Cauvery College for Women (Autonomous)

Nationally Accredited (III Cycle) with ‘A’ Grade by NAAC

Annamalai Nagar, Tiruchirappalli-18.

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	: History, Needs, Features, Key Advantage

and Versions 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

History of HADOOP

- Hadoop was created by Doug Cutting and Mike Cafarella in 2005.

- It was originally developed to support distribution for the Nutch search engine project.

- In 2006, Hadoop was released by Yahoo and today is maintained and distributed by Apache Software Foundation (ASF).

Features

- Handles massive quantities of structured, semistructured and unstructured data using commodity h/w
- Has shared nothing architecture
- Replicates data across multiple computers-Replica
- For high throughput rather than latency
- Batch processing therefore response time is not immediate
- Complements OLTP and OLAP
- Not a replacement for RDBMS
- Not good when work cannot be parallelized
- Not good for processing small files

Key Advantages of Hadoop

1 Stores data in its native form(HDFS)

- No structure that is imposed in keying or storing data
- Schema less
- Only when data needs to be processed that structure is imposed on new data

2 Scalable

- Can store and distribute very large data sets across hundred of inexpensive servers that operate in parallel

3 Cost Effective

- Has a much reduced cost/terabyte of storage and processing

Key Advantages of Hadoop (ctd..)

4 Resilient to Failure

- Fault tolerant. Practices replication of data. When data is sent, it is replicated.

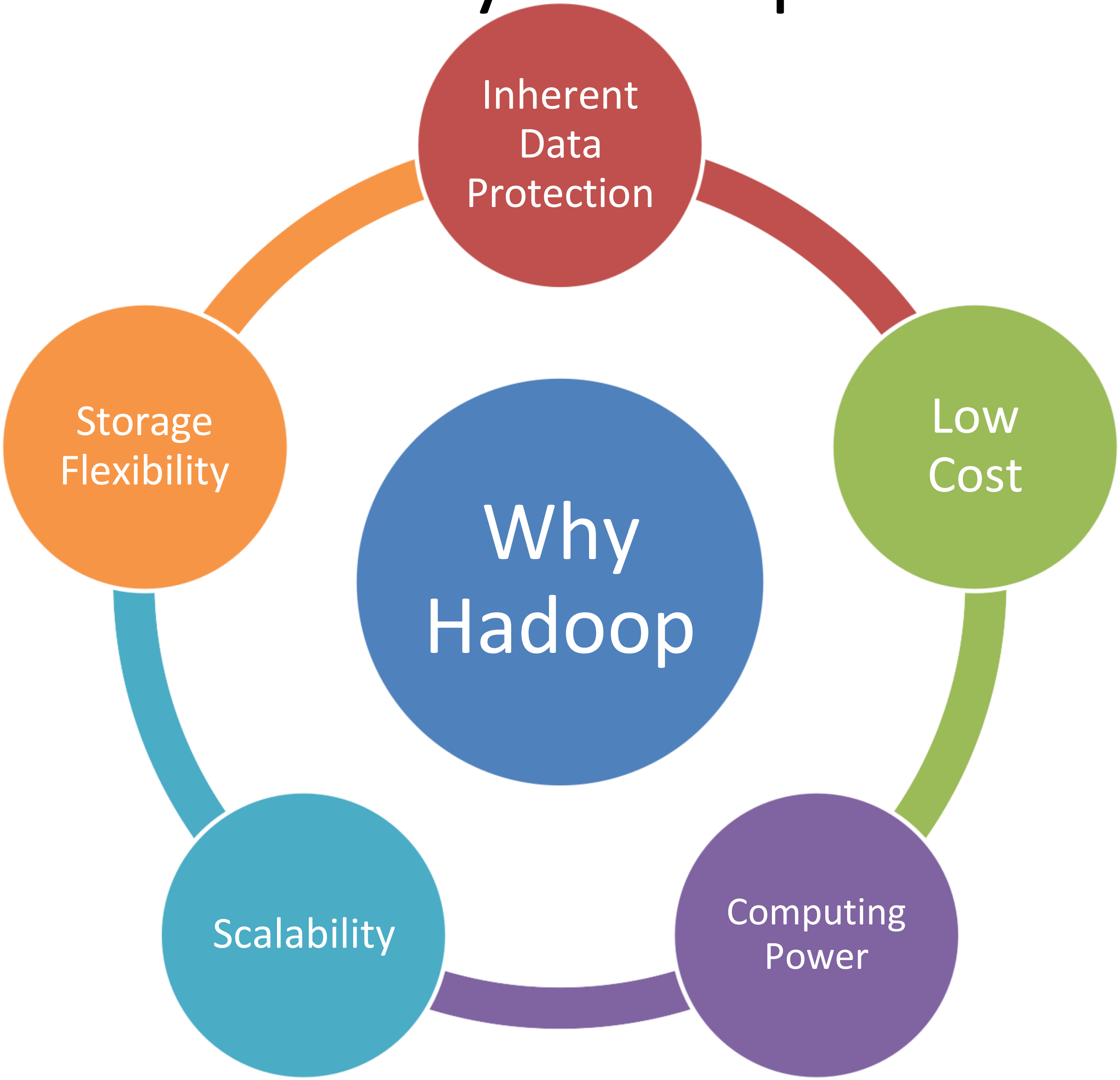
5 Flexibility

- Works with all type of data structures. Helps drive meaningful information from email, social media. ClickStreamData.
- Put to several purpose such as log analysis, data mining, recommendation systems, market campaign analysis etc.

6 Fast

- Extremely fast. Moves code to data.

Why Hadoop



Inherent Data Protection

Low Cost

Computing Power

Scalability

Storage Flexibility

Why Hadoop

Versions of Hadoop

Hadoop 1.0

Hadoop 2.0

Hadoop 1.0

- Data storage Framework
- Data processing framework:

Versions of Hadoop

Hadoop 1.0

Data storage Framework

- HDFS is schemaless. Stores data files in data format. Stores files close to original form.

Data processing framework:

- Uses two functions MAP and REDUCE to process data.
- “Mappers” take in a set of key value pairs and generate intermediate data.
- “Reducers” act on this input to produce the output data. Two functions work in isolation enabling high distributed in a high parallel, fault tolerant and scalable way.

Hadoop 1.0

Limitations:

- Requires MapReduce programming expertise with proficiency required in other programming languages like Java
- Supported batch processing suitable for tasks such as log analysis, large scale data mining projects.
- Tightly computationally coupled with MapReduce. Either rewrite their functionality in MapReduce so that it could be executed in Hadoop or extract the data from HDFS and process it outside of Hadoop. None of the options were viable as a Hadoop. Led to process inefficiencies caused by the data being moved in and out of Hadoop cluster.

Hadoop 2.0

- HDFS continues to be the data storage framework.
- Yet Another Resource Negotiator(YARN) has been added
- Any application capable of dividing itself into parallel tasks is supported by YARN
- YARN co ordinates the allocation of the subtasks of the submitted applications thereby enhancing flexibility, scalability and efficiency of the applications

Hadoop 2.0 ctd..

- It works by having ApplicationMaster in place of the JobTracker , Running applications on resources governed by a new NodeManager
- MapReduce programming expertise is no longer required
- It supports Batch Processing and also Real time processing
- Data Processing Functions such as Data Standardisation, Master Data Management can now be performed in HDFS.