Big Data Architect Masters Program

Big Data Masters Program makes you proficient in tools and systems used by Big Data experts. It includes training on Hadoop and Spark stack, Cassandra, Talend and Apache Kafka messaging system. The curriculum has been determined by extensive research on 5000+ job descriptions across the globe.

Career Related Program:

Extensive Program with 9 Courses

200+ Hours of Interactive Learning

Capstone Project

Key Learning

- All About Bigdata & Hadoop Drive
- Linux, SQL, ETL, & Datawarehouse Refresh
- Hadoop HDFS, Map Reduce, YARN Distributed Framework
- NOSQL - For realtime data storage and search using HBASE & ELASTIC SEARCH
- Visualization & Desktop - Jibana with Elastic search Integration using Spark
- Robotic Process Automation (RPA) Using Linux & Spark
- In Memory stream for Fast Data, Realtime Streaming & Data Formation using Spark, Kafka, Nifi.
- Reusable Framework creation with logging Framework
- Cluster formation creation in Cloud environments
- SDLC, Packaging & Deployment in Bigdata Platform
- Project execution with Hackathon & Test.
- Job submission & Orchestration with Scheduling using Oozie

High Level Eco System Overview

- All About Bigdata & Hadoop Deep Drive
- Linux, SQL, ETL, & Datawarehouse Refresh

- Hadoop HDFS, Map Reduce, YARN Distributed Framework
- SQOOP - Data ingestion Framework
- Hive - SQL & OLAP Layer on Hadoop
- HBASE & Elastic SEARCH - Real Time Random Read/Write NOSQL
- PHOENIX - SQL Layer on Top of HBASE
- KIBANA - Realtime Visualization on top Elastic Search
- OOZIE - Workflow Scheduling & Monitoring tool
- NIFI- Data Flow Tool for Mediation & Routing og large dataset
- KAFKA - Distributed & Scalable Messaging queue
- SPARK - Fast & Distributed In-Memory engine for largescale data
- SCALA/PYTHON - Scalable, Function based Highlevel Language
- HUE - GUI for Hadoop Eco System
- AMBARI - Provisioning, Managing and Monitoring Hadoop Cluster
- Google Cloud based - Hadoop & Spark Cluster setup
- HORTONWORKS - Distribution for provisioning Hadoop Cluster
- AWS Services - EMR, EC2, 53, IAM, SG, ATENA
- MAVEN & GITHUB - DevOps Continuous Build & Version control
- Frameworks for Data Masking, Data Validation & Sanitation

Overview of BIGDATA

We have to first have know all about Big-Data & its Characteristics.

- Evolution of Data
- Introduction
- Classification
- Size Hierarchy
- Why Big data is Trending
- 10T, Devops, Cloud Computing, Enterprise Mobility

- Challenges in Big Data
- Characteristics
- Tools for Big Data
- Why Big Data draws attention in IT Industry
- What do we do with Big data
- How Big Data can be analyzed
- Typical Distributed System
- Draw backs in Traditional
- Distrubited Systems
- Bigdata tools

Linux Foundation

In this module you will be learning Introduction & Key Components of Linux Dev & Admin

- History and Evolution
- Architecture
- Development Commands
- Env Variables
- File Management
- Directories Management
- Admin Commands
- Advance Commands
- Shell Scripting
- Groups and User managements
- Permissions
- Important directory structure
- Disk utilities
- Compression Techniques

- Misc Cornmands
- Kernel, Shell
- Terminal, SSH, GUI
- Hands On Exercises

Linux Scripting

In this module you will be lipux shell scripting and automation techniques

- Automation process using shell& scripting
- Integration of hadoop Eco systems with Linux scripting
- Looping, conditional, vars methods
- Key Differences between Linux & Windows
- Kernel
  - What is the Purpose of Kernel?
  - How Kernel Works?
  - Find Kernel
- Shell
  - What is the Purpose of Shell?
  - Types of Shell
  - Environment Variables in Shell
  - Hands On Exercises

Linux Scripting

In this module you will be lipux shell scripting and automation techniques

- Automation process using shell scripting
- Integration of hadoop Eco systems with Linux scripting
- Looping, conditional, vars methods

