Hadoop is an open-source framework for storing and processing large-scale data across distributed clusters using commodity hardware. The Hadoop Ecosystem is a suite of tools and technologies built around Hadoop's core components (HDFS, YARN, MapReduce and Hadoop Common) to enhance its capabilities in data storage, processing, analysis and management.

Components of Hadoop Ecosystem
Hadoop Ecosystem comprises several components that work together for efficient big data storage and processing:

HDFS (Hadoop Distributed File System): Stores large datasets across distributed nodes.
YARN (Yet Another Resource Negotiator): Manages cluster resources and job scheduling.
MapReduce: A programming model for batch data processing.
Spark: Provides fast, in-memory data processing.
Hive & Pig: High-level tools for querying and analyzing large datasets.
HBase: A NoSQL database for real-time read/write access.
Mahout & Spark MLlib: Libraries for scalable machine learning.
Solr & Lucene: Tools for full-text search and indexing.
Zookeeper: Manages coordination and configuration across the cluster.
Oozie: A workflow scheduler for managing Hadoop jobs.
hadoopEcosystem
Key Components of Hadoop Ecosystem
Note: Apart from above mentioned components, there are many other components too that are part of Hadoop ecosystem.

All these components revolve around a single core element "Data". That’s the beauty of Hadoop, it is designed around data, making its processing, storage and analysis more efficient and scalable.

Let’s explore these key components of the Hadoop ecosystem in detail.

HDFS
HDFS is a core component of Hadoop ecosystem, designed to store large volumes of structured or unstructured data across multiple nodes. It manages metadata through log files and splits storage tasks between two main parts:

NameNode (master): Stores metadata (data about data) and requires fewer resources.
DataNodes (slaves): Store actual data on commodity hardware, making Hadoop cost-effective.
HDFS handles coordination between clusters and hardware, serving as the backbone of entire Hadoop system.

YARN
YARN (Yet Another Resource Negotiator) is resource management layer of Hadoop, responsible for scheduling and allocating resources across the cluster. It has three key components:

ResourceManager: Allocates resources to various applications in the system.

NodeManager: Manages resources (CPU, memory, etc.) on individual nodes and reports to ResourceManager.
ApplicationMaster: Acts as a bridge between ResourceManager and NodeManager, handling resource negotiation for each application.
Together, they ensure efficient resource utilization and smooth execution of jobs in the Hadoop cluster.

MapReduce
MapReduce enables distributed and parallel data processing on large datasets. It allows developers to write programs that transform big data into manageable results.

Map(): Processes input data by filtering, sorting and organizing it into key-value pairs.
Reduce(): Takes the output from Map(), aggregates the data and summarizes it into a smaller, consolidated set of results.
Together, they efficiently handle large-scale data transformations across the Hadoop cluster.

PIG
Pig is a platform developed by Yahoo for analyzing large datasets using Pig Latin, a SQL-like scripting language designed for data processing.

It simplifies complex data flows and handles MapReduce operations internally.
The processed results are stored in HDFS.
Pig Latin runs on Pig Runtime, similar to Java on JVM.
It enhances programming ease and optimization, making it a key part of Hadoop ecosystem.
HIVE
Hive uses a SQL-like interface (HQL: Hive Query Language) to read and write large datasets.

It supports both real-time and batch processing, making it highly scalable.
Hive supports all standard SQL datatypes, easing query operations.
It has two main components:

JDBC/ODBC drivers: manage data connections and access permissions.
Hive Command Line: used for query execution and processing.
Mahout
Mahout brings machine learning capabilities to Hadoop-based systems by enabling applications to learn from data using patterns and algorithms.

It provides built-in libraries for clustering, classification and collaborative filtering.
Users can invoke algorithms as needed through Mahout's scalable, distributed libraries.
Apache Spark
Apache Spark is a powerful platform for batch, real-time, interactive, iterative processing and graph computations.

It uses in-memory computing, making it faster and more efficient than traditional systems.
Spark is ideal for real-time data, while Hadoop suits batch or structured data, so both are often used together in organizations.
Apache HBase

HBase is a NoSQL database in Hadoop ecosystem that supports all data types and handles large datasets efficiently, similar to Google's BigTable.

It is ideal for fast read/write operations on small portions of data within massive datasets. HBase offers a fault-tolerant and efficient way to store and retrieve data quickly, making it useful for real-time lookups.
Other Components
Apart from core components, Hadoop also includes important tools like:

Solr & Lucene: Used for searching and indexing. Lucene (Java-based) offers features like spell check and Solr acts as its powerful search platform.

Zookeeper: Handles coordination and synchronization between Hadoop components, ensuring consistent communication and grouping across the cluster.

Oozie: A job scheduler that manages workflows. It supports two types of jobs:

Workflow jobs (executed in sequence)
Coordinator jobs (triggered by data or time-based events).