TRENDY TECH
UPLIFT YOUR CAREER!

www.trendytech.in

Big Data

# Engineering Masters Program

## TRENDYTECH

5 STAR GOOGLE RATED
BIG DATA COURSE

# CURRICULUM
# BIG DATA MASTERS PROGRAM

## 01 WEEK

### INTRODUCTION TO BIG DATA & HDFS CONCEPTS ALONG WITH LINUX COMMANDS

»» INTRODUCTION TO BIG DATA
»» WHAT IS BIG DATA AND WHY BIG DATA
»» BIG DATA SYSTEM REQUIREMENTS
»» MONOLITHIC VS DISTRIBUTED SYSTEM
»» DISTRIBUTED SYSTEM ARCHITECTURE
»» WHAT IS HADOOP AND EVOLUTION OF HADOOP
»» GOOGLE FILE SYSTEM (GFS)
»» DISTRIBUTED PROCESSING (MAPREDUCE)
»» HADOOP 1.0 VS HADOOP 2.0
»» WHAT IS YARN
»» CORE COMPONENTS OF HADOOP
»» HADOOP ECOSYSTEMS TOOLS
»» BRIEF INTRODUCTION TO SPARK
»» HADOOP CLUSTER VS SPARK CLUSTER
»» HDFS ARCHITECTURE:
»» WHAT IS NODE AND WHAT IS CLUSTER
»» DATA BLOCK & BLOCK SIZE
»» SLAVE NODE, MASTER NODE, DATA NODE & NAME NODE
»» METADATA AND REPLICATION FACTOR
»» HEART BEAT & FAULT TOLERANCE
»» HANDLING NAMENODE FAILURE

»» WHAT IS SPOF
»» FSIMAGE & EDIT LOGS
»» SECONDARY NAMENODE
»» NAME NODE RECOVERY
»» CHECK POINTING
»» UNDERSTANDING REPLICATION FACTOR
»» WHAT IS RACK AND RACK FAILURE
»» RACK AWARENESS MECHANISM
»» BLOCK REPORT
»» NAMENODE HIGH AVAILABILITY
»» QUORUM JOURNAL MANAGER & QUORUM JOURNAL NODE
»» UNDERSTANDING LINUX FILE SYSTEM
»» LIST & PARAMETERS OF LIST COMMAND
»» TOUCH, MKDIR, RMDIR & OTHER LINUX COMMANDS
»» HDFS COMMANDS:
»» LIST FILES & DIRECTORIES
»» HOW HDFS COMMANDS WORK
»» 'LS' COMMAND WITH VARIOUS PARAMETERS
»» CREATE, REMOVE FILE/DIRECTORY
»» COPY & GET FILES/FOLDERS FROM LOCAL TO HDFS & VICE VERSA
»» MOVE FILES/FOLDERS FROM HDFS TO HDFS
»» CHANGE REPLICATION FACTOR DYNAMICALLY
»» VIEW FILE METADATA INFORMATION
»» WEEK1: QUIZ
»» WEEK1: ASSIGNMENT

## 02 WEEK

### MAPREDUCE - DISTRIBUTED COMPUTING FRAMEWORK

»» INTRODUCTION TO MAPREDUCE
»» WHAT IS MAPREDUCE
»» STAGES IN MAPREDUCE
»» WHAT IS KEY-VALUE
»» WHAT IS MAP & WHAT IS REDUCE
»» EXAMPLE TO UNDESTAND MAP&REDUCE
»» WORD COUNT EXAMPLE IN MAPREDUCE
»» RECORD READER
»» MAPPER PHASE
»» REDUCER PHASE
»» MAPREDUCE SHUFFLE & SORT
»» INSIDE MAP & REDUCE PHASE
»» WORDCOUNT EXAMPLE IN MAPREDUCE
»» TYPICAL MAPREDUCE FLOW
»» BLOCKS IN MAPREDUCE
»» DEFAULT NUMBER OF MAPPERS & REDUCERS
»» UNDERSTANDING NUMBER OF MAPPERS/REDUCERS
»» MAPREDUCE FRAMEWORK BEHIND THE SCENES
»» ROLE OF HASH FUNCTION IN MAPREDUCE
»» PARTITIONING IN MAPREDUCE
»» HOW TO CHOOSE NUMBER OF REDUCERS
»» HOW HASH FUNCTION WORKS
»» UNDERSTANDING SHUFFLE & SORT
»» EXAMPLE: CALCULATING MAX TEMPERATURE IN A DAY
»» COMBINER FUNCTION IN MAPREDUCE
»» ADVANTAGES OF COMBINERS
»» WHEN TO USE OR NOT TO USE COMBINER
»» EXAMPLE1: FILTERING DATA USING MAPREDUCE
»» EXAMPLE2: FINDING DISTINCT VALUES
»» EXAMPLE3: FINDING TOP 3 MOST INFLUENTIAL USERS