- Key Differences between Linux & Windows
- Kernel
  - What is the Purpose of Kernel?
  - How Kernel Works?
  - Find Kernel
- Shell
  - What is the Purpose of Shell?
  - Types of Shell
  - Environment Variables in Shell
  - Hands On Exercises

Hadoop In Depth

In this module you will be leaning all about Hadoop

- What is Hadoop?
- Evolution of Hadoop
- Features of Hadoop
- Characteristic of Hadoop
- Hadoop compared with Traditional Dist. Systems
- When to use Hadoop
- When not to use Hadoop Components of Hadoop IHDFS, MapReduce, YARN)
- Hadoop Architecture
- Daemons in Hadoop Version 1 & *2* How Data is stored in Hadoop Cluster, Datacenter, Spilt, Block. Rack Awareness, Replication, Heart beat)
- Hadoop 1.0 Limitation
- Name Node High Availability

Hadoop HDFS

Hadoop distributed file system concepts with architecture, commands, options, advance options, data management

- Name node Federation
- Hadoop version s
- Anatomy of File Read/Write
- Hadoop Ouster Formation in VM, Sandbox & GCP Cloud
- Cluster formation & sizing guide
- Hadoop Commands Hands-on
- Hadoop admin hands-on
- HDFS integration with Lima shell
- HDFS additional Use cases
- Data Integrity
- Serialization
- Compression techniques
- Data ingestion to HDFS using different ecosystems

SQOOP Data Acquisition

Data ingestion or data acquisition tool for transporting bulk data between RDBMS -> Hadoop & Vice versa

- Sqoop Introduction & History
- Technical & Business benefits
- Installation and configuration
- Why Sqoop
- In-depth Architecture
- Import & Export Properties
- Sqoop Export Architecture

- Commands (Import HOSE, HIVE, HBase from MYSCIL, ORACLE)
- Export Command Options
- Incremental Import
- Saved Jobs, Sqoop Merge
- Import All tables, Excludes
- Best practices & performance tuning
- Sqoop import/export use cases
- Advance Sqoop commands
- Sqoop Realtime use cases
- Sqoop Hive HBbase Integration

Hive

SQL Layer on top of Hadoop for analytical and declarative querie

- Introduction to Hive
- Architecture
- Hive Vs RDBMS Vs NOSQL
- Detailed Installation (Metastore, Integrating with Hue)
- Starting Metastore and Hive Server2
- Data types (Primitive, Collection Array, Struct, Map)
- Create Tables (Managed, External, Temp)
- DML operations (load, insert, export)
- Exploring Indexes
- HQL Automation using shell scripts
- Managed Vs External tables
- HOL Queries using end to end usecases
- Hive analytical and Hierarchial queries

Hive Components

Hive Components such as partition, bucketing, views, indexes, joins, handlers, udfs etc

- Hive access through Hive Client, Beeline and Hue
- File Formats (RC, ORC, Sequence)
- Partitioning (static and dynamic)
- partition with external table
- Drop, Repair Partitions
- Hive Sqoop, HBase, Integration
- Hive Storage Handler implementation
- Bucketing, Partitioning Vs Bucketing
- Views, different types of joins
- Aggregation, normalization Queries
- Add files to the distributed cache, jars to the class path
- UDF using Python & Scala
- Generic UDF, UDAF

Advance Hive

usecases & POCs on serdes, file formats, schema evolution, SCD concepts etc,

- Optimized joins (Mapside,join,SMB Bucketing join)
- Compressions on tables (LZO, Snappy)
- Serde (XML Serdq, JsonSerde, CSV, Avro, Regex)
- Parallel execution
- Sampling data
- Speculative execution
- Installation &Configuration
- Two POCs using the large dataset on the above topics

- Hive Slowly changing dimension implementation
- Hive Schema evolution use case using Avro dataset
- Hive Usecase with retail and banking dataset

Map Reduce Framework

Hadoop Processing framework for Distributed processing with multitasking capabilities

- Introduction to MapReduce
- Hadoop Ecosystems roadmap
- Map Reduce Flow
- Types of Input and Output Format
- MapReduce in details
- Different types of files supported (Text, Sequence, map and Awo)
- MapReduce lob submission in YARN Cluster in details
- Role of Mappers and reducers
- Identity Mapper, Identity Reducer
- Zero Reducer, Custom Partitioning
- Combiner, Sequence file format
- Tweaking mappers and reducers
- Mapreduce package and deployment
- Code component, walk through
- Mine, Sequence file format

