Table of Contents

[**1.Apache Spark** 2](#_Toc122090705)

[**2.Apache Hadoop** 3](#_Toc122090706)

[**3.Similarities** 4](#_Toc122090707)

[**4.Differences** 4](#_Toc122090708)

[**5.References** 5](#_Toc122090709)

### **1.Apache Spark**

Apache Spark is an open-source Data Processing Engine that is meant for big datasets. Large jobs are divided up by Spark employing several nodes, and data is processed and cached in Random Access Memory (RAM). A unified engine that supports SQL queries, streaming data, machine learning, and graph processing is another example of an Apache Spark use case. APIs in Spark are made to be simple to use when processing and manipulating semi-structured data (GFG). Additionally, Apache Spark is the only processing framework that integrates both AI and data, enabling users to run cutting-edge machine learning (ML) and AI algorithms after performing extensive data transformations and analysis (IBM). As seen in Figure 1, the Spark Ecosystem consists of the following modules:

* Spark Core: It oversees duties like task dispatching, input and output activities, and scheduling.
* Spark SQL: Collects data on structured data so that users can optimise the processing of that data.
* Spark Streaming: These modules increase the capacity for stream processing. Data from various streaming sources, such as HDFS, Kafka, and Kinesis, is divided into micro-batches by Spark Streaming in order to depict a continuous stream. A more recent strategy built on Spark SQL called "Structured Streaming" aims to decrease latency and make programming easier.
* MLib: Scalability and increased accessibility are the objectives of this component.
* GraphX: This has a collection of APIs that are employed to make graph analytics activities easier.

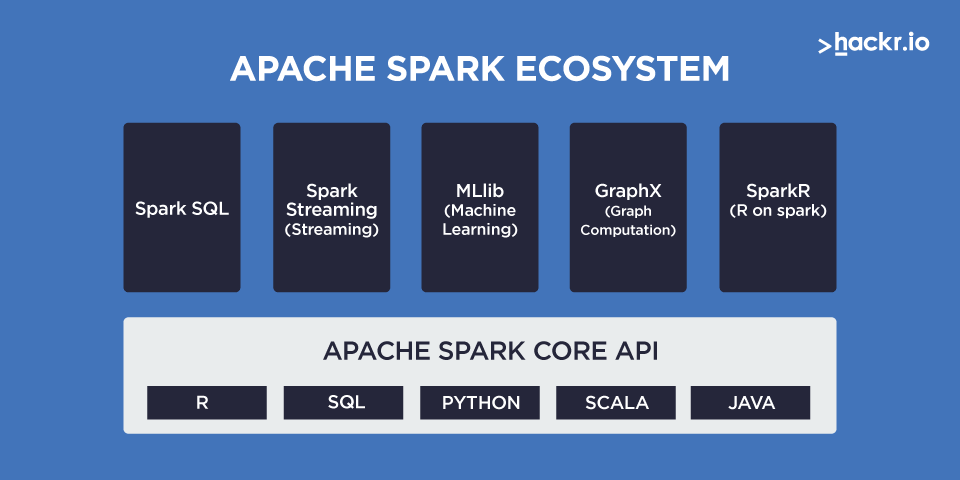


Figure : Apache Spark Ecosystem

# **2.Apache Hadoop**

With the help of nodes, Apache Hadoop is an open-source software tool that enables users to manage large data collections. It is a solution that stores, and processes structured, semi-structured, and unstructured data that is extremely scalable and economical (GFG). The benefits of Hadoop include Data Security in the event of hardware failure, incredibly large scalability and real-time analytics for reviewing the past and making decisions. Advanced analytics, such as predictive analysis, data mining, machine learning, etc., are supported by Hadoop for stored data. It makes it possible to divide up huge data analytics processing duties into smaller activities. The little tasks are dispersed over a Hadoop cluster and completed in parallel using an algorithm (like MapReduce) (IBM). According to Figure 2, the Hadoop Ecosystem consists of the following:

* Hadoop Distributed File System (HDFS): This parallelizes file storage across a cluster in a Hadoop-native style. It controls the distribution of huge data collections among a Hadoop Cluster. Both structured and unstructured data can be handled using Hadoop.
* Yet Another Resource Negotiator (YARN): A timetable for coordinating application runtimes.
* Hadoop MapReduce: To put the parts together for the desired outcome, the algorithm analyses the input in parallel.
* Hadoop Common (Hadoop Core): Also known as Hadoop Core, it supports all other components. All other modules rely on its collection of shared libraries and utilities.