»» REALTIME USE CASE: GOOGLE WEB SEARCH
»» HOW GOOGLE SEARCH WORKS
»» MAPREDUCE PROGRAMMING
»» MR CODE EXPLANATION
»» HOW TO WRITE MAP REDUCE CODE
»» MAPPER CODE
»» REDUCER CODE
»» MAIN CODE
»» FINDING THE FREQUENCY OF EACH WORD IN A FILE
»» MAPREDUCE JARS
»» MAPREDUCE PRACTICAL SESSIONS
»» WORD COUNT PROGRAM - PRACTICAL SESSION1
»» JAR CREATION & EXECUTION - PRACTICAL SESSION2:
»» HOW TO CREATE A JAR
»» HOW TO EXECUTE THE JAR
»» HOW TO TRACK A JOB
»» HOW TO TRACK ALL PREVIOUS JOBS
»» MR PROGRAM VARIATIONS - PRACTICAL SESSION3:
»» HOW TO CHANGE NUMBER OF REDUCERS
»» WRITING CUSTOM PARTITIONER LOGIC
»» CHANGING NUMBER OF REDUCERS TO ZERO
»» INTRODUCING COMBINER
»» WRITING CUSTOM COMBINER LOGIC
»» WEEK2: QUIZ
»» WEEK2: ASSIGNMENT
»» WEEK1 ASSIGNMENT SOLUTION

## 03 Week

### Apache Sqoop - Data Ingestion to Hadoop

- »» Sqoop Fundamentals
- »» Sqoop Basics
- »» What is sqoop
- »» Sqoop Workflow
- »» Key Features of Sqoop
- »» Sqoop Import
- »» Sqoop Export
- »» Connecting to MySQL
- »» Acessing MySQL Databases from Hadoop
- »» Acessing MySQL Tables from Hadoop
- »» Sqoop Eval
- »» Sqoop Import Practicals
- »» Sqoop Export Practicals
- »» Sqoop Job
- »» Sqoop Password Management
- »» Sqoop Incremental Load
- »» Sqoop Default Import
- »» Sqoop Free-Form Query Import
- »» Sqoop Direct import
- »» Importing Data Into Hive
- »» Importing Data Into HBase
- »» Sqoop Validate
- »» When a Sqoop Export May Fail
- »» Week3: Quiz
- »» Week3: Assignment
- »» Week2 Assignment Solution

Apache Sqoop

## Apache Hive Basics

## 04 Week

### Apache Hive Basics - Process Structure Data in Hadoop

»» Hive Overview:
»» Transactional System and Analytical System
»» Examples of Transactional Systems
»» Examples of Analytical Systems
»» What is Hive
»» Hive Query Language (HQL)
»» Understanding Hive Table
»» Introduction to Hive Metadata
»» Why Hive over traditional databases
»» Transactional and Analytical Processing
»» What is Data Warehouse
»» Hive Architecture
»» Hive on top of Hadoop
»» How Hive Works
»» Transactional vs Analytical Processing
»» Data Warehouse Concept
»» The Hive Metastore
»» Hive vs RDBMS
»» HQL vs SQL
»» Hive Subqueries Views & Index
»» Transactional and Analytical Processing
»» What is Data Warehouse
»» Hive Architecture
»» Hive on Hadoop
»» Hive Metastore
»» Hive vs. RDBMS
»» Hive Complex Data Types
»» Hive Array, Map & Struct
»» Hive Built-in Functions
»» Hive UDF, UDAF & UDTF
»» Hive Lateral Views
»» Hive Subqueries
»» Hive Views
»» Hive Normalization vs Denormalization
»» Week4: Quiz
»» Week3 Assignment Solution

# 05 WEEK

## APACHE HIVE ADVANCE - PART 1

»» HIVE STRUCTURE LEVEL OPTIMIZATIONS:
»» HIVE PARTITIONING
»» HIVE PARTITIONING WITH 2 COLUMNS
»» HIVE BUCKETING
»» HIVE PARTITIONING WITH BUCKETING
»» HIVE QUERY LEVEL OPTIMIZATIONS:
»» HIVE JOIN OPTIMIZATIONS
»» HIVE BUCKET MAP JOIN OPTIMIZATIONS
»» HIVE WINDOW FUNCTIONS
»» HIVE RANKING
»» HIVE SORTING
»» WEEK5: QUIZ
»» WEEK5: ASSIGNMENT

# 06 WEEK

## APACHE HIVE ADVANCE - PART 2

- HIVE FILE FORMAT
- ROW VS COLUMN FILE FORMATS
- SPECIALIZED FILE FORMATS
- INTERNALS OF ORC FILE FORMATS
- INTERNALS OF PARQUET FILE FORMATS
- ORC VS PARQUET FILE FORMATS
- HIVE COMPRESSION TECHNIQUES
- HIVE VECTORIZATION
- CHANGING THE HIVE ENGINE
- HIVE THRIFT SERVER
- HIVE MSCK REPAIR
- HIVE SCD
- WEEK6: QUIZ
- WEEK6: ASSIGNMENT
- WEEK5: ASSIGNMENT SOLUTION