Yarn

Hadoop Resource management component for containerization, scheduling with multi tenant feature

- Introduction to YARN
- YARN Architecture

- YARN Components
- YARN Longlived & Shortlived Daemons
- YARN Schedulers
- Job Submission under YARN
- Multi tenancy support of YARN
- YARN High Avalability
- YARN Fault tolerance handling
- MapReduce job submission using YARN
- YARN UI
- History Server
- YARN Dynamic allocation
- Containerisation of YARN

NOSQL - HBASE

Think beyond SQL with the column oriented datastore for realtime random read write of differential data sets

- Introduction to NoSQL
- Types of NOSOL
- Characteristics of NoSQL
- CAP Theorem
- Columnar Datastore
- What is HBase
- Brief History
- Row vs Column oriented
- HOES vs HBASE
- RDBMS vs HBASE
- Storage Hierarchy ->Characteristics

- Table Design
- HMaster & Regions

HBase Contd

Think beyond SQL with the column oriented datastore for realtime random read write of differential data sets

- Region Server & Zookeeper
- Inside Region Server (Memstore, Blockcache,HFile, WAL)
- HBase Architecture (Read Path, Write Path, Compactions, Splits)
- Minor/Major Compactions
- Region Splits
- Installiation &Configuration
- Role of Zookeeper
- HBase Shell
- Introduction to Filters
- Row Key Design
- Map reduce Integration
- Performance Tuning
- Hands on with Medical domain
- Hive HBase Handler
- SQoop HBase Integration

Phoenix

SQL Layer on top of HBASE for low latency, real time aggregation queries with joining capabilities

- Overview of Phoenix
  - Introduction

- Architecture
- History
- Phoenix Hbase Integration
- HBase table, view creation
- SQL & UDEs
- SQL Line & PLSQL Line of Phoenix
- Phoenix Load & Query engine
- Understanding coprocessor Configurations
- Hive -> Mask -> Phoenix integration
- Creation of views in phoenix
- Load bulk data using plsql
- Serverlog Aggregation usecase

Oozie

In this module, you will do the Hands on and Exploration of the Integration of components

- Introduction
- History - Why Oozie
- Components
- Architecture
- Workflow Engine
- Nodes
- Workflow
- Coordinator
- Action (MapReduce, Hive, Spark, Shell & Sqoop) Introduction to Bundle
- Email Notification
- Error Handling
- Installation

- Workouts
- Orchestration of end to end tools
- Scheduling of data pipeline
- Invoking shell script. Sqoop. Hive

Scala

Learn a scalable, Function based & Object oriented high level programming language

- Scala Introduction
- History Why Scala , Scala Installation
- Function based programming features
- Variable / Values
- Conditional structure
- Looping constructs
- Execute Pattern Matching in Scala
- Exception Handling
- Method creation
- 00Ps concepts (Classes, Objects. Collections, Inheritance, Abstraction and Encapsulation)
- Functional Programming in Scala (Closures. Currying, Expressions, Anonymous Functions)
- Know the concepts of classes in Scala Object Orientation in Scala (Pnmary, Auxiliary Constructors, Singleton Objects, Companion Objects)
- Trans, Moms & Abstract classes

Python

In this module, you will learn about the Git Workflow and case

- Python Introduction
- Evolution

- Application
- Features
- Installation &Configuration
- Objectives
- Flow Control
- Variables
- Data types
- Functions
- Modules
- OOPS
- Python for Spark
- Structures
- Collection types
- Looping Constructs
- Dictionary & Tuples
- File I/O

Spark

Learn the most advanced in- memory, fast, scalable market needed framework for large scale computation

- Spark Introduction
- History
- Overmew
- MR vs Spark
- Spark Libraries
- Why Spark
- RDDs

- Spark Internals
- Pillars of Spark
- Transformations & Actions
- DAG , Lazy evaluation & execution
- Fault Tolerance
- Lineage
- TermInologies
- Ouster types
- Hadoop Integration
- Spark SQL
- Data frames, DataSets
- Optimizers- Catalyst,Tungsten, AST

