

Kafka internals

- Kafka Architecture
- Producer API
- Consumer API
- Write producer code to get data from sources (Scala, Python)
- Write consumer code to get data from Kafka and flush data to sink.
- Spark Kafka integration
- Get data from web server and process data using spark
- Spark Streaming end to end spark workflow
- How to submit a project using AWS EMR, Azure, Databricks, Cloudera

Apache Nifi introduction

- Nifi Internals
- Different Procedures
- Import/export Templates
- Get data from Rest APIi and process
- Spark Kafka Nifi integration

Other important topics:

- Git commands
- Commit your IntelliJ code to GitHub
- How to improve Ur skills using Google, GitHub, LinkedIn
- Resume preparation
- Mock Tests
- Interview tips