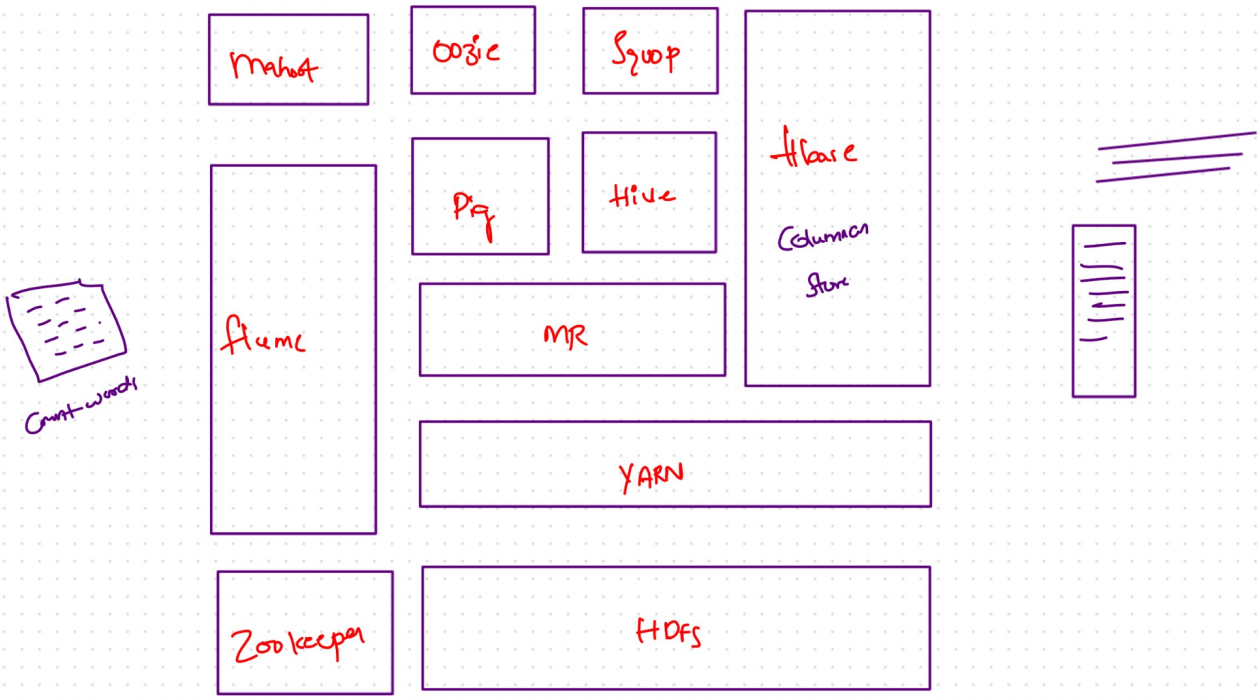
Hadoop Ecosystem Components



Extended components

4. **Hive:**
   - * developed by facebook in 2009
   - * Query engine (not a database)

   key feature: Abstracts MapReduce by translating SQL queries into MapReduce jobs

   Solves: Avoids writing java code while retaining MapReduce performance

   Use Case: Data warehousing solutions

5. **Pig:**
   - * created by yahoo

   key feature: High-level scripting language (PigLatin) for MapReduce

   Use case: data transformation without writing Java or SQL

   Abstracts MapReduce in the backend

6. Sqoop:

Key feature: It helps to import & export data between Hadoop & relational databases

Use Case: Data migration without writing MapReduce code

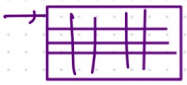
7. Oozie:

* Created by Cloudera
* Key feature: XML-based workflow scheduling & automation

Use case: manage job dependencies using xml for scheduling complex workflows

8. HBase:

row {key: _ }



* NoSQL database modeled after Google's Big Table
* Columnar store (i.e stores data by columns rather than rows)

Key feature: Allows real-time reads & writes on HDFS

Use Case: Time series data, messaging system

9. Mahout:

* Data science component
* Key feature: provides machine learning libraries

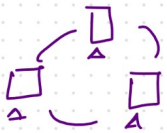
Use Case: implements algorithms for clustering, classification, collaborative filtering

10 flume:

data ingestion & streaming

* messaging queue (similar to kafka)
* Key feature: collects log or event data from various sources
* Use Case: Real-time analysis, monitoring

11 Zookeeper:



* Developed by Yahoo

* Key feature: co-ordinates distributed system to maintain consistency

Use case: Critical for ensuring reliability in Hadoop Cluster

Note:

Hadoop framework Properties

(1) Loosely coupled framework:

* components can be removed or replaced while the system working

eg: MapReduce being replaced by Spark for processing

(2) Easy Integration:

* Can connect with other big data technologies (eg. Spark)