Spark SQL & Streaming

Learn the Spark SQL & Streaming data Wrangling and Munging techniques for end to end processing framework

- Session
- Structured Streaming
- SQL Contexts
- Hive Context
- RDDs to Relations
- Spark Streaming
- Windowing function
- Why Spark Streaming
- Insurance Hackathon
- Data masking techniques
- Introduction to Spark ML

- Spark UI
- lob Submission mto different cluster managers
- Reusable framework creation
- SDK implementation of Spark
- Building of Fat & ean Jars

Spark Use Cases

Learn the real time data processing with different source and destination system integration

- PYSPARK integration
- Working with PYSPARK Functions
- Developing applications with PYSPARK
- Maven Git Eclipsce integration
- Spark -> NOSQL integration
- Spark options
- Integration with multiple sources & targets
- SCD implementation - Real time use LAWS
- Ebay auction analysis
- US customer data analysis
- End to end real-time integration with NIFI -> Kafka ->; Spark Streaming Amazon 53 -> EC2 -> RDBMS Different Filesystems Hive -: Oozie & Hbase

Kafka

Publish — Subscriber Distributed Message Queue Cluster creation & integration

- Kafka Introduction
- Applications, Cluster Setup
- Broker fault tolerance

- Architecture
- Components
- Partitions & Replication
- Distribution of messages
- Producer & Consumer workload Distribution
- Topics management
- Brokers
- Installation
- Workouts
- Console publishing
- Console Consuming
- Topic options
- Offset Management Cluster deployment in cloud

NIFI

NIFI is a Data flow tool for real time data ingestion into Bigdata platform with tight integration with Kafka & Spark

- NIFI Introduction
- Core Components
- Architecture
- NIFI Installation &Configuration
- Fault tolerance
- Data Provenance Routing,
- Mediation,transformation & routing
- Nifi -> Kafka -> Spark integration
- Workouts
- Scheduling

- Real time streaming
- Kafka producer & consumer
- File streaming with HDFS integration
- Data provenenance
- Packaging NIFI templates
- Rest Api integration
- Twitter data capture

Hue & Ambari

UI tools for working and managing Hadoop and Spark eco systems in a self driven way for development and administration

- Introduction
- Setting up of Ambari and HDP
- Cluster formation guide and Implementation
- Deployment in Cloud
- Full Visibility into Cluster Health
- Metrics & Dashboards
- Heat Maps
- Configurations
- Services, Alerts, Admm activities
- Provisioning, Managing and Monitoring Hadoop Clusters
- Hue Introduction
- Access Hive
- Query executor
- Data browser
- Access Hive. HCatalog, Oozie, File Browser

Hortonworks/Cloudera

The top level distributions for managing Hadoop and spark ecosystems

- Installing and configuring HDP using Ambari
- Configuring Cloudera manager & HDP in sandbox
- Cluster Design
- Different nodes (Gateway, Ingestion, Edge)
- System consideration
- Commands(fsck,job,dfs admin, distcp,balancer)
- Schedulers in RM (Capacity, Fair, FIF0)

Elastic Search

Full Document search store for NOSQL solution with rich real time visualization & analytics capabilities

- History
- Components
- Why ES
- Cluster Architecture/Framework
- All about REST APIs
- Index Request
- Search Request
- Indexing a Document
- limitations
- Install/Config
- Create / Delete / Update
- Get /Search
- Realtime data ingestion with hive
- NIFI integration

- Spark streaming integration
- Hands-on Exercises using REST APIs
- Batch & Realtime Usecases

Kibana

A Raltime integrated Dashboard with rich Visualization &Dashboards with creation of lines, trends, pies, bars, graphs, word cloud

- History
- Components
- Why Kibana
- Trend analysis
- Install/Config
- Creation of different types of visualizations
- Visualization integration into dashboard
- Setting of indexes, refresh and lookup
- Discovery of index data with search
- Sense plugin integration
- Deep Visualizations
- Deep Dashboards
- Create custom Dashboards
- End to end flow integration with Nift,
- Kafka, Spark, ES & Kibana

GitHub & Maven

Repository & Version controller for code management and package generation for dependency Management & collaboration of different components used in TLC

