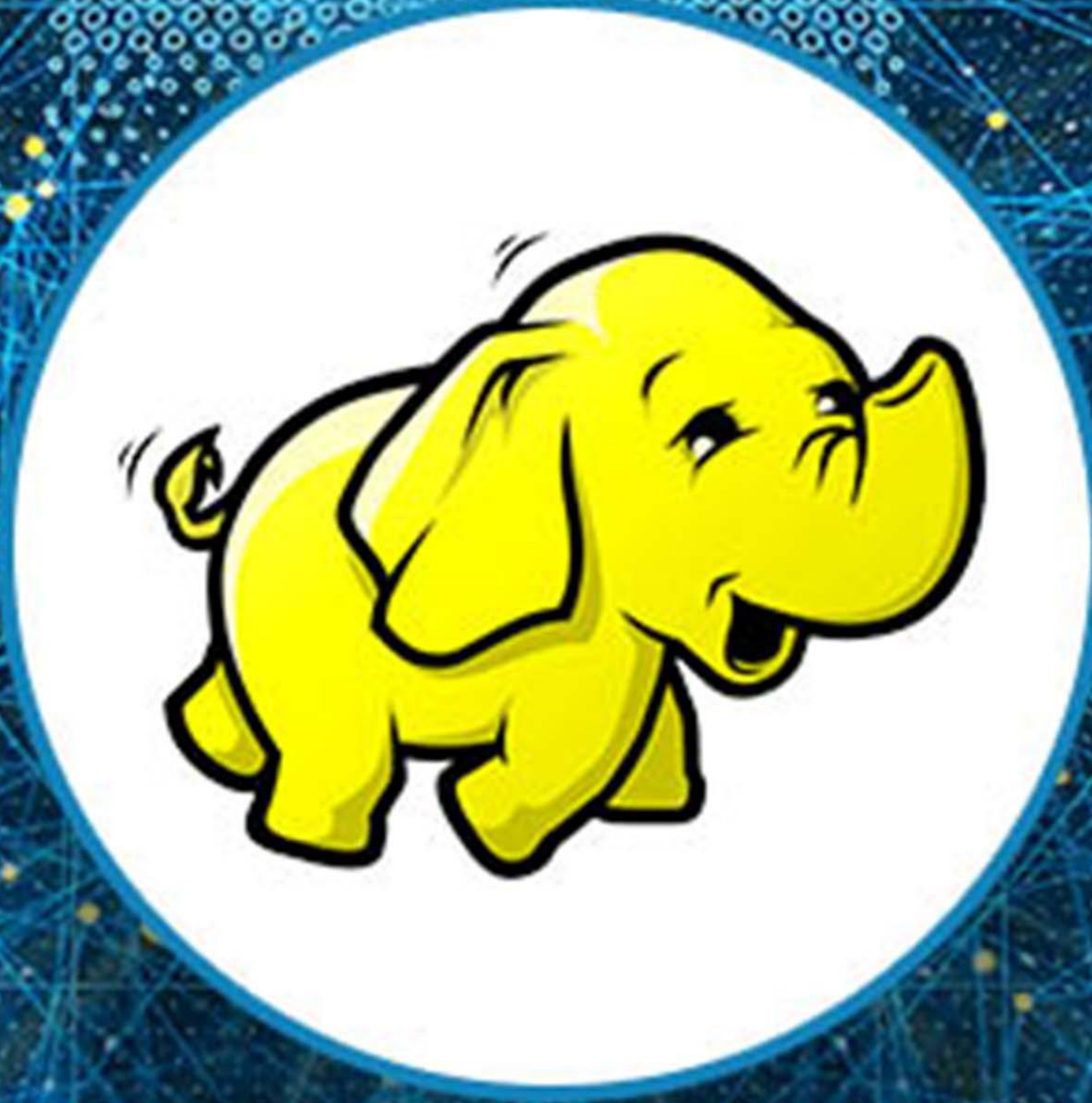