## 07 WEEK

### NOSQL DATABASES - HBASE

»» HBASE BASICS
»» KEY REQUIREMENTS OF DATABASE
»» LIMITATIONS OF HADOOP
»» GOOGLE BIGTABLE CONCEPT FOR QUICK SEARCHING
»» IMPLEMENTATION OF BIGTABLE AS HBASE
»» PROPERTIES OF HBASE
»» WHAT HBASE CAN OFFER
»» ROW BASED STORAGE VS COLUMNAR STORAGE
»» ADVANTAGES OF COLUMNAR STORAGE
»» NORMALIZATION VS DENORMALIZATION
»» CRUD OPERATION
»» RDBMS VS HBASE
»» HBASE DATA MODEL
»» 4-DIMENSIONAL DATA MODEL
»» CAP THEOREM
»» HBASE ARCHITECTURE
»» HBASE REGION SERVER
»» REGION, MEMSTORE, WAL & BLOCK CACHE
»» HFILE
»» ZOOKEEPER
»» HMASTER & META TABLE
»» HBASE ARCHITECTURE COMPONENTS IN DETAILS
»» HBASE READ/WRITE OPERATIONS
»» COMPACTION
»» HBASE DATA UPDATE
»» HBASE DATA DELETION
»» HANDLING SERVER FAILURES
»» HBASE PRACTICALS
»» HANDLING HBASE FAILURE SERVICES

»» CREATE & LIST TABLE
»» INSERT RECORDS IN TABLE
»» SCAN(VIEW) & GET RECORDS FROM TABLE
»» DELETE A COLUMN
»» DESCRIBE A TABLE
»» CHECK TABLE EXISTS OR NOT
»» DROP TABLE - UNDERSTANDING HOW IT WORKS
»» PARAMETERS OF GET COMMAND
»» PARAMETERS OF SCAN COMMAND
»» HBASE FILES STRUCTURE IN HDFS
»» HOW TO DISABLE/ENABLE A TABLE
»» VARIOUS FILTERS IN HBASE
»» COUNT RECORDS
»» CASSANDRA OVERVIEW
»» WHAT IS CASSANDRA
»» HOW CASSANDRA CLUSTER LOOK LIKE
»» TUNABLE READ/WRITE CONSISTENCY
»» HBASE VS CASSANDRA
»» INTEGRATION WITH HADOOP (MINI PROJECT)
»» HBASE-HIVE INTEGRATION
»» WEEK7: QUIZ
»» WEEK7: ASSIGNMENT
»» WEEK6 ASSIGNMENT SOLUTION

## 08 WEEK

### LEARNING SCALA - A GUIDE TO FUNCTIONAL PROGRAMMING

»» WHY SCALA
»» WHERE TO RUN SCALA CODE
»» SCALA CODE USING IDE
»» SCALA BASICS»» VAR VS VAL
»» TYPE INFERENCE
»» DATA TYPES IN SCALA
»» STRING INTERPOLATION
»» STRING COMPARISON
»» FLOW CONTROL: IF ELSE
»» MATCH CASE
»» FOR LOOP
»» WHILE LOOP
»» SCALA FUNCTIONAL PROGRAMMING
»» HOW TO DEFINE A FUNCTION
»» HIGHER ORDER FUNCTION
»» ANONYMOUS FUNCTION
»» SCALA COLLECTIONS
»» ARRAY
»» LIST
»» TUPLE
»» RANGE
»» SET
»» MAP
»» SCALA FUNCTIONAL PROGRAMMING:
»» WHY SCALA
»» MODES OF WRITING SCALA CODE
»» WHAT IS A FUNCTIONAL PROGRAMMING

»» WHAT IS A FUNCTION
»» WHAT IS A PURE FUNCTION?
»» FIRST CLASS FUNCTION
»» HIGHER ORDER FUNCTION
»» ANONYMOUS FUNCTION
»» IMMUTABILITY
»» LOOP
»» RECURSION
»» TAIL RECURSION
»» STATEMENT VS EXPRESSION
»» CLOSURE
»» SCALA TYPE SYSTEM
»» SCALA OPERATORS
»» ANONYMOUS FUNCTION
»» PLACEHOLDER SYNTAX
»» PARTIALLY APPLIED FUNCTIONS
»» FUNCTION CURRYING
»» WEEK8: QUIZ
»» WEEK8: ASSIGNMENT
»» WEEK7 ASSIGNMENT SOLUTION

