Hadoop-Spark-Scala-AWS Syllabus

1) Hadoop Fundamentals:

- What is Bigdata?
- Google Papers & Introduction to Hadoop by Doug Cutting
- Evolution of Bigdata
- Types of Data and their Significance
- DFS
- What is Hadoop
- Hadoop Distributions
- Features/Limitations of Hadoop
- Hadoop Ecosystem
- RDBMS Vs Hadoop

2) Hadoop HDFS Architecture:

- Hadoop 1.x Architecture
  1) HDFS
     a) NameNode
     b) Secondary NameNode
     c) DataNode
  2) MapReduce
     a) Job Tracker
     b) Task Tracker
- Limitation of Hadoop 1.x
- Hadoop 2.x Architecture
  a) High Availability
  b) Federation
  c) Resource Manager
  d) Node Manager
- Hadoop 2.x Core Components
- Block Size in 1.x & 2.x
- HDFS Federation
- Replication Factor
- Rack Awareness

3a) HDFS Commands:

- Diff Between hdfs dfs & hadoop fs
- -ls
- -mkdir -p
- -put
- -copyFromLocal
- -cat
- -touchz
- -get
- -copyToLocal
- -appendToFile
- -rm
- -rm [-skipTrash]
- -count [-h|-q|-v]
- -du
- -df
- -moveFromLocal
- -cp | -mv
- -chmod
- -stat [%u|%g|%b|%r|%n|%y|%Y]
- -tail
- -head
- -help
- -help -usage
- -test [-d|-e|-f|-s|-w|-r|-z]
- -setrep [-w|-R]
- -find
- -getfacl
- -getfacl -R
- truncate
- checksum
- -text
- -getmerge

3b) HDFS Admin Commands:

- -report [-live] [-dead] [-decommissioning] [-enteringmaintenance] [-inmaintenance]
- -safemode [-enter] [leave] [get] [wait] [forceExit]

MySQL & Sqoop:

4) MySQL:

- How to Login
- Create User
- Grant Privileges
- Flush Privileges
- Users
- Logged-in Users
- DB Name
- show DB
- create DB
- use DB
- create table
- show tables
- describe table
- Insert query
- ctas query
- drop table
- drop DB
- table count query
- find all pk's on DB tables

5) Sqoop:

- What is Sqoop
- Why is Sqoop Used
- Features of Sqoop
- Sqoop Architecture

Sqoop Commands:

- version
- list-databases
- list-tables
- eval
  a) show databases
  b) show tables
  c) Insert into
- --boundary-query
- Import
  a) table data with PK
  b) table data without PK
  c) -m1
  d) --split-by
- Protecting Password
  a) Standard Input
  b) From file
- --direct
- --target-dir
- --delete-target-dir
- --append
- --warehouse-dir
- --columns
- -m1
- --num-mappers
- import-all-tables
- --exclude-tables
- --autoreset-to-one-mapper
- --compress
- --compression-codec
- --as-sequencefile
- --as-avrodatafile
- --fields-terminated-by
- incremental imports
  a) --append
  b) --lastmodified
- Sqoop Job
  a) --create
  b) --delete
  c) --exec
  d) --show
  e) --list
- Sqoop Export
- --null-string
- --null-non-string
- --map-column-java
- --update-mode updateonly
- --update-mode allowinsert
- --staging-table

Hive:

6) Hive

- What is Hive/Why Hive
- Hive Architecture
  - a) HDFS/Map Reduce
  - b) Metastore
  - c) Driver
  - d) Hive Clients
- **Hive Metastore:**
  - a) Embedded
  - b) Local
  - c) Remote
- **Data Types:**
  - a) Numeric
  - b) Date/Time
  - c) String
  - d) Miscellaneous
- **Complex Data Types:**
  - a) Array
  - b) Map
  - c) Struct
  - d) Union
- **Hive Queries:**
  - a) show DB
  - b) create DB
  - c) describe
  - d) use DB
  - e) current DB
  - f) drop DB
  - g) create table
  - h) view table
  - i) alter table
  - j) truncate table
  - k) describe table
  - l) load
    - a) LFS to Hive
    - b) HFS to Hive
  - m) insert
  - n) multi-insert
- Managed Table
- External Table
- Diff between Managed/External
- **Functions:**
  - a) unix_timestamp
  - b) from unix_time
  - c) year/quarter/month/day/hour/min
  - d) to_date
  - e) weekofyear
  - f) datediff
  - g) date_add
  - h) date_sub
  - i) current_date
  - j) last_day
  - k) ceil
  - l) floor
  - m) round
  - n) concat
  - o) length
  - p) lower/upper
  - q) lpad/rpad
  - r) trim/ltrim/rtrim
  - s) reverse/split
  - t) substr/instr
  - u) nvl/coalesce/if-else
  - v) rank/dense_rank/row_number
  - w) explode/lateral
- **Hive Partitions:**
  - a) Static
  - b) Dynamic

  desc/alter/add/drop partitions
- Bucketing
- **Hive Joins:**

  Inner/left outer/right outer/full outer
- Views/Index
- Map Join
- Bucket Map Join
- Sort-Merge-Bucket Map Join
- UDF
- ACID Transactions
- Variables in Hive
- File Compressions
- **Window Functions:**
- a) lead/lag
- b) first_value/last_value
- c) count/sum/min/max/avg
- d) rank/dense_rank/row_number
- e) percent_rank
- f) ntile
- Read Sequence File Data from Hive
- Read Avro File Data from Hive