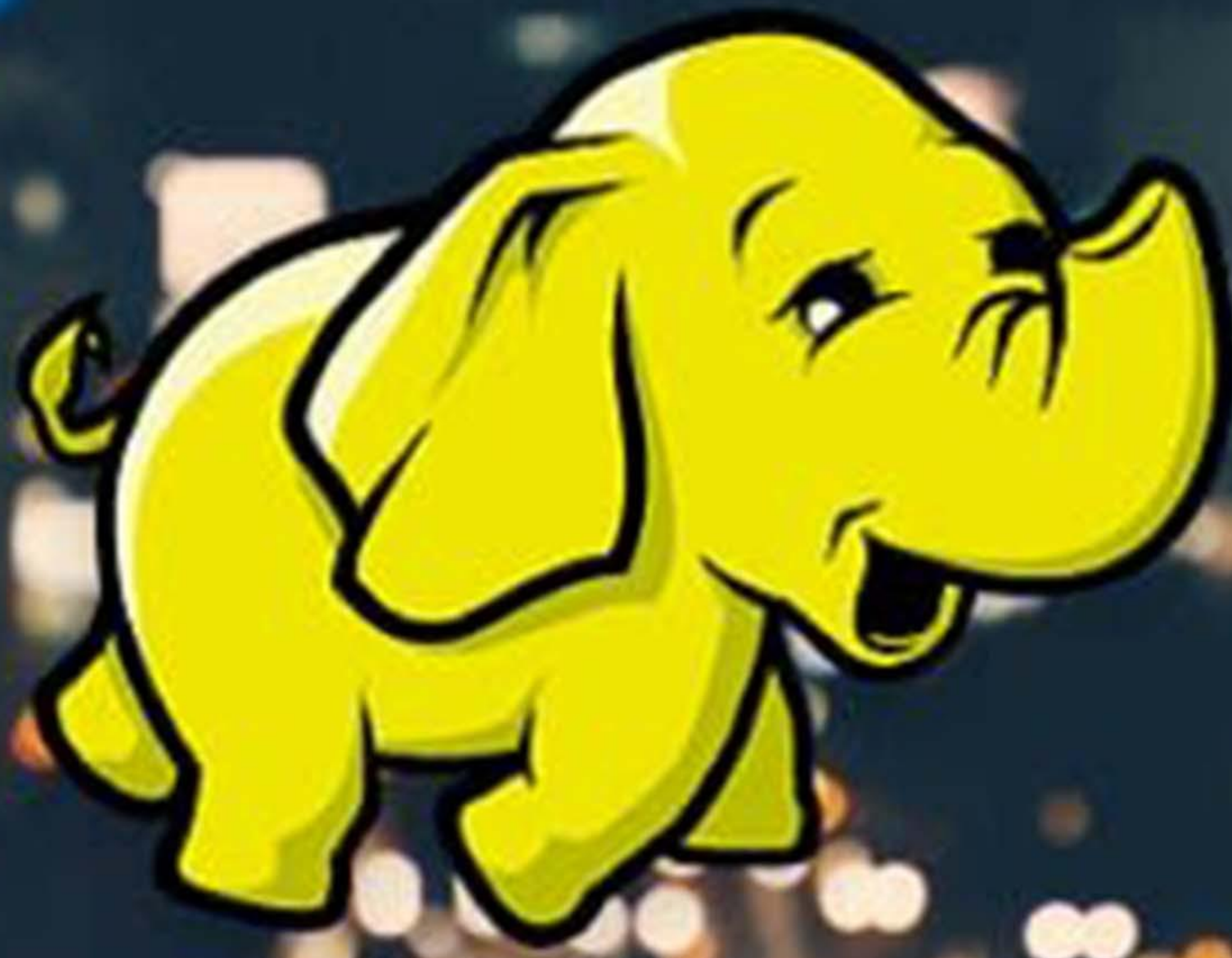