## 09 WEEK

### APACHE SPARK - GENERAL PURPOSE CLUSTER COMPUTING FRAMEWORK

»» SCALA INTERVIEW PREPARATION SERIES
»» WHAT IS APP CLASS IN SCALA
»» DEFAULT ARGS, NAMED ARGS & VARIABLE ARGS
»» DIFFERENCE BETWEEN NIL, NULL, NONE & NOTHING
»» WHAT IS OPTION IN SCALA
»» WHAT IS UNIT IN SCALA
»» DEALING WITH NULLS IN SCALA
»» WHAT IS YIELD
»» WHAT IS VECTOR
»» SCALA IF GUARDS & PATTERN GUARDS
»» WHAT IS "FOR COMPREHENSIONS"
»» DIFFERENCE BETWEEN "==" IN JAVA AND SCALA
»» DIFFERENCE BETWEEN STRICT VAL VS LAZY VAL
»» WHAT ARE DEFAULT PACKAGES IN SCALA
»» WHAT IS SCALA APPLY METHOD
»» WHAT IS A DIAMOND PROBLEM IN SCALA
»» WHAT IS A TRAIT
»» WHY SCALA IS THE TOP MOST CHOICE FOR A BIG DATA DEVELOPER OVER PYTHON AND JAVA
»» WHAT IS APACHE SPARK
»» UNDERSTANDING SPARK CLUSTER
»» IS SPARK A REPLACEMENT TO HADOOP
»» WHY SPARK IS FASTER THAN MAPREDUCE
»» HOW DATA STORE IN SPARK
»» WHAT IS RDD
»» WHAT IS DAG

»» RDD LINEAGE
»» RESILIENCY
»» IMMUTABILITY
»» TRANSFORMATION & ACTION
»» LAZY EVALUATION
»» WORD COUNT PROGRAM IN SPARK
»» WORD COUNT PROGRAM IN PYSPARK
»» WORD COUNT PROBLEM REAL-TIME EXAMPLE
»» WEEK9: QUIZ
»» WEEK9: ASSIGNMENT
»» WEEK8 ASSIGNMENT SOLUTION

## 10 WEEK

### APACHE SPARK - IN DEPTH

»» SPARK REAL-TIME EXAMPLE
»» BROADCAST VARIABLE
»» ACCUMULATORS
»» HOW SPARK EXECUTES PROGRAM ON THE CLUSTER
»» SPARK DRIVER AND EXECUTORS
»» CLIENT MODE, CLUSTER MODE AND LOCAL MODE
»» ANALYZING LOG MESSAGES - HANDS ON
»» NARROW VS WIDE TRANSFORMATIONS
»» STAGES IN SPARK
»» DIFFERENCE BETWEEN REDUCEBYKEY & REDUCE
»» DIFFERENCE BETWEEN GROUPBYKEY & REDUCEBYKEY
»» PAIR RDD
»» PAIR RDD VS MAP
»» UNDERSTANDING DEFAULT PARALLELISM
»» DIFFERENCE BETWEEN REPARTITION & COALESCE
»» WHEN TO INCREASE/DECREASE PARTITIONS
»» SPARK ON YARN ARCHITECTURE

### YARN - YET ANOTHER RESOURCE NEGOTIATOR

»» LIMITATIONS OR DRAWBACKS OF MR1
»» RESOURCE MANAGER
»» NODE MANAGER
»» APPLICATION MASTER
»» CONTAINERS
»» WEEK10: QUIZ
»» WEEK10: ASSIGNMENT
»» WEEK9 ASSIGNMENT SOLUTION

## 11 WEEK

### APACHE SPARK - STRUCTURED API PART-1

»» CACHE VS PERSIST
»» SPARK STORAGE LEVELS
»» DIFFERENCE BETWEEN DAG & LINEAGE
»» HOW TO SUBMIT A SPARK JOB
»» REAL-TIME EXAMPLE - FINDING TOP MOVIES BASED ON RATINGS

## 12 WEEK

### APACHE SPARK - STRUCTURED API PART-2

»» WRITING OUTPUT TO SINK (SPARK.WRITE)
»» SPARK FILE LAYOUT
»» BENEFITS OF REPARTITIONS
»» PARTITIONBY & BUCKETBY
»» SAVING FILE IN VARIOUS FILE FORMAT
»» INTRODUCTION TO SPARKSQL
»» STORING DATA IN PERSISTENT MANNER
»» HANDLING SPARK METADATA
»» LOW & HIGH LEVEL TRANSFORMATIONS
»» REFERING TO A COLUMN IN DATAFRAME/DATASET
»» COLUMN STRING
»» COLUMN OBJECT
»» COLUMN EXPRESSION
»» SPARK UDF USING STRUCTURED API
»» ADDING COLUMN IN DATAFRAME
»» DATAFRAME TO DATASET USING CASE CLASS.
»» DATASET TO DATAFRAME CONVERSION
»» SPARK CATALOG
»» REGISTRING UDF WITH DRIVER
»» TRANSFORMATIONS HANDS ON EXAMPLES
»» AGGREGATE TRANSFORMATIONS

»» SIMPLE AGGREGATIONS
»» GROUPING AGGREGATIONS
»» WINDOW AGGREGATIONS
»» JOINS ON DATAFRAME
»» SIMPLE JOIN (SHUFFLE SORT MERGE JOIN)
»» BROADCAST JOIN
»» DEALING WITH AMBIGUOES COLUMN NAMES
»» DEALING WITH NULL'S
»» INTERNALS OF JOIN OPERATIONS
»» WHEN TO USE SIMPLE JOIN WHEN USE BROADCAST JOIN
»» GROUPING AGGREGATION REAL-TIME EXAMPLE
»» INFERING DATA IN SPARKSQL
»» WEEK12: QUIZ
»» WEEK12: ASSIGNMENT
»» WEEK11 ASSIGNMENT SOLUTION