Scala & Spark:

Scala:

- Scala Introduction
- Basic Syntax & First Program
- Variables
- String Interpolation
- Data Types
- OOPs Concept
- Functions
- Closures
- Strings
- Arrays
- **Collections:**
  - a) List
  - b) Set
  - c) Map
  - d) Tuple
  - e) Option
  - f) Iterators
- If-Else
- Loops
- Traits
- Access Modifiers
- Pattern Matching
- Regular Expressions
- Exception Handling
- Extractors
- File I/O

Spark:

- What is Spark
- Why Spark
- Spark Ecosystem:
  - a) Spark Core
  - b) Spark SQL
  - c) Spark Streaming
  - d) MLib
  - e) GraphX
  - f) SparkR
- What is RDD
- Different Ways to create RDD
  - a) Parallelized Collections
  - b) External Datasets
  - c) Existing RDD's
- Features of Apache Spark
- Limitations of Apache Spark
- SparkContext/SparkConf
- Spark Architecture
- Spark Shell Commands:
- a) Read File & Create RDD
- b) Create RDD via Parallelized Collection
- c) Create RDD from Existing RDD
- Count/Filter
- take/partitions/cache
- saveAsTextFile
- RDD Operations:
- a) Transformations
- b) Actions
- map/flatMap/filter/mapPartitions
- Intersection/distinct/groupByKey
- reduceByKey/sortByKey/join/coalesce
- repartition/count/collect/take/top
- countByValue/countByKey/reduce
- union/foreach
- Map Vs FlatMap
- In-Memory Computation:
- a) Memory Only
- b) Memory And Disk
- c) Memory Only Ser
- d) Memory And Disk Ser
- e) Disk Only
- f) Memory And Disk_2
- Spark SQL Introduction
- Spark SQL Architecture
- Dataframe/SqlContext/HiveContext
- Dataset/JDBC DS/Catalyst Optimizer
- Different Ways of Creating Dataframe:
- a) textfile
- b) spark session without schema
- c) spark session with schema
- d) sqlContext
- e) collections
- Spark SQL Queries:
- a) printSchema
- b) select
- c) concat
- d) filter
- e) groupBy
- f) like
- g) in
- h) orderby
- i) distinct
- j) join

Hbase-Cassandra-Kafka-AWS-Git-GitHub

Hbase:

- Introduction
- Shell Commands
- Create Table
- List Tables
- Describe Tables
- Insert Rows to Tables
- Exists and Count
- Read All Rows
- Filtering Rows
- Get Single Row
- Disable or Enable Table
- Delete Rows

Cassandra:

- What is Cassandra
- Features of Cassandra
- History of Cassandra
- Components of Cassandra
- Data Replication
- Read/Write Operation
- CQL
- Diff between NoSql & RDBMS
- CQL Data Types
- Cassandra Vs RDBMS
- Cassandra TTL
- DDL/DML/Clauses
- Create Keyspace
  a) Diff components of Cassandra
  b) Replication
  c) Durable Writes
  d) Verification
- Alter Keyspace
- Drop Keyspace
- Create Table
- Alter Table
- Adding/Dropping Column

Kafka:

- Introduction
- Terminology
- Architecture
- Replication
- Producer
- Consumer
- Broker

AWS:

- EC2
- S3
- Glacier
- EBS
- RDS
- Redshift
- Athena
- EMR
- Lambda
- AWS CLI

Git & GitHub:

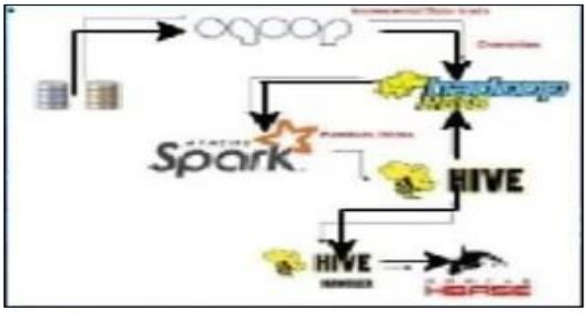
- What is Git & GitHub
- Repository
- Staging Files
- Commits
- Undoing Things
- Branches
- Merging Branches
- Git Commands

Hadoop-Spark Integrations

- Sqoop-Hive
- Hive-Sqoop
- Spark-HBase
- Spark-HBase-Hive
- Spark-MySQL-MySQL
- Spark-Hive-Hive
- Spark-Kafka

Project:

Project 1



Project 2