Cauvery College for Women (Autonomous)

Nationally Accredited (III Cycle) with 'A' Grade by NAAC

Annamalai Nagar, Tiruchirappalli-18.


CAUVERY COLLEGE FOR WOMEN
TIRUCHIRAPPALLI
கற்க நிற்க

Name of the Faculty	: Dr Sinthu Janita
Designation	: Professor & Head
Department	: Computer Science
Contact Number	: 9894484436
Programme	: MSc Computer Science
Batch	: 2016-2017 Onwards
Semester	: IV
Course	: Big Data Analytics
Course Code	: P16CSE5A
Unit	: IV
Topics Covered	: Essential of Hadoop ecosystem, RDBMS versus Hadoop, Key Aspects and Components of Hadoop,

Unit IV

Hadoop Foundation for Analytics:

Hadoop Foundation for Analytics

Unit IV

Hadoop Foundation for Analytics:

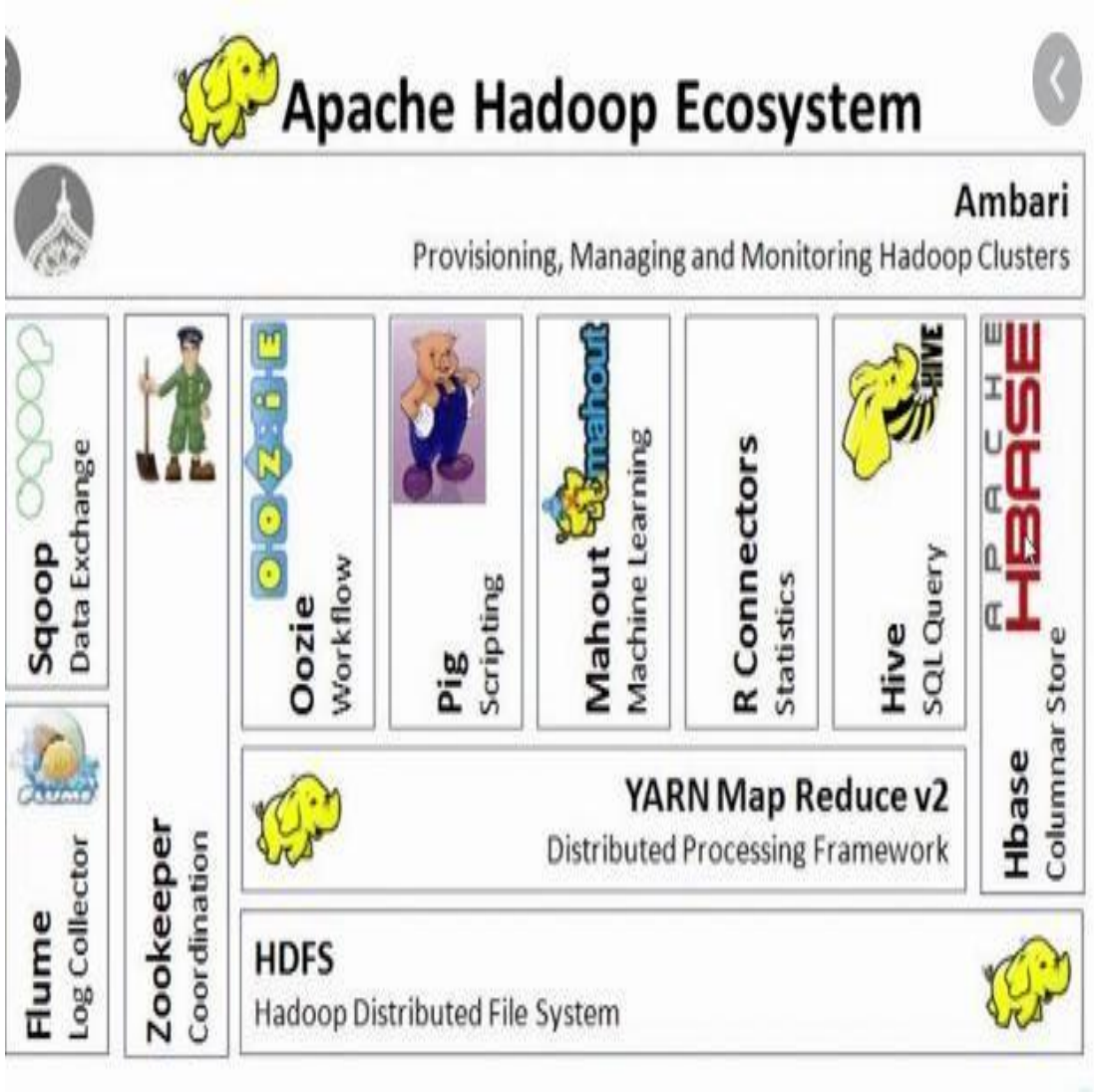
History, Needs, Features, Key Advantage and Versions of Hadoop, Essential of Hadoop ecosystem, RDBMS versus Hadoop, Key Aspects and Components of Hadoop, Hadoop Architectures

Essential of Hadoop Ecosystems

Supports projects to enhance the functionality of Hadoop Core Components

. The Eco projects are

- HIVE
- PIG
- SQOOP
- HBASE
- FLUME.
- OOZIE
- MAHOUT



Essential of Hadoop Ecosystems

Supports projects to enhance the functionality of Hadoop Core Components.

