Parul® University

BIG DATA ANALYTICS
CHAPTER 4

Prepared By: Prof. Himadri Vegad

CHAPTER-4

Hadoop Related Tools:

Overview of HBase

Column-oriented NoSQL database
Runs on HDFS for real-time read/write
Suitable for big, sparse tables

Pig Introduction

High-level framework for Hadoop
Uses Pig Latin scripting language
Simplifies ETL pipelines

Pig Data Model

Atom
Tuple
Bag
Map

Hive

Data warehouse tool on Hadoop

Uses HiveQL (SQL-like)

Transforms queries into MapReduce/Spark jobs

Hive: Data Types & File Formats

Primitive: INT, STRING, FLOAT
Complex: ARRAY, MAP, STRUCT
File formats: Text, ORC, Parquet

HiveQL Data Definition

CREATE DATABASE/TABLE
ALTER TABLE
DROP TABLE
Partition management

HiveQL Data Manipulation

LOAD DATA
INSERT INTO
UPDATE (limited)
DELETE (limited)

HiveQL Queries

SELECT, WHERE, GROUP BY
JOIN, ORDER BY, SORT BY
Aggregation functions

Pig Latin Overview

LOAD, FILTER, FOREACH, GROUP

JOIN, ORDER, DUMP, STORE

Procedural data pipelines

Pig vs Hive

Pig → ETL, script-based, procedural

Hive → DW, SQL-like, declarative

Pig for programmers, Hive for analysts

Using JSON

Semi-structured data format
Common in APIs & NoSQL databases
Easy to parse in Hadoop ecosystem

Overview of Cassandra

Distributed NoSQL database
Peer-to-peer architecture
High availability and scalability

Jasper Reports

Reporting engine generating PDF/HTML
Uses XML templates
Integrates with databases & Hadoop