## 13 WEEK

### APACHE SPARK - OPTIMIZATION PART-1

»» LEVEL OF OPTIMIZATIONS
»» RESOURCE LEVEL OPTIMIZATIONS
»» APPLICATION LEVEL OPTIMIZATIONS
»» CLUSTER LEVEL OPTIMIZATIONS
»» HOW TO CALCULATE NO OF EXECUTORS
»» THIN EXECUTOR
»» FAT EXECUTOR
»» HOW TO CALCULATE NO OF EXECUTORS
»» HOW TO CALCULATE MEMORY ALLACATION
»» HOW TO CALCULATE NO OF CORES
»» HEAP MEMORY
»» OFF-HEAP MEMORY
»» HANDS ON WITH REAL-TIME CLUSTER
»» UNDERSTANDING CLUSTER CONFIGUARATIONS
»» REALTIME EXAMPLE:

### MOVING ATA TO HDFS USING A EDGE NODE AND WORK AROUND IT IN A REALTIME CLUSTER

»» STATIC RESOURCE ALLOCATION
»» DYNAMIC RESOURCE ALLOCATION
»» UNDERSTANDING MEMORY USAGE IN SPARK
»» EXECUTION MEMORY
»» STORAGE MEMORY
»» PRACTICAL DEMONSTRATION:

### CACHE & PERSIST

»» JAVA SERIALIZER VS KRYO SERIALIZER
»» WEEK12: QUIZ
»» WEEK12: ASSIGNMENT
»» WEEK11 ASSIGNMENT SOLUTION

## 14 WEEK

### APACHE SPARK - OPTIMIZATION PART-2

»» BROADCAST JOIN PRACTICAL DEMONSTARTIONS
»» BROADCAST JOIN USING RDD
»» WHEN TO USE BROADCAST JOIN
»» BROADCAST JOIN USING DATAFRAME
»» VISUALIZING BROADCAST JOIN WITH STRUCTURED API
»» PRACTICAL DEMO ON REPARTITION VS COALESCE
»» CLIENT MODE VS CLUSTER MODE WHEN USING SPARK SUBMIT
»» SPARK JOIN OPTIMIZATIONS
»» SPARK ADVANCE OPTIMIZATIONS: SORT AGGREGATE VS HASH AGGREGATE
»» SPARK CATALYST OPTIMIZER
»» WEEK14: QUIZ
»» WEEK14: ASSIGNMENT
»» WEEK13 ASSIGNMENT SOLUTION

## 15 WEEK

### APACHE SPARK - STREAMING PART-1

- »» KIND OF PROCESSING
- »» WHAT IS REAL-TIM PROCESSING
- »» THE IMPORTANCE OF REAL-TIME PROCESSING
- »» BATCH PROCESSING VS REAL-TIM STREAM PROCESSING
- »» SPARK STREAMING DATA
- »» SPARK DISCRETIZED STREAM OR DSTREAM
- »» BATCH & BATCH INTERVAL
- »» DO SPARK IS A REAL-TIME STREAMING ENGINE
- »» STREAM PROCESSING IN SPARK
- »» TRANSFORMED DSTREAM
- »» UNDERSTANDING PRODUCER & CONSUMER
- »» PRACTICAL ON REAL-TIME PROCESSING
- »» STREAM TRANSFORMATIONS
- »» STATELESS TRANSFORMATIONS
- »» STATEFUL TRANSFORMATIONS
- »» WINDOW OPERATIONS
- »» BATCH INTERVAL
- »» WINDOW SIZE
- »» SLIDING INTERVAL
- »» PRACTICAL ON STATELESS TRANSFORMATION
- »» PRACTICAL ON STATEFUL TRANSFORMATION
- »» REDUCEBYKEY VS UPDATESTATEBYKEY
- »» WORKING WITH SLIDING WINDOW
- »» REDUCEBYKEYANDWINDOW TRANSFORMATION
- »» REDUCEBYWINDOW TRANSFORMATION
- »» COUNTBYWINDOW TRANSFORMATION
- »» WEEK15: QUIZ
- »» WEEK15: ASSIGNMENT
- »» WEEK14 ASSIGNMENT SOLUTION