* Can integrate with non-big data technologies

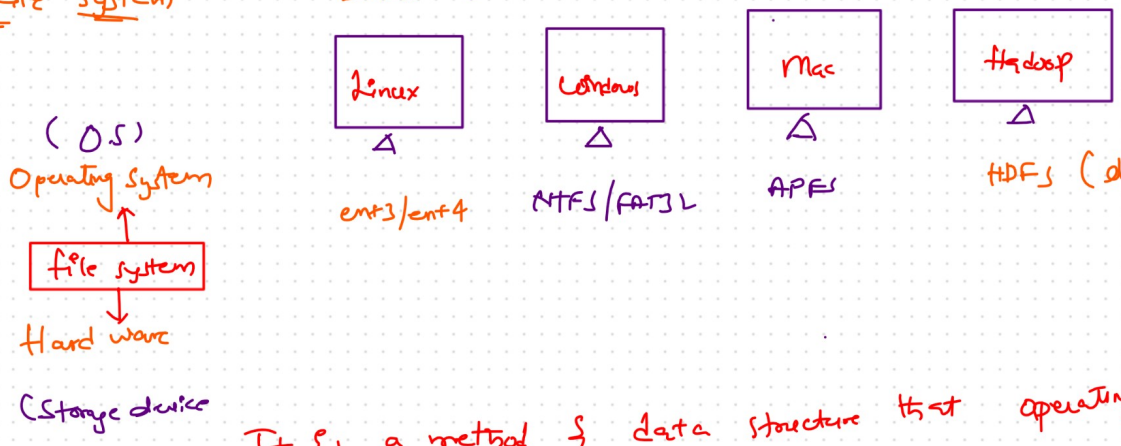
eg: (MySQL, other relational databases)

Key Terminology

HDFS = Hadoop Distributed file System

1. File System

Servers → Speed, Second

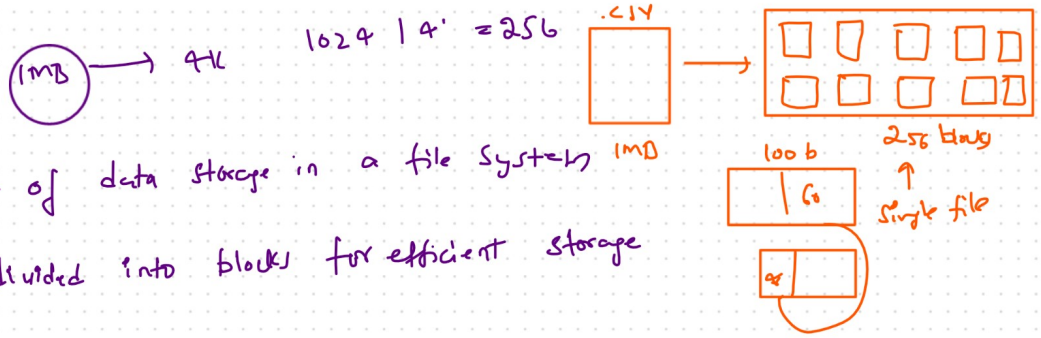


→ It is a method & data structure that operating systems use to manage files on storage devices

→ Acts as a layer b/w software (O.S) & hardware (Storage devices)

→ determines how data is stored, retrieved and organized

2. Blocks



* Smallest unit of data storage in a file system
* files are divided into blocks for efficient storage

eg: Windows (NTFS) : 4KB block size

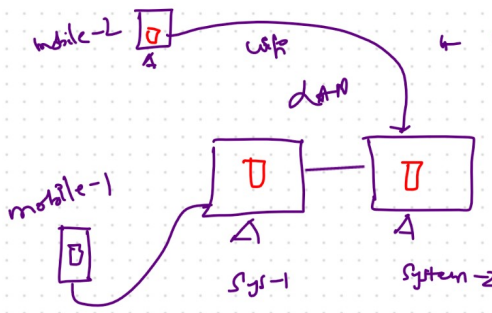
HDFS : 128 MB default block size

Benefits :

* Quick & efficient for data storage & retrieval
* efficient use of storage space

3. Types of File Systems

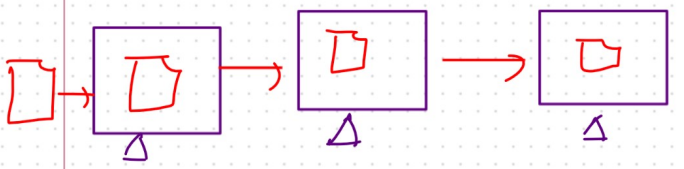
* Standalone File Systems
* Distributed File Systems



→ files are stored & managed on a single machine
→ even in connected environment, files stay on one machine

Standalone File System

Distributed File System (HDFS)



→ files are stored across multiple machines in a cluster

Node A Node B Node C

3 Node cluster

more scalable — can add hundreds of machine to increase storage