Hadoop Admin

1. 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

2. 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

3. Hadoop High-Availability

NameNode Availability
Architecture of HA
Implementation of HA
Apache Zookeeper Service

Quorum Journal

- Active NameNode and Standby NameNode
- Zookeeper Fail Over controller
- Quorum Journal Manager
- Quorum Journal Node(s)

Namespace federation (NFS)

- Namespace Information
- Zookeeper Fail Over controller

4. 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

5. PLANNING YOUR HADOOP CLUSTER

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

6. 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

7. Accessing Hadoop

Access HDFS using command line

- Hadoop fs
- Hadoop dfsadmin

Access Cloudera Manager (Admin)
Access HUE (Developer)

8. 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

9. HADOOP SECURITY.

Why Hadoop Security Is Important
Hadoop's Security System Concepts
What Kerberos is and how it Works
Setting up production kerbros
Securing a Hadoop Cluster with Kerberos

10. MANAGING AND SCHEDULING JOBS

Why scheduler
Type of scheduler

- FIFO scheduler
- Fair scheduler
- Capacity scheduler

Scheduling Hadoop Jobs

Managing Running Jobs
Configuring the Fair Scheduler
Introduction of Apache OozieSeeting up jobs with Oozie
Design Workflow in Oozie
Coordinator in Oozie

11. CLUSTER MAINTENANCE

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

12. SQOOP, FLUME & HDFS CLIENT

Sqoop & Flume installation
Ingesting Data from External (RDBMS) Sources with Sqoop
Ingesting Data from/to Relational Databases with Sqoop
Ingesting Data from External Sources with Flume
Integration of Sqoop and Hbase
Integration of Flume and Hbase
Integration of Sqoop and Hive
Best Practices for Importing Data

13. CONCLUSION & FAQS

Hadoop Developer

1 .THE MOTIVATION & LIMITATION FOR HADOOP

Motivation of Hadoop
Big data features and challenges
Problems with Traditional Large-Scale Systems
Why Hadoop & Hadoop Fundamental Concepts
Comparison between Hadoop and RDBMS
Is Hadoop replacing RDBMS?
History of Hadoop with Hadoopable problems
Limitation of Hadoop

2. HDFS CONCEPTS

HDFS Design & Goals
Understand Blocks and Configuration of block size
Block replication and replication factor
Understand Hadoop Rack Awareness and configure racks in Hadoop
File read and write anatomy in HDFS
Enable HDFS Tash
Configure HDFS Name and space Quota
Configure and use WebHDFS (Rest API For HDFS)

3. HADOOP ECOSYSTEM & CLUSTER

Hadoop 2.x & 1.x
Available Distributions of Hadoop (Cloudera, Hortonworks)
Hadoop Projects & Components
Architecture of Hadoop & Planning for cluster
The Hadoop Distributed File System (HDFS)
Cluster Daemons & Its Functions

Name Node
Secondary Node
Data Nodes
Application Master and Task Tracker
YARN Responsibilities
Deployment of Hadoop Cluster

4. APACHE ZOOKEEPER

Introduction to Apache Zookeeper
Zookeeper stand alone installation
Zookeeper Clustered installation
Understand Znodes and Ephemeral nodes
Manage Znodes using Java API
Zookeeper four letter word commands

5. SANDBOX / QUICK START VMS)

Overview of SandBox

Different flavours (Virtual Box / VMware) of SandBox
Installation of SandBox
Start Working with SandBox

6. 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 and flume

7. HADOOP SHELL AND COMMANDS

Hadoop Developer/Admin commands using shell
NameNode & Secondary NameNode Commands
HDFS DFSAdmin and File system shell commands
Hadoop NameNode / DataNode directory structure
HDFS permissions model
HDFS Offline Image Viewer
Map-Reduce job deployment
Oozie workflow design
Different Components Jobs design

8. MAPREDUCE CONCEPTS

Introduction to MapReduce
Architecture of Map-Reduce
Understanding the concept of Mappers & Reducers
Anatomy of Mapreduce program
Phases of a MapReduce progam
Data-types in Hadoop MapReduce
Driver, Mapper and Reducer classes
InputSplit and RecordReader
Input format and Output format in Hadoop
Concepts of Combiner and Partitioner
Running and Monitoring MapReduce jobs
Writing your own MapReduce job using MapReduce API
Writing Mappers and Reducers with the Streaming API
Different interview questions raised for Map-Reduce

9. SPARK

Problems with Traditional Large-Scale Systems
Introducing Spark & Spark Basics
Apache MapReduce vs Apache Spark
Using the Spark Shell
Resilient Distributed Datasets (RDDs)
Functional Programming with Spark
Working with RDDs
Key-Value Pair RDDs
Types of RDDs
MapReduce and Pair RDD Operations
Read data from text file
Using HDFS with Spark

10. Spark DataFrame Dates And Timestamps

Introduction to Date & timestamps
Working with Dates & timestamps

Spark DataFrame Aggregate Operations
Aggregate and GroupBy concepts
Sorting & Ordering

11. HCATALOG OR METASTORE TABLES

Introduction of apache Hcatalog
Creating tables using Hcatalog
Bulk uploads using MetaStore Tables
Play with semi-structured data
Integration of Hcatalog with Hive
Hive SQL query analysis

12. 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, including
Map-side joins and Sort-Merge-Bucket joins
Integration to HBase & Cassandra
Sentiment Analysis and N-Grams
Hive Thrift Service

13. Spark SQL

Introduction to Spark SQL
Inferring a schema
Applying a schema
Loading and Writing schema
SQL caching and UDF
Spark SQL queries to perform computations

14. FLUME

Installation of Flume
Ingesting Data from External Sources with Flume
Configuration for flume
REST Interfaces
Best Practices for Importing Data

15. SQOOP

Installation of Sqoop
Ingesting Data from External (RDBMS) Sources with Sqoop
Ingesting Data from/to Relational Databases with Sqoop
Integration of Sqoop and Hbase
Integration of Sqoop and Hive
Best Practices for Importing Data

16. MANAGING AND SCHEDULING JOBS

Managing Running Jobs
Scheduling Hadoop Jobs

Configuring the FairScheduler

Machine Learning

INTRODUCTION OF MACHINE LEARNING WITH TENSORFLOW

Motivation of Machine Learning?
Use Cases of Machine Learning
Future Scope of Machine Learning
Real World Domain using ML
Types of Machine Learning
Different tools/framework available for ML
Limitation of Machine Learning tensorflow

BASICS OF MACHINE LEARNING

Understanding supervised & unsupervised learning Understanding bases associated with any machine learning algorithm

Ways of reducing bias and increasing generalisation Introduction to scikit-learn & SciPy in Python

MACHINE LEARNING TECHNIQUES

Supervised & Unsupervised Learning
Recommender Systems

- User Based recommendation Engine
- Item Based recommendation Engine Logistic Regression

Regularization

Linear Regression with One Variable Linear Algebra Review
Linear Regression with Multiple Variables

NAIVE BAYES

Use Naive Bayes with scikit learn in python/Mahout.
Splitting data between training sets
Testing sets with scikit learn (Python/Mahout).
Calculate the posterior probability
Pior probability of simple distributions

SUPPORT VECTOR MACHINES (SVM)

Learn the simple intuition behind Support Vector Machines.
Implement an SVM classifier in SKLearn/scikit-learn Choose the right kernel for your SVM Learn about RBF and Linear Kernels

NEURAL NETWORKS: REPRESENTATION

NEURAL NETWORKS: LEARNING

Advice for Applying Machine Learning