

Big Data & Hadoop

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

Available version Hadoop 1.x & 3
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. HIGH AVAILABILITY AND NAMESPACE FEDERATION

Introduction to HDFS Federation
Understand Name service ID and Block pools
Introduction to HDFS High Availability
Failover mechanisms in Hadoop 2.x
Concept of Active and StandBy NameNode
Configuring Journal Nodes and Split brain scenario
Automatic and manual failover techniques in HA
HDFS HA admin commands
Understand Name Node Safe mode, File system image and edits

5. YARN - YET ANOTHER RESOURCE NEGOTIATOR

YARN Architecture in Hadoop 2.x

Yarn Components

- Resource Manager
- Node Manager
- Job History Server
- Application Time Line Server
- MR Application Master

YARN Application execution flow

Running and Monitoring YARN Applications

Understand and Configure Capacity / Fair Schedulers in YARN

YARN Rest API

Writing and executing YARN applications

6. 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

7. INSTALLATION & DEPLOYMENT OF HADOOP

Setting up Local Cloudera/Hortonworks repository

Installation for Cloudera Manager/Ambari

Overview of Cloudera Manager/Apache Ambari

Installing Multi (atleast 10 Machine) Node Hadoop

(Cloudera/Hortonworks)

Setting up Cloudera/Hortonworks Hadoop environment

Specifying the Hadoop Configuration

Performing Initial HDFS Configuration

Performing Initial YARN and Map-Reduce Configuration

Hadoop Logging & Cluster Monitoring

8. SANDBOX / QUICK START VMS

Overview of SandBox

Different flavours (Virtual Box / VMware) of SandBox

Installation of SandBox

Start Working with SandBox

9. HUE OR HADOOP UI

Introduction of HUE

Getting started with HUE

Deployment of Map-Reduce

Functional Execution of Hive/HBase

Design of work-flow using Job Designer

Data transfer in Sqoop and flume

10. MAP REDUCE CONCEPTS

Introduction to MapReduce
Architecture of Map-Reduce
Understanding the concept of Mappers & Reducers
Anatomy of MapReduce program
Phases of a MapReduce program
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

11. 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

12. HBASE: THE HADOOP DATABASE

Problems with RDBMS
Introduction to HBase
HBase components - Hbase master and Region servers
Non-RDBMS, Not-Only SQL or No-SQL
Installation HBase & Deployment Types
CRUD & Batch Operations
Filters, Counters, Pool
Rest Interface & Web-UI

13. 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

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



mapping
MINDS

Python

1. PROGRAMMING LANGUAGE?

What is Programming Language?
Overview of Python, Version and why use Python?
Hardware Overview
Installing Python on Windows OR Linux
Using Python and Writing A Program
Installing Python and Writing A Program
Writing the "Hello World" Assignment
Variables types and properties
Quiz, questions and queries

2. CONTROL FLOW AND LOOPS

if-elif-else Statements
Logical, Boolean Expressions
Making Decision, Flow Control
Loops and Iteration
for & while Statements
Break, Continue & Pass statements
range() Function
Use if range() function in Loops

3. DEFINING FUNCTIONS

What are functions
User-defined & Pre-defined functions
Functions Argument Values
Default & Keyword Argument Values
Init function & Self Argument
Arbitrary & Unpacking Argument Lists
Lambda Expressions
Function Annotation

4. DATA STRUCTURE

What Are Linear Structures?
The Queue, Stack, and Deque Interfaces
The List Interface: Linear Sequences
The USet Interface: Unordered Sets
The SSet Interface: Sorted Sets
Dictionary, Tuples, Range, xrange
Stacks, Queues, DQueues
Use of Strings