## 16 WEEK

### APACHE SPARK - STREAMING PART-2

»» WHAT IS STRUCTURED STREAMING
»» REQUIREMENT OF STRUCTURE STREAMING
»» LIMITATIONS OF SPARK STREAMING
»» BENEFITS OF SPARK STRUCTURE STREAMING
»» PRACTICAL - WORDCOUNT EXAMPLE ON STRUCTURED STREAMING
»» DYNAMICALLY SETTING THE SHUFFLE PARTITIONS
»» DATA STREAM WRITER OUTPUT MODES
»» DATASTREAM OUTPUT MODES - APPEND, UPDATE & COMPLETE
»» SPARK STREAMING GRACEFUL SHUTDOWN
»» HOW DOES SPARK STREAMING CODE EXECUTES INTERNALLY
»» HOW A JOB CONVERTED TO MICRO BATCHES
»» TRIGGER POINT FOR MICRO BATCHES
»» TYPES OF TRIGGERS - UNSPECIFIED, TIME INTERVAL, ONE TIME, CONTINUOUS
»» TYPES OF DATA SOURCES - SOCKET SOURCE, RATE SOURCE, FILE SOURCE, KAFKA SOURCE
»» LIMITATIONS OF SOCKET SOURCE
»» PRACTICAL ON FILE DATA SOURCE
»» TYPES OF SPARK STREAMING OUTPUT DATA OPTIONS
»» FAULT TOLERANCE AND EXACTLY ONCE GUARANTEE
»» UNDERSTANDING CHECKPOINT LOCATION
»» STATEFUL VS STATELESS TRANSFORMATIONS
»» MANAGED STATEFUL OPERATIONS VS UNMANAGED STATEFUL OPERATIONS
»» TYPES OF AGGREGATIONS - CONTINUOUS AGGREGATIONS VS TIME BOUND AGGREGATIONS

»» WINDOW TRANFORMATIONS
»» UPDATESTATEBYKEY, REDUCEBYKEYANDWINDOW, REDUCEBYWINDOW, COUNTBYWINDOW
»» TYPES OF WINDOWS - TUMBLING TIME WINDOW, SLIDING TIME WINDOW
»» DEALING WITH LATE COMING RECORDS USING WATERMARK
»» STATE STORE CLEANUP
»» CALCULATING THE WATERMARK BOUNDARY
»» STREAMING JOINS
»» STREAMING DATAFRAME TO STATIC DATAFRAME
»» STREAMING DATAFRAME WITH ANOTHER STREAMING DATAFRAMES
»» WEEK16: QUIZ
»» WEEK16: ASSIGNMENT
»» WEEK15 ASSIGNMENT SOLUTION

## 17 WEEK

### APACHE KAFKA - DISTRIBUTED EVENT STREAMING PLATFORM

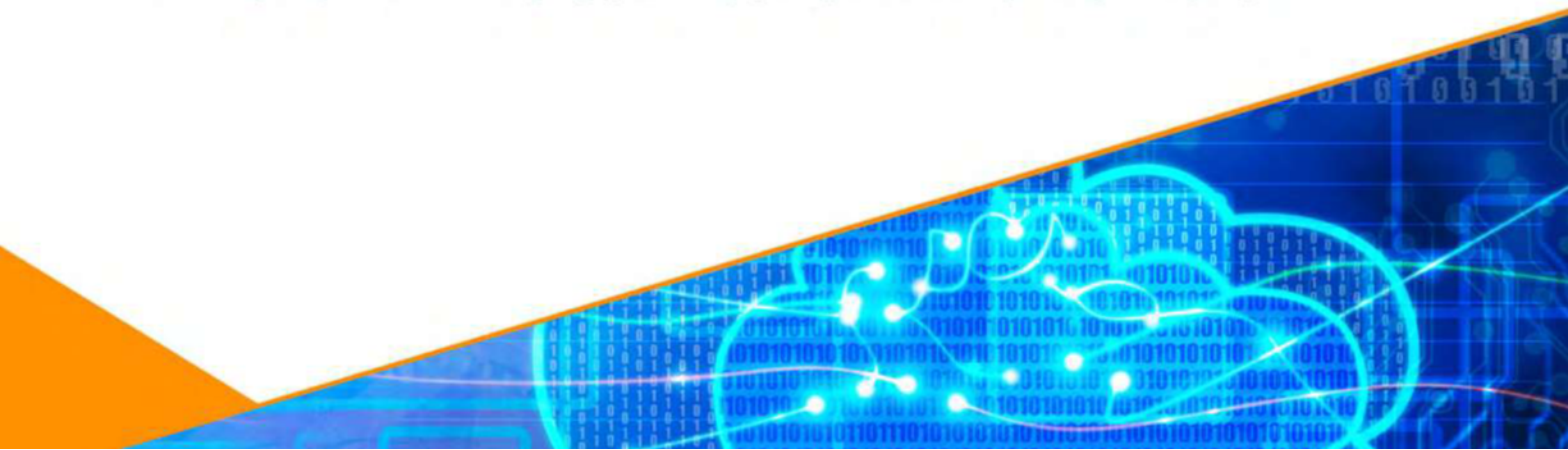
»» INTRODUCTION TO KAFKA
»» KAKFA ARCHITECTURE
»» KAFKA KEY CONCEPTS/FUNDAMENTALS
»» OVERVIEW OF ZOOKEEPER AND IT'S ROLE IN KAFKA CLUSTER
»» CLUSTER, NODES, BROKERS, TOPICS
»» CONSUMER, PRODUCERS, LOGS, PARTITIONS
»» CONCEPT OF CONSUMER GROUPS
»» LEADER & FOLLOWER PARTITION
»» INSTALLING ONE NODE KAFKA CLUSTER ON LOCAL
»» INSTALLING MULTINODE KAFKA CLUSTER ON LOCAL
»» COMMAND LINE PRODUCER AND CONSUMER
»» REPLICATION CONCEPT FOR FAULT TOLERANCE
»» HOW DATA IS STORED IN BROKERS
»» LOG SEGMENTS, MESSAGE OFFSETS, MESSAGE INDEX
»» ISR LIST / MINIMUM ISR
»» COMMITTED VS UNCOMMITED MESSAGES
»» WRITING A KAFKA PRODUCER IN JAVA
»» WRITING A KAFKA CONSUMER IN JAVA
»» SCALING UP THE KAFKA CLUSTER
»» ACHIEVING EXACTLY ONCE SEMANTICS
»» INTEGRATING KAFKA WITH SPARK STRUCTURED STREAMING.
»» WEEK16: QUIZ
»» WEEK16: ASSIGNMENT
»» WEEK15 ASSIGNMENT SOLUTION

