Hadoop Master

HadoopAdministrative

Developer

Spark

Projects

9811434634/ 9711198265

Hadoop All in One includes Hadoop Administrator, Hadoop Developer & Apache Spark training programs. Which basically start with building block for Hadoop and takes us on edge for career development. Hadoop is more about the cost-effective, open source framework for distributed storage and processing of extremely large, multi-source data sets. With Hadoop All in One training program, you learn to manage & maintain large amount of data, both structured and unstructured, to extract more meaningful business insights from more of your data.

Administration

Big Data & Hadoop

Big data features and challenges
3Vs for Big Data
Problems with traditional large-scale systems
Why Hadoop? & Hadoop fundamental concepts
Hadoop vs RDBMS vs No-SQL
History of Hadoop with Hadoopable problems
Hadoop distributed file system (HDFS)
Limitation of Hadoop

Hadoop Architecture

Hadoop Version - 2.x & 1.x
Distributions of Hadoop (Cloudera, Hortonworks)
Architecture of Hadoop
Rack awareness and topology
Cluster storage daemons Name Node
Secondary Node
Data Nodes
YARN Responsibilities
Resource Manager
Job History Server
Node Manager
Application Manager
Application Master

Hadoop High-Availability

Name Node availability
Architecture of HA
Implementation of HA
Apache Zookeeper service
1. Quorum Journal
Active Name Node and Standby Name Node
Zookeeper Fail Over controller
Quorum Journal Manager
Quorum Journal Node(s)
2.Namespace federation (NFS)
Namespace information
Zookeeper fail over controller

Linux Initials

Installation of Linux (Red Hat)
Basic Linux configurations
Basic Linux commands
Password less ssh
IP address and hostname
Firewall and selinux
Yum and creating yum repository
NTP configurations

EAGLETFLY SOLUTIONS

981434634/ 9711198265

Planning your Hadoop Cluster

Installation Prerequisites
General planning considerations
Choosing the right hardware
Network considerations
Configuring nodes
Planning for cluster management

Installation & Deployment of Hadoop

Choosing deployment types
Setting up multi-nodes
Setting up Cloudera yum repository
Installation for Cloudera Manager
Installing multi-node Hadoop (Cloudera) environment
Specifying the Hadoop configuration
Performing Initial HDFS configuration
Performing Initial YARN and Map Reduce configuration
Hadoop logging & cluster monitoring

Accessing Hadoop

Access HDFS using command line
Hadoop fs
Hadoop HDFS admin
Access Cloudera Manager (Admin)
Access HUE (Developer)

Configuration for Admin

Add and remove services
Configuring HDFS properties like Block size
Seeting up zookeeper on multi node
Configuring Hadoop operating system (YARN) & Map-Reduce
Configuring schedulers
Hadoop logging & monitoring
Advanced configuration parameters
Configuring Hadoop ports
Explicitly including and excluding hosts

Cluster Maintenance

Checking HDFS status
Copying Data between clusters
Adding and removing cluster nodes
Rebalancing the cluster
Cluster upgrading

BigDataAnalytics

Sendbox/ Quik Start VMs

Overview of sandBox
Different flavours (Virtual Box / VMware) of sandBox
Installation of sandbox
Start working with sandBox

Hue or Hadoop UI

Introduction of HUE
Getting started with HUE
Deployment of jobs
Functional execution of Hive/HBase
Design of work-flow using job designer
Data transfer in Sqoop

Hadoop Shell & Commands

Hadoop developer/admin commands using shell
NameNode & Secondary Name Node commands
HDFs dfsadmin and file system shell commands
Hadoop Name Node / Data Node directory structure
HDFS permissions model
Map-Reduce job deployment
Oozie workflow design
Different components jobs design

Apache Sqoop

Installation of Sqoop
Ingesting data from external (DB) sources with Sqoop
Ingesting data from/to relational databases with Sqoop
Integration of Sqoop and Hive
Best practices for importing data

Map Reduce Concepts

Introduction to map reduce
Architecture of map reduce
Understanding the concept of mappers & reducers
Anatomy of map reduce program
Phases of a map reduce progam
Data-types in hadoop map reduce
Driver, mapper and reducer classes
Input split and record reader
Input format and output format in hadoop
Concepts of combiner and partitioner
Running and monitoring mapreduce jobs
Writing your own map reduce job using map reduce API
Different interview questions raised for map reduce

Apache Hive

Problems with No-SQL database
Introduction & installation Hive
Hive schema and data storage
Data types & introduction to SQL
Hive-SQL: DML & DDL
Hive-SQL: views & indexes
Explain and use the various Hive file formats
Use Hive to run SQL-like queries to perform data analysis
Use Hive to join data sets using a variety of techniques
Map-side joins and Sort-Merge-Bucket joins

Spark

Spark Introducation

Introduction & objectives
Evolution of distributed systems
Limitations of mapreduce in Hadoop
Batch vs. real-time processing
Application of stream & in-memory processing
Components of a Spark
History of Spark
Spark execution architecture
Automatic parallelization of complex flows
Apache Spark a unified platform of big data apps
More benefits of Apache Spark
Running Spark in different modes
Installing spark as a standalone & cluster environment
Overview of Spark on a cluster

Spark SQL

Introduction & objectives
Importance of Spark SQL
Benefits of Spark SQL
SQL Context
Data Frames
Creating a Data Frame
Using datacframe operations
Run Spark SQL with a data frame
Using the reflection-based approach
Using the programmatic approach
Run Spark sql programmatically
Data sources
Save modes
Saving to persistent tables
Schema merging
Spark SQL using JSON & Hive data
DML operation Hive queries

Spark RDDs

Introduction of RDDs
Features of RDDs
Creating RDDs
Creating RDDs referencing an external dataset
Creating RDDs using text files
Creating RDDs using other hadoop input formats
RDD operations & transformations
Features of RDD persistence
Storage levels Of RDD persistence
Choosing the correct RDD persistence storage level
Invoking the Spark shell
Creating the Spark context
basic operations on files in Spark shell RDD
Demo-build a Spark Python project
Build a Spark Java project
Shared variables broadcast & accumulators
Double RDD methods
Pair RDD methods Join
Pair RDD methods Others
General RDD methods
Transformations in RDD
Actions in RDD
Key Value pair RDD in Python
Reading text & sequence file from HDFS
Using groupby operation
Python application performing groupby operation