Spark – Real Time Streaming Integrated with BigData Hadoop & also covering Cassandra

The Case for Apache Hadoop

- Why Hadoop?
- Core Hadoop Components
- Fundamental Concepts

HDFS

- HDFS Features
- Writing and Reading Files
- NameNode Memory Considerations
- Overview of HDFS Security
- Using the Namenode Web UI
- Using the Hadoop File Shell

Getting Data into HDFS

- Ingesting Data from External Sources with Flume
- Ingesting Data from Relational Databases with Sqoop
- REST Interfaces
- Best Practices for Importing Data

YARN and MapReduce

- What Is MapReduce?
- Basic MapReduce Concepts
- YARN Cluster Architecture
- Resource Allocation
- Failure Recovery
- Using the YARN Web UI
- MapReduce Version 1

Planning Your Hadoop Cluster

- General Planning Considerations
- Choosing the Right Hardware
- Network Considerations
- Configuring Nodes
- Planning for Cluster Management

Hadoop Installation and Initial

- Configuration
- Deployment Types
- Installing Hadoop
- Specifying the Hadoop Configuration
- Performing Initial HDFS Configuration
- Performing Initial YARN and MapReduce Configuration
- Hadoop Logging

Hadoop Clients

- What is a Hadoop Client?
- Installing and Configuring Hadoop Clients

Hive

- Comparison with RDBMS
- HQL
- Data types
- Tables
- Importing and Exporting
- Partitioning and Bucketing – Advanced.
- Joins and Join Optimization.
- Functions- Built in & user defined
- Advanced Optimization of HQL
- Storage File Formats – Advanced
- Loading and Storing Data
- SerDes– Advanced

Sqoop

- Important basics
- Import – Deep dive
- Export – Deep dive
- Sqoop Optimization – Incremental Load

PIG

- Important basics
- Pig Latin
- Data types

- Functions – Built-in, User Defined
- Loading and Storing Data

Flume

- Configure Flume and Import data
- Architecture and LAB

Oozie

- Different workflow jobs
- Ooze scheduler.

Managing and Scheduling Jobs

- Managing Running Jobs
- Scheduling Hadoop Jobs
- Configuring the FairScheduler

Cluster Maintenance

- Checking HDFS Status
- Copying Data Between Clusters
- Adding and Removing Cluster Nodes
- Rebalancing the Cluster
- Cluster Upgrading

Spark –Real Time Streaming

- Introduction to Apache Spark
- What is Spark? Explain about the modules in spark
- Spark-Shell - Scala and python REPL
- Spark Internals - The Driver program, master, workers, executors and the tasks
- Running spark in a standalone mode
- Spark UI and monitoring a job
- Functional programming with spark
- Map-reduce and spark advantages over mapreduce.

RDD

- What is an RDD ?
- Laziness in RDD Evaluation
- Different ways of creating an RDD
- Types of RDD's - PairRDD, DoubleRDD
- RDD Operations
- Partitions - The core of RDD
- textFiles, wholeFiles

Running Spark on a Cluster

- Overview
- A Spark Standalone Cluster
- The Spark Standalone Web UI
- Installing and configuring a cluster

Cassandra - Understanding the architecture

- Key components for configuring
- Data distribution and replication

DataBase Design

- Data modeling
- Compound keys and clustering
- Collection columns (collections set, list, map)
- Expiring columns
- Counter columns
- Using natural or surrogate primary keys
- Indexing
- About indexing (When to use an index and When not to use an index)
- Using multiple indexes
- Using the database
- Querying Cassandra
- Creating a table
- Using a compound primary key
- Inserting data into a table
- Using the keyspace qualifier
- Determining time-to-live for a column
- Determining the date/time of a write
- Adding columns to a table
- Altering the data type of a column
- Removing data
- Expiring columns
- Dropping a table or keyspace
- Deleting columns and rows