The Eco projects are

- HIVE: It enables analysis of large data sets using a language similar to standard ANSI SQL. Enables to access data stored on a Hadoop Cluster
- PIG: Easy to understand data flow language. Helps with the analysis of large data sets. Even without the proficiency in MapReduce, the data in the Hadoop cluster can be analysed as PIG scripts are automatically converted into MapReduce jobs by the PIG interpreter
- SQOOP: Used to transfer bulk data between Hadoop and structured data stores as RDBMS

Essential of Hadoop Ecosystems ctd..

- HBASE: It is Hadoop's database and compares well with an RDBMS. It supports structured data storage for large tables
- FLUME:Is a distributed, reliable and available software for efficiently collecting, aggregating and moving large amounts of log data. Has simple and flexible architecture.
- OOZIE: It is a workflow scheduler system to manage Apache Hadoop jobs
- MAHOUT: It is a scalable machine learning and data mining library

RDBMS versus HADOOP

PARAMETERS	RDBMS	HADOOP
System	Relational database Management System	Node Based Flat Structure
Data	Suitable for structured data	Suitable for structured, unstructured data, Supports variety of data formats in real time such as XML, JSON, text based flat file formats etc.
Processing	OLTP	Analytical, Big Data Processing

RDBMS versus HADOOP ctd..

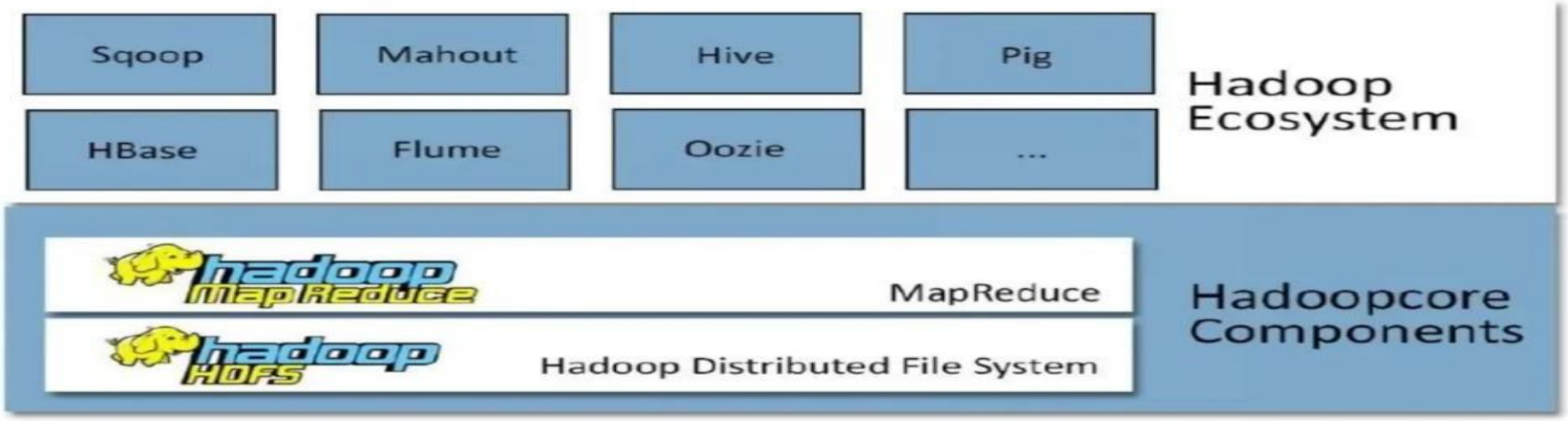
PARAMETERS	RDBMS	HADOOP
Choice	When the data needs consistent relationship	Big Data processing, which does not require any consistent relationships between data
Processor	Needs expensive hardware or high-end processors to store huge volumes of data	In a HADOOP cluster, a node requires only a processor, a network card and few hard drives
Cost	Cost around $10,000 to $14,000 per terabytes of storage	Cost around $4,000 per terabytes of storage

Key Aspects of Hadoop

1. • Open Source Software
   It is free to download, use and contribute
2. • Framework
   The requirements to develop and execute and application is provided-program tools etc.
3. • Distributed
   Divides and stores data across multiple computers.
   Computation/Processing is done in parallel across multiple connected nodes
4. • Massive Storage
   Stores colossal amounts of data across nodes of low-cost commodity hardware

- Faster Processing
  Large amounts of data is processed in parallel yielding quick reponse

Components of Hadoop



Core Components	Hadoop Ecosystem
HDFS Storage Components Distribute data across several nodes Natively redundant	•HIVE •PIG •SQOOP •HBASE •FLUME. •OOZIE •MAHOUT
MapReduce Computational framework Splits a task across several nodes Process data in parallel	