5. CLASSES

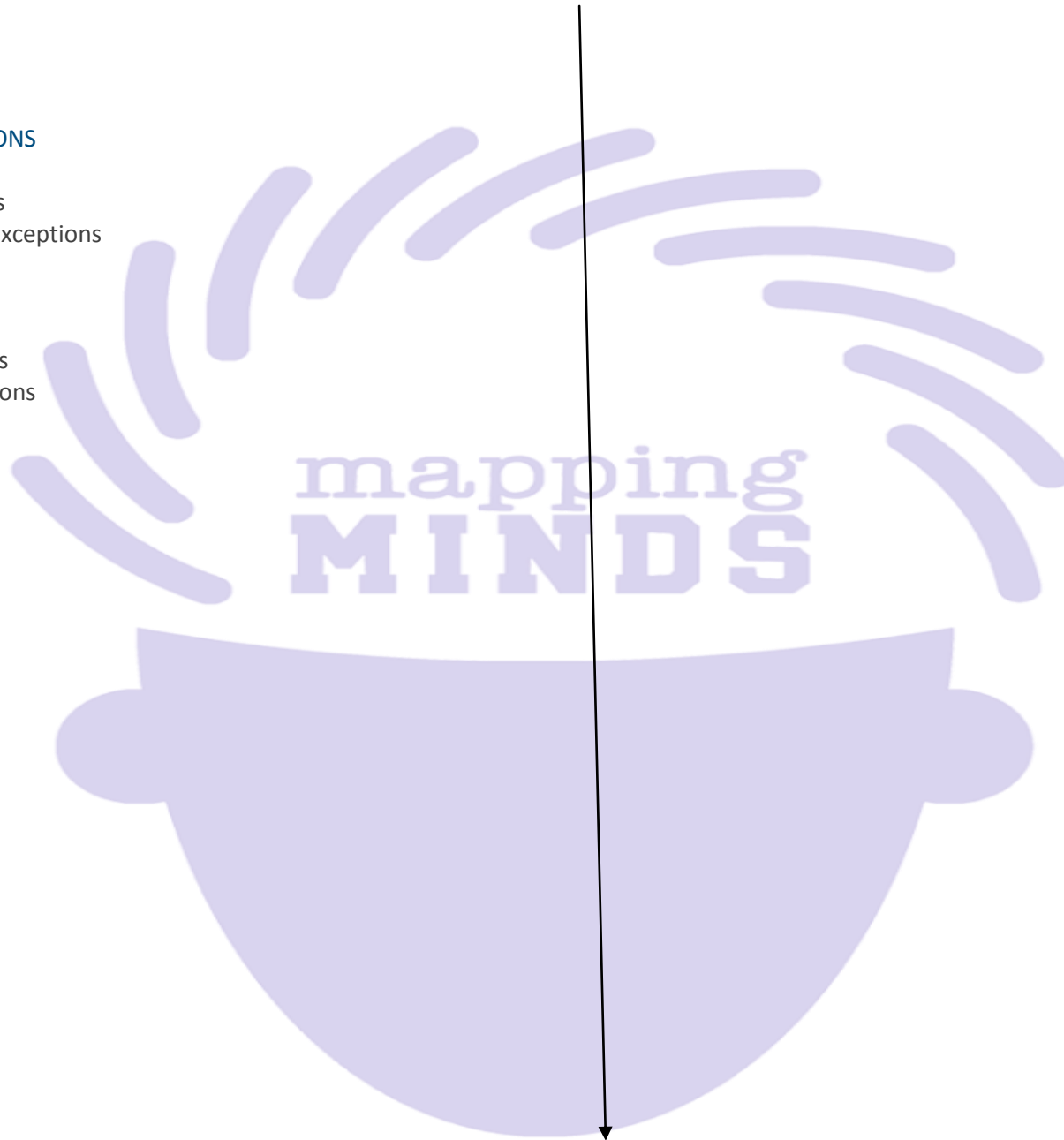
Names and Objects
Python Scopes and Namespaces
Classes
Inheritance
Private Variables
Odds and Ends
Iterators
Generators
Generator Expressions

6. FILES & EXCEPTION HANDLING

Working with Files
Opening a text file
Reading & Writing a file
File Operations
Dealing with errors
Modules & Importing Modules
Regular Expressions
Introduction to List Comprehensions
List Comprehension Operations

7. ERRORS AND EXCEPTIONS

- Syntax OR Compile Errors
- Exceptions OR Runtime Exceptions
- Handling Exceptions
- Raising Exceptions
- User-defined Exceptions
- Defining Clean-up Actions
- Predefined Clean-up Actions



Machine Learning

INTRODUCTION OF MACHINE LEARNING WITH TENSORFLOW

1. 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

2. 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

3. 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

4. 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

5.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

R with Data Science

1: FUNDAMENTALS OF 'R'

First steps with R
Discover the data types & variable in R
Installing R on personal machines. retrieving R packages.
Basics of R, R-Studio, R Markdown.
Data types, variable assignment

2: VECTORS

Analyze gambling behaviour using vectors.
Create, name and select elements from vectors.
Comparing Vectors
Selection from Vectors
Sorting of Vectors

3: MATRICES

Work with matrices in R
Computations with matrices
Demonstrate your knowledge by analyzing the any data figures
Comparing Matrices
Selection from Matrices
Sorting of Matrices

4: FACTORS

R stores categorical data in factors
Learn how to create subset and compare categorical data.
Comparing Factors
Selection from Factors
Sorting of Factors

5: DATA FRAMES

Learn how to create data frames
Data sets and structure
Selection from data frames
Sorting of data frames

6: LISTS

Learn how to create list
list and data structure
Selection and Sorting from/of list

7: MODULE

If/else statements.
For/while loops.
Functions
Apply() family over data
Utilities like with(), grepl(), sub() to specify environment

8:MODULE

Writing Functions in R
A quick refresher for functions
Functional programming
Advanced inputs and outputs
Robust functions

9: IMPORTING DATA INTO R

Importing data from flat files
Importing data from Excel
Importing data from Databases
Importing data from the web

ADVANCE R DATA VISUALIZATION

1: MODULE

The Grammar of Graphics
Lines and Syntax
Transformations
Interactivity and Layers
Customizing Axes, Legends.

2: MODULE

Data Visualization with ggplot2
Introduction & Data
Aesthetics
Geometries
qplot and wrap-up
Statistics
Coordinates and Facets
Themes

DATA VISUALIZATION - BEST PRACTICES & CASE STUDY

Statistical Modelling

Intro to Statistics with R:
Histograms and Distributions
Scales of Measurement
Measures of Central Tendency
Measures of Variability

MACHINE LEARNING

Introduction to Machine Learning

Supervised Learning
Unsupervised Learning
Semi-Supervised Learning Performance
Measures Classification
Regression Clustering