Hadoop

- An open source framework to handle massive amount of data in a distributed and scalable ways

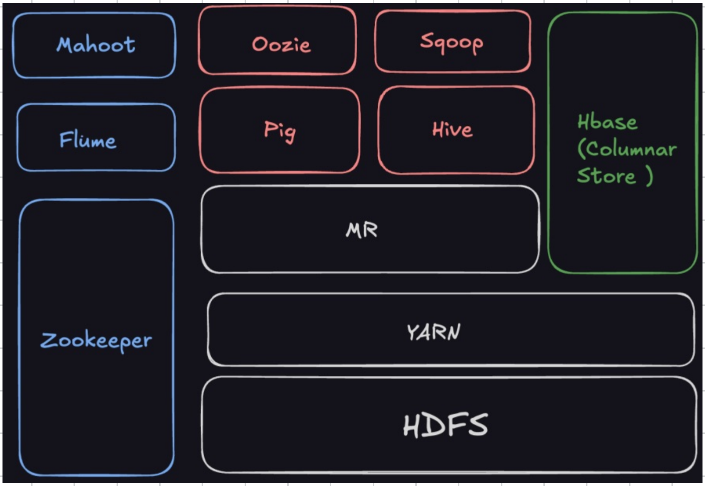
GFS → Storing
MapReduce → Processing

- 2 tasks
  - massive data storage
  - faster parallel processing

* Properties

- Scalability : Hadoop can scale horizontally
- Fault tolerance : Hadoop maintains copies/replicas to avoid failure if any single machine failed
- Distributed processing : Hadoop can process the data where it is stored
- Cost effectiveness: Inexpensive hardware / commodity machines can be used
- open source : Free to use and modify

* Hadoop Ecosystem



* Note : Hadoop properties

- Loosely coupled framework: we can remove components and it's still going to work
- Integration : can be connected to other frameworks easily

HDFS - distributed storage
MapReduce - divide into smaller tasks
YARN - decouple the resource management

Hive - Query engine (not a database)
Abstract MR by translating SQL queries into MR jobs

Pig - Abstraction of MR (MR performance but no Java/SQL)
High level scripting language for MR

Sqoop - Facilitate import/export between Hadoop and Relational DTB

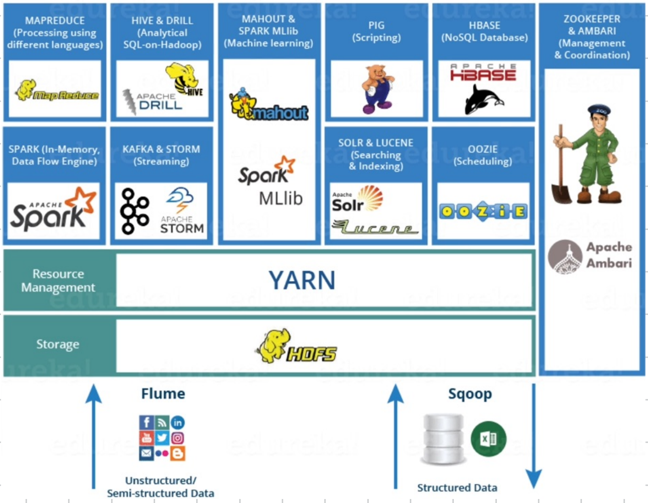
Oozie - Abstraction of MR
Use xml files for scheduling and automating our work
→ essential for scheduling complex workflows in ETL

Hbase - NoSQL DTB which allows real time read and write on HDFS

Mahoot - DS component
provide ML libraries

Flume - Messaging queue
Collect logs or event data from various sources and deliver them to Hadoop and Hbase
Real-time analytics for monitoring and ingesting streaming data

Zookeeper - Coordinate distributed system to maintain consistency
Critical for ensuring reliability in Hadoop clusters


MAPREDUCE (Processing using different languages)
MapReduce
HIVE & DRILL (Analytical SQL-on-Hadoop)
APACHE DRILL
HIVE
MAHOUT & SPARK MLlib (Machine learning)
mahout
Spark MLlib
PIG (Scripting)
HBASE (NoSQL Database)
APACHE HBASE
ZOOKEEPER & AMBARI (Management & Coordination)
Apache Ambari
SPARK (In-Memory, Data Flow Engine)
APACHE Spark
KAFKA & STORM (Streaming)
APACHE STORM
SOLR & LUCENE (Searching & Indexing)
Apache Solr
Lucene
OOZIE (Scheduling)
OOZIE
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/ Semi-structured Data
Sqoop
Structured Data

Storage: HPFS, Hbase
Processing: MR, Pig, Hive, Spark
Data ingestion: Flume, Sqoop
Coordination: Zookeeper
Workflow management: Oozie