Administering Hadoop (5 Days)

This 5 day training course provides System Administrators with a detailed understanding of all the skills required to operate and manage Hadoop clusters. It covers Installation, Configuration, Monitoring and Performance Tuning of Hadoop clusters in diversified Environments.

The course uses lectures and hands-on labs to explore the design, installation, setup, and administration of Apache Hadoop clusters. As well as Hadoop itself, the course introduces Big Data concepts and other members of the Hadoop ecosystem, including Hive, MapReduce, and YARN, as well as Pig, Oozie, Sqoop, HBase, and Flume.

Course Outline

Hadoop Introduction:

Big Data and Hadoop.

- Understanding Big Data and its impact on Application Architectures.
- Hadoop: An Introduction.
- Hadoop Versions
- Apache Hadoop and Commercial Hadoop Flavors.
- Hadoop Enhancement in Hadoop 2.x
- LAB 1: Installing, Configuring and Starting Hadoop 2.X
- LAB 2: Migrating from Hadoop 1.X to Hadoop 2.x

Exploring Hadoop building blocks: HDFS and MapReduce:

- A technical overview of Hadoop.
- Understanding Configuration files.
- Planning Hadoop Cluster installation.
- Introduction to MapReduce and HDFS.
- Setting Up multi Node Hadoop Cluster.
- Working with HDFS command Shell.
- Using Administrative HDFS commands.
- Understanding logs and directory structures in Hadoop.
- Introduction to MapReduce Next Generation (YARN).
- LAB 3: Working With HDFS Command.
- LAB 4: Using HDFS Administrative Command.
- LAB 5: Configuring a single node YARN Cluster and managing YARN Component.

HDFS Deep Dive:

- Understanding key HDFS Features.
- High Availability.
- Automatic Failover
- Using REST interface.

Working with MapReduce and YARN:

- Setting up and configuring MapReduce parameters to execute parallel task.
- MapReduce Configuration in Multinode cluster.
- Understand benefits of NextGen MapReduce (YARN).
- Job management using YARN.
- Working with Capacity Scheduler and Fair Scheduler.
- Using YARN Webservice to manage Cluster resources.
- Implementing HA cluster using shared storage Device.
- Handling single point of failure.

Hadoop Nodes and Topology and Securing Hadoop:

- Understanding Rack and using Rack Topology.
- Commissioning and decommissioning Nodes.
- Securing Hadoop Nodes and processes.
- Understanding Authentication and Authorization.
- LAB 6: MultiNode MultiRack Hadoop Configuration.
- LAB 7: Configuring High Availability for failover.
- LAB 8: Authentication using Kerberos.
- LAB 9: Implementing Service Level Authorization.
- LAB 10: Using Hadoop Auth to enable Kerberos SPNEGO authentication for HTTP.

Setting up Hadoop Ecosystem – 1:

- Installing and integrating Flume with Hadoop.
- Understanding source/sink architecture of flume and work with data ingestion.
- Handling RDBMS data using Sqoop.
- Installing and configuring Sqoop server on Hadoop cluster.
- Working with import/export.
- LAB 11: Flume Data Collection onto HDFS with Avro Serialization.
- LAB 12: Managing RDBMS data using Sqoop.

Setting up Hadoop Ecosystem - 2

- Setting up HIVE in local and MapReduce Mode.
- Controlling HIVE behavior using Mapred configuration variables.
- Configuring Logging for HIVE.
- Setting up Pig and understanding run modes.
- Configuring Pig environment variables.
- Setting up Oozie for workflow Management.
- Submitting, starting, running, suspending, resuming and killing a workflow.
- Understanding Hbase and its benefit.
- LAB 13: Installing Hive and Configuring Hive with MySql
- LAB 14: Install and Configure Pig.
- LAB 15: Working with Workflow using Oozie.
- LAB 16: Setting up HBase on a Hadoop cluster and working with HBase metadata.

Hadoop Cluster Monitoring and Optimization:

- Using basic HDFS admin commands to get the statistics.
- Understanding Log files and log entries.
- Setting up Chukwa for large-scale log collection and analysis.
- Identify Hadoop QoS Metrics.
- Setting up monitoring system using ganglia.
- Understand limitations of monitoring system.
- Evaluate various monitoring Tools (Nagios, Chukwa, System Administrator etc.)
- LAB 17: Data collection, monitoring and analysis system for large clusters Using Chukwa.
- LAB 18: Installing and Setting up Ganglia to monitor MultiNode Hadoop Cluster
- LAB 19: Monitoring Tool Analysis and Using Nagios to configure Alerts.

Maintaining Hadoop Cluster:

- Checking HDFS Health.
- Using Rebalancer.
- Working with Backup and Restore.
- Memory management and YARN Resource Management.
- Managing Hadoop Cluster using Ambari
- LAB 20: Rebalancing a Hadoop cluster.
- LAB 21: Managing Hadoop Cluster Health.
- LAB 22: Managing Memory of Resources.
- LAB 23: Managing and Provisioning Hadoop Cluster using Ambari.

Hadoop (Special Webage Solutions Offering)

- Evaluate Benefits of Hadoop on Cloud.
- Hadoop on Google Cloud Platform. Integrate Hadoop and MongoDB.
- Spark Introduction and Benefits.
- Hadoop and Data warehouse-Dual power Case study.
- LAB 24:Hadoop Installation and Configuration on Amazon E C2.
- LAB 25: Hadoop and MongoDB integration.
- LAB 26: Working with Spark.