## 18 WEEK

### BIG DATA ON CLOUD PART-1

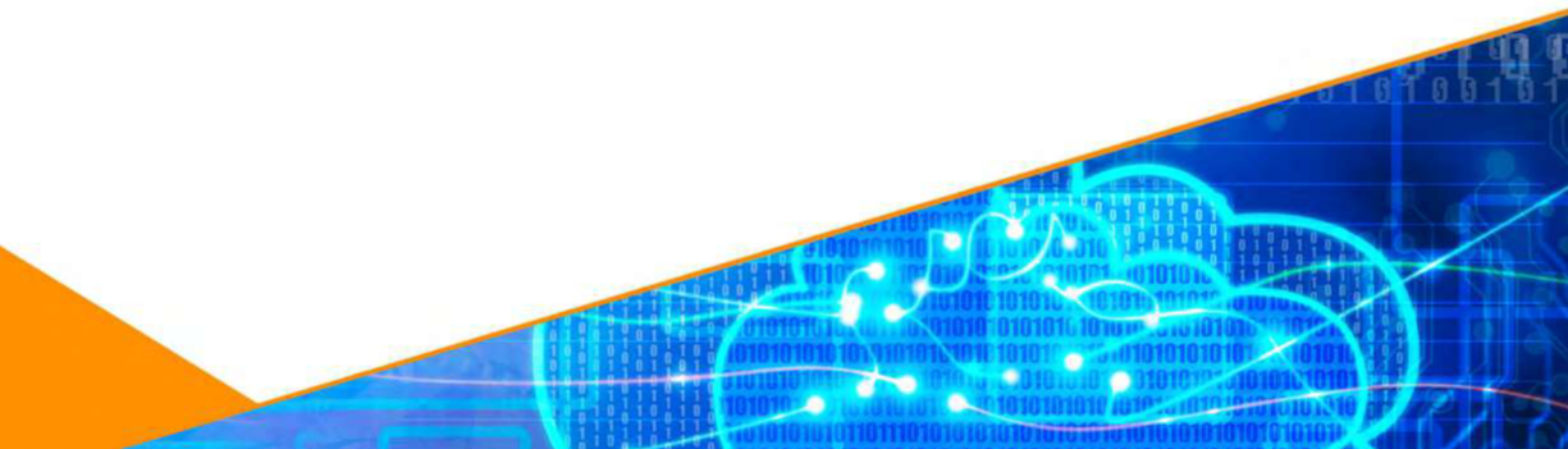
### AWS EMR (ELASTIC MAPREDUCE):

- »» WHAT IS A VM (VIRTUAL MACHINE)
- »» ON-PREMISE VS CLOUD SETUP
- »» MAJOR VENDORS OF HADOOP DISTRIBUTION
- »» WHY CLOUD & BIG DATA ON CLOUD
- »» MAJOR CLOUD PROVIDERS OF BIGDATA
- »» WHAT IS EMR
- »» HDFS VS S3
- »» WHAT IS S3
- »» IMPORTANT INSTANCES IN AWS
- »» KINDS OF NODES IN CLUSTER
- »» TRANSIENT VS LONG RUNNING CLUSTER
- »» RUNNING SPARK CODE ON EMR
- »» HOW TO TRACK YOUR JOB
- »» COPY FILE FROM S3 TO LOCAL
- »» ZEPPELIN NOTEBOOK
- »» TYPES OF EC2 INSTANCES
- »» HOW TO CREATE A VM
- »» WHAT IS A KEYPAIR
- »» ELASTIC IP
- »» AWS STORAGE, NETWORKING & CLI
- »» INSTANCE STORE
- »» S3 & EBS
- »» PUBLIC IP VS PRIVATE IP
- »» NETWORK SWITCHES
- »» SECURITY GROUP
- »» AWS COMMAND LINE INTERFACE
- »» LAUNCH A EMR CLUSTER USING ADVANCED OPTIONS