Big Data Hadoop with Python



About Course

Hadoop framework is written in Java language; however, Hadoop programs can be coded in Python or C++ language. We can write programs like MapReduce in Python language, while not the requirement for translating the code into Java jar files.

BIG DATA HADOOP WITH PYTHON

CURRICULUM

- 1 Understanding Big Data And Hadoop**
 - Introduction to Big Data & Big Data Challenges
 - Limitations & Solutions of Big Data Architecture
 - Hadoop & its Features
 - Hadoop Ecosystem
 - Hadoop 2.x Core Components
 - Hadoop Storage: HDFS (Hadoop Distributed File System)
 - Hadoop Processing: MapReduce Framework
 - Different Hadoop Distributions
- 2 Hadoop Architecture And HDFS**
 - Hadoop 2.x Cluster Architecture
 - Federation and High Availability Architecture
 - Typical Production Hadoop Cluster
 - Hadoop Cluster Modes
 - Common Hadoop Shell Commands
 - Hadoop 2.x Configuration Files
 - Single Node Cluster & Multi-Node Cluster set up
 - Basic Hadoop Administration

3

Hadoop MapReduce Framework

Traditional way vs MapReduce way

Why MapReduce

YARN Components

YARN Architecture

YARN MapReduce Application Execution Flow

YARN Workflow

Anatomy of MapReduce Program

Input Splits, Relation between Input Splits and HDFS Blocks

MapReduce: Combiner & Partitioner

Demo of Health Care Dataset

Demo of Weather Dataset

4

Advanced Hadoop MapReduce

Counters

Distributed Cache

MRunit

Reduce Join

Custom Input Format

Sequence Input Format

XML file Parsing using MapReduce

5

Apache Pig

Introduction to Apache Pig

MapReduce vs Pig
Pig Components & Pig Execution
Pig Data Types & Data Models in Pig
Pig Latin Programs
Shell and Utility Commands
Testing Pig scripts with Punit
Aviation use-case in PIG
Pig Demo of Healthcare Dataset

6

Apache Hive

Introduction to Apache Hive
Hive vs Pig
Hive Architecture and Components
Hive Metastore
Limitations of Hive
Comparison with Traditional Database
Hive Data Types and Data Models
Hive Partition
Hive Bucketing
Hive Tables (Managed Tables and External Tables)
Importing Data
Querying Data & Managing Outputs
Hive Script & Hive UDF
Retail use case in Hive
Hive Demo on Healthcare Dataset

7

Advanced Apache Hive And HBase

Hive QL: Joining Tables, Dynamic Partitioning

Custom MapReduce Scripts

Hive Indexes and views

Hive Query Optimizers

Hive Thrift Server

Hive UDF

Apache HBase: Introduction to NoSQL Databases and HBase

HBase v/s RDBMS

HBase Components

HBase Architecture

HBase Run Modes

HBase Configuration

HBase Cluster Deployment

8

Advanced Apache HBase

HBase Data Model

HBase Shell

HBase Client API

Hive Data Loading Techniques

Apache Zookeeper Introduction

ZooKeeper Data Model

Zookeeper Service

HBase Bulk Loading

Getting and Inserting Data

HBase Filters

9

Processing Distributed Data With Apache Spark

What is Spark

Spark Ecosystem

Spark Components

What is Scala

Why Scala

SparkContext

Spark RDD

10

Dozie And Hadoop Project

Dozie

Dozie Components

Dozie Workflow

Scheduling Jobs with Dozie Scheduler

Demo of Dozie Workflow

Dozie Coordinator

Dozie Commands

Dozie Web Console

Dozie for MapReduce

Combining flow of MapReduce Jobs

Hive in Dozie

Hadoop Project Demo

Hadoop Talend Integration



www.softcrayons.com



(+91) 854 501 2345



@softcrayons



info@softcrayons.com



693, Sector 14-A, Vasundhara,
Ghaziabad (U.P.), 201012