BATCH AND REAL TIME ANALYTICS WITH APACHE SPARK.

WEEK 1:SCALA (Object Oriented and Functional Programming)

- Getting started With Scala.
- Scala Background, Scala Vs Java and Basics.
- Interactive Scala – REPL, data types, variables, expressions, simple functions.
- Running the program with Scala Compiler.
- Explore the type lattice and use type inference
- Define Methodsand Pattern Matching.

Scala Environment Set up.

- Scala set up on Windows.
- Scala set up on UNIX.

Functional Programming.

- What is Functional Programming.
- Differences between OOPS and FPP.

Collections (Very Important for Spark)

- Iterating, mapping, filtering and counting
- Regular expressions and matching with them.
- Maps, Sets, group By, Options, flatten, flat Map
- Word count, IO operations,file access, flatMap

Object Oriented Programming.

- Classes and Properties.
- Objects, Packaging and Imports.
- Traits.
- Objects, classes, inheritance, Lists with multiple related types, apply

Integrations

- What is SBT?
- Integration of Scala in Eclipse IDE.
- Integration of SBT with Eclipse.

Week: 2SPARK CORE

- Batch versus real-time data processing
- Introduction to Spark, Spark versus Hadoop
- Architecture of Spark.
- High-level Architecture
  Workers,Cluster Managers,Driver Programs,Executors,Tasks

- Coding Spark jobs in Scala
- Data Sources
- Exploring the Spark shell -> Creating Spark Context.
- RDD Programming
- Operations on RDD.
  Transformations, Actions, Loading Data and Saving Data.
  Key Value Pair RDD.
  Persistence.
- Lazy Operations
  Action Triggers Computation
- Caching
- RDD Caching Methods,RDD Caching Is Fault Tolerant,Cache Memory Management
- Spark Jobs
- Shared Variables,Broadcast Variables,Accumulators
- Configuring and running the Spark cluster.
- Exploring to Multi Node Spark Cluster.
- Cluster management
- Submitting Spark jobs and running in the cluster mode.
- Developing Spark applications in Eclipse
- Tuning and Debugging Spark.
- **Two Projects using Core Spark**

WEEK:3 ->SPARK STREAMING

- Introduction of Spark Streaming.
- Architecture of Spark Streaming.
- Processing Distributed Log Files in Real Time
- Introducing Spark Streaming
  Spark Streaming Is a Spark Add-on
  High-Level Architecture
  Data Stream Sources, Receiver, Destinations
- Application Programming Interface (API)
  StreamingContext
  Basic Structure of a Spark Streaming Application
  Discretized Stream (DStream)
  Creating a DStream
  Processing a Data Stream
  Output Operations
  Window Operation
- Discretized streams RDD.
- Applying Transformations and Actions on Streaming Data
- Integration with Flume and Kafka.
- Integration with Cassandra.
- Monitoring streaming jobs.
- **Use case with spark core and spark Streaming**

WEEK-4 ->SPARK SQL

- Introduction to Apache Spark SQL
- Understanding the Catalyst optimizer
  How it works…,Analysis, Logical plan optimization,Physical planning,Code generation
- Creating HiveContext
- Inferring schema using case classes
- Programmatically specifying the schema
- The SQL context
- Importing and saving data
- Processing the Text files,JSON and Parquet Files
- Data Frames
- Using Hive
- Application Programming Interface (API)
  Key Abstractions,Creating DataFrames,Processing Data Programmatically with SQL/HiveQL
- Processing Data with the DataFrame API
- Saving a DataFrame
- Built-in Functions
  Aggregate,Collection,Date/Time,Math,String,Window
- UDFs and UDAFs
- Interactive Analysis Example
- Interactive Analysis with Spark SQL JDBC Server
- Local Hive Metastore server
- Loading and saving data using the Parquet format
- Loading and saving data using the JSON format
- Loading and saving data from relational databases
- Loading and saving data from an arbitrary source
- Integrating With Hive
- Integrating With MySQl.

WEEK-5 ->SPARK MLIB.

- Introduction to Machine Learning
- Types of Machine Learning.
- Introduction to Apache Spark MLLib Algorithms.
- Machine Learning Data Types and working with MLLib.
- Regression and Classification Algorithms.
- Decision Trees in depth.
- Classification with SVM, Naïve Bayes
- Clustering with K-Means
- Getting Started with Machine Learning Using MLlib
- Creating vectors
- Creating a labeled point
- Calculating summary statistics
- Calculating correlation
- Doing hypothesis testing
- Creating machine learning pipelines using ML
- Supervised Learning with MLlib – Regression

- Using linear regression
- Supervised Learning with MLlib – Classification
- Doing classification using logistic regression
- Doing classification using decision trees
- Doing classification using Random Forests
- Doing classification using Gradient Boosted Trees
- Doing classification with Naïve Bayes
- Unsupervised Learning with MLlib
- Clustering using k-means
- Dimensionality reduction with principal component analysis
- Building the Spark server

WEEK -6 ->SPARK GRAPHX AND CLUSTER MANAGERS

- Introducing Graphs
- Introducing GraphX
- Graph Processing with Spark
- Undirected Graphs,Directed Graphs,Directed Multigraphs,Property Graphs
- Introducing GraphX
- GraphX API
- Data Abstractions
- Creating a Graph,Graph Properties,Graph Operators
- Cluster Managers
- Standalone Cluster Manager
- Architecture
- Setting Up a Standalone Cluster
- Running a Spark Application on a Standalone Cluster
- Apache Mesos
- Architecture
- Setting Up a Mesos Cluster
- Running a Spark Application on a Mesos Cluster
- YARN
- Architecture
- Running a Spark Application on a YARN Cluster

CASSANDRA (NOSQL DATABASE)

- Learning Cassandra
- Getting started with architecture
- Installing Cassandra.
- Communicating with Cassandra.
- Creating a database.
- Create a table
- Inserting Data
- Modelling Data.
- Creating an Application with Web.
- Updating and Deleting Data.

SPARK INTEGRATION WITH NO SQL (CASSANDRA) and AMAZON EC2

- Introduction to Spark and Cassandra Connectors.
- Spark With Cassandra -> Set up.
- Creating Spark Context to connect the Cassandra.
- Creating Spark RDD on the Cassandra Data base.
- Performing Transformation and Actions on the Cassandra RDD.
- Running Spark Application in Eclipse to access the data in the Cassandra.
- Introduction to Amazon Web Services.
- Building 4 Node Spark Multi Node Cluster in Amazon Web Services.
- Deploying in Production with Mesos and YARN.

- **Two REAL TIME PROJECTS Covering all the above concepts.**

QMinds TECHNOLOGIES

Hyderabad | Bangalore | Canada

+91-9036180777, 9686238647, 9618407774

www.qminds.in

Email: info@qminds.in, sivaqminds@gmail.com