## AWS ATHENA:

»» WHAT IS ATHENA
»» WHEN DO WE REQUIRE ATHENA
»» WHAT PROBLEM ATHENA SOLVE
»» HOW ATHENA WORKS
»» ATHENA PRICING
»» ATHENA PRACTICAL DEMONSTRATION:
»» HOW TO CREATE A NORMAL TABLE MANUALLY ON CSV DATA RESIDING IN S3
»» HOW TO MINIMIZE DATA SCANNING IN ATHENA
»» HOW TO CREATE PARTITION TABLE ON PARQUET FILE
»» INFERING SCHEMA AUTOMATICALLY USING AWS GLUE
»» GLUE CATALOG
»» WEEK18: QUIZ
»» WEEK18: ASSIGNMENT
»» WEEK17 ASSIGNMENT SOLUTION



## 19 WEEK

### BIG DATA ON CLOUD PART-2

### AWS GLUE

»» WHAT IS AWS GLUE?
»» INTRODUCTION TO GLUE
»» FEATURES OF GLUE
»» AWS GLUE BENEFITS
»» AWS GLUE TERMINOLOGY
»» POINTING TO SPECIFIC DATA STORES AND ENDPOINTS
»» GLUE DATA CATALOGUE
»» CRAWLERS
»» CONNECTING TO YOUR DATA STORE
»» USING CRAWLERS FOR CATALOGUE TABLES
»» OVERVIEW AND WORKING OF GLUE JOBS
»» ADDING NEW JOBS IN GLUE
»» TRIGGERING JOBS AND THEIR SCHEDULING

### AWS REDSHIFT

»» DATABASE VS DATA WAREHOUSE VS DATA LAKE
»» INTRODUCTION TO AMAZON REDSHIFT
»» BENEFITS OF AMAZON REDSHIFT
»» USE CASES OF AMAZON REDSHIFT
»» REDSHIFT MASTER SLAVE ARCHITECTURE
»» TYPES OF NODES
»» REDSHIFT SPECTRUM
»» REDSHIFT FAULT TOLERANCE
»» REDSHIFT SORT KEYS
»» REDSHIFT DISTRIBUTION STYLES
»» PRACTICAL DEMONSTRATION
»» WEEK19: QUIZ
»» WEEK19: ASSIGNMENT
»» WEEK18 ASSIGNMENT SOLUTION

## 20 WEEK

## APACHE AIRFLOW - WORKFLOW MANAGEMENT PLATFORM

»» INTRODUCTION TO AIRFLOW AND ITS USAGE
»» WHAT IS WORKFLOW
»» CRON-JOB CREATION EXAMPLE
»» AIRFLOW ADDITIONAL FEATURES
»» AIRFLOW ARCHITECTURE AND COMPONENTS
»» AIRFLOW INSTALLATION DEMO
»» DAGS-CREATING A SIMPLE HELLOWORLD DAG
»» INTRODUCTION TO TASKS AND OPERATORS
»» VIEWING THE DAG IN UI-GRAPH VIEW, TREE VIEW, LOGS VIEWING
»» EXAMPLE SHOWCASING BASH OPERATORS USAGE
»» SETTING PRECEDENCE AMONG VARIOUS TASKS
»» LIFECYCLE OF A TASK-UNDERSTANDING VARIOUS STAGES
»» ABOUT TRIGGER_RULES & UNDERSTANDING WITH EXAMPLE
»» AIRFLOW ARTIFACT - MORE ON OPERATORS
»» WRITING OUR OWN CUSTOM OPERATORS
»» WALKTHROUGH OF AIRFLOW UI
»» CONNECTIONS TO VARIOUS DATASTORES & VARIABLES
»» WORKING WITH CONNECTIONS, UNDERSTANDING SENSORS – DEMO
»» BUILDING AN END-TO-END CUSTOMER-360 PIPELINE USING AIRFLOW INVOLVING DATA COLLECTION FROM VARIOUS SOURCES, PROCESSING IN SPARK, LOADING THE PROCESSED DATA IN HIVE AND UPLOADING THE SAME TO HBASE AND GENERATING A NOTIFICATION ABOUT SUCCESS OF THE PIPELINE TO THE DOWNSTREAM APPLICATIONS.

**PLUS**

**ONE END-TO-END PIPELINE PROJECT INVOLVING ALL MAJOR COMPONENTS LIKE SQOOP, HDFS, HIVE, HBASE, SPARK... ETC.**

**INTERVIEW PREPARATION TIPS:**

**SAMPLE RESUME**

**15+ MOCK INTERVIEW RECORDINGS**

**MOCK INTERVIEW QA**

**INTERVIEW QUESTIONS**

**HOW TO HANDLE MANAGERIAL ROUND QS**

Google Reviews ★★★★★

# 5 STAR GOOGLE RATED BIG DATA COURSE

## LEARN FROM THE EXPERT