- DevOps Basics

- Versioning
- Create and use a repository
- Start and manage a new branch
- Make changes to a file and push them to GitHub as commits
- Open and merge a pull request
- Create Story boards
- Desktop integration
- Maven integration with Git
- Create project in Maven
- Add scala nature
- Maven operations
- Adding and updating POM
- Managing dependencies with the maven
- Building and installing maven repository
- Maven fat & lean jar build with submit

AWS Cloud

Amazon Web Service components of EC2, 53 storage, access control, Subnets, Athena, Elastic Mapreduce components with Hadoop framework integration

- Introduction to AWS & Why Cloud Managing keys for password less connection
- All about EC2 instance creation till the management
- Amazon Virtual Private Cloud creation Managing the roles with identity Access management
- Amazon object simple storage service (S3) creation with static file uploads and exposure.
- Athena - SQL on top of S3 creation and managing
- Managing AWS EMR cluster with the formation.
- Spark & Hive Integration for data pipeline with S3, Redshift/Dynamo DB, EC2 instance
- Kafka integration

Google Cloud Platform

identify the Platform as a service with the creation and management of Hadoop and Spark cluster in the Google cloud platform

- Registering and managing cloud account
- Key generation
- Cloud compute engine configuration and creation
- Enabling Ambari
- Multi Node cluster setup
- Hardware consideration Software Consideration
- Commands (fsck, job, dfsadmin)
- Schedulers in Resource Manager
- Rack Awareness Policy
- Balancing
- NameNode Failure and Recovery
- Commissioning and Decommissioning a Nodes
- Managing other GCP services
- Cluster health management

Value Added Services

Lets do a smart effort of learning how to prepare resume, interview, projects, answering cluster size, daily activities, roles, challenges faced, data size, growth rate, type of data worked etc.,

- Resume Building & flavoring
- Daily Roles & Responsibilitres
- Cluster formation guidelines
- Interview Questions
- Project description & Flow Execution of end to end 5134.0 practices
- Framework integration with log monitor

- Data size & growth rate
- Architectures of Lambda, Kappa, Master slave. Peer to peer with types of data handled
- Datalake building guide
- Projects discussion
- Package & Development

Use Cases (We cover beyond this)

- Setting up of Single node pseudo Distributed mode Cluster, Hortonworks Sandbox & Cloud based multimode Hortonworks cluster setup and Admin.
- Customer - Transaction data movement using Sqoop.
- Customer - Transaction Data analytics using Hive.
- Profession segmentation, Weblog analysis & Student career analysis using Hive
- Unstructured course data and Students processing using MapReduce.
- Medical and Patient data handling using HBase, Web Statistics low latency data processing using Phoenix.
- Web Server and HDFS data integration with Kafka using NIFI.
- eBay Auction data analytics and SF Police Department data processing using Spark Core.
- Retail Banking data processing using Spark core.
- Server Log Analysis using spark core,Sensus data analysis using Spark SQL.
- Realtime Network, HDFS and Kafka data processing using Spark Streaming.
- Create rich Visualization 8. Dashboard using Kibana with eBay & Trans data
- Managing twitter open data, RESTAP1 data using NIFI-> KAFKA->SPARK

Projects (We cover beyond this.)

- Project 1: Sentimental Analytics - Web event analytics using Linux, HDFS, Hive, Hbase & Oozie.
- Project 2: Server log analysis for view ship pattern, threat management and error handling - Sqoop, Hive, HCatalog, HBase, Phoenix.

- Project 3: Datalake for Usage Pattern Analytics & Frustration scoring of customer - Data Warehouse Migration/consolidation using Sqoop, HDFS, Masking UDF Hive, Oozie, HBase,
- Phoenix.
- Project 4: Realtime Streaming analyrics using Vehicle fleet data using I0T, RPA, Kafka, Spark, NIFI, Kafka, Hive, HBASE/ES, Phoenix.
- Project 5: DataLake exploration using Spark SQL, Hive, HBASE/ES;
- Project 6: Fast Data Processing for Customer segmentation using Kafka, Spark, NIFI, AWS S3, Hive, HBASE/ES.
- 2 Hackathons
- 1 Exams
- 1 Production packaging and deployment
- 1 Cloud formation
- 1 Live Project execution
- 1 Job Support video
- 1 Chat & text mining