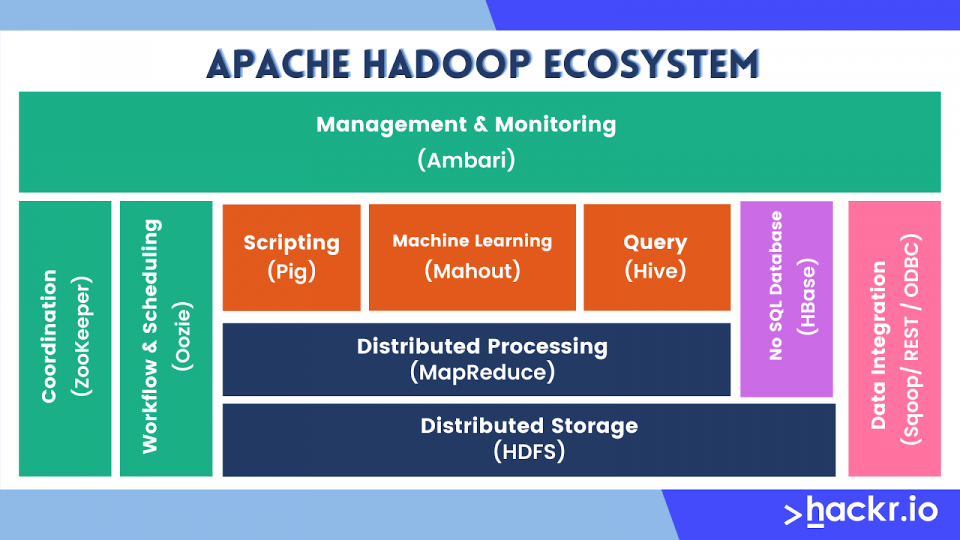


Figure : Apache Hadoop Ecosystem

# **3.Similarities**

The Apache Software Foundation is responsible for the development of both Spark and Hadoop, two open-source platforms for big data. Each framework includes a broad ecosystem of open-source solutions for handling, managing, and analysing large data collections. SQL queries can be run on both Spark and Hadoop's respective systems by their users. Spark employs SparkSQL, which is intimately linked with the Spark Core, whereas Hadoop uses Apache Hive and Impala to accomplish the same task (GFG). The fact that both platforms, Spark through MLib and Hadoop through Apache Mahout, allow users to do machine learning is another similarity between them. Moreover, both platforms have strong fault tolerance. Hadoop is extremely fault tolerant; failing applications don't need to be relaunched from scratch. Without further code or configuration, Spark streaming recovers previous work and provides exactly one semantic out of the box (GFG).

# **4.Differences**

Despite the similarities, Spark and Hadoop are more different than they are alike. The first regarding their Processing Speed. As seen in Figure 3, Spark is 100 times faster than Hadoop, due to its in-memory processing system. When it comes to Stream Processing, Hadoop doesn’t support it whereas Spark uses micro batches for every streaming workload. Hadoop also costs way less as it runs on less expensive hardware, whereas Spark requires a lot of RAM to run (GFG). Hadoop also cannot cache the data in-memory for future needs, while Spark is able to do so. On the other hand, Hadoop can run on commodity hardware, but Spark needs mid to high-end hardware to do so. Hadoop also cannot support Graph Processing, but Spark can using GraphX. In terms of Scheduling, Hadoop requires external schedulers like Apache Ozzie, whereas Spark has its own scheduler. Hadoop also offers configurable memory management, while Spark has automatic memory management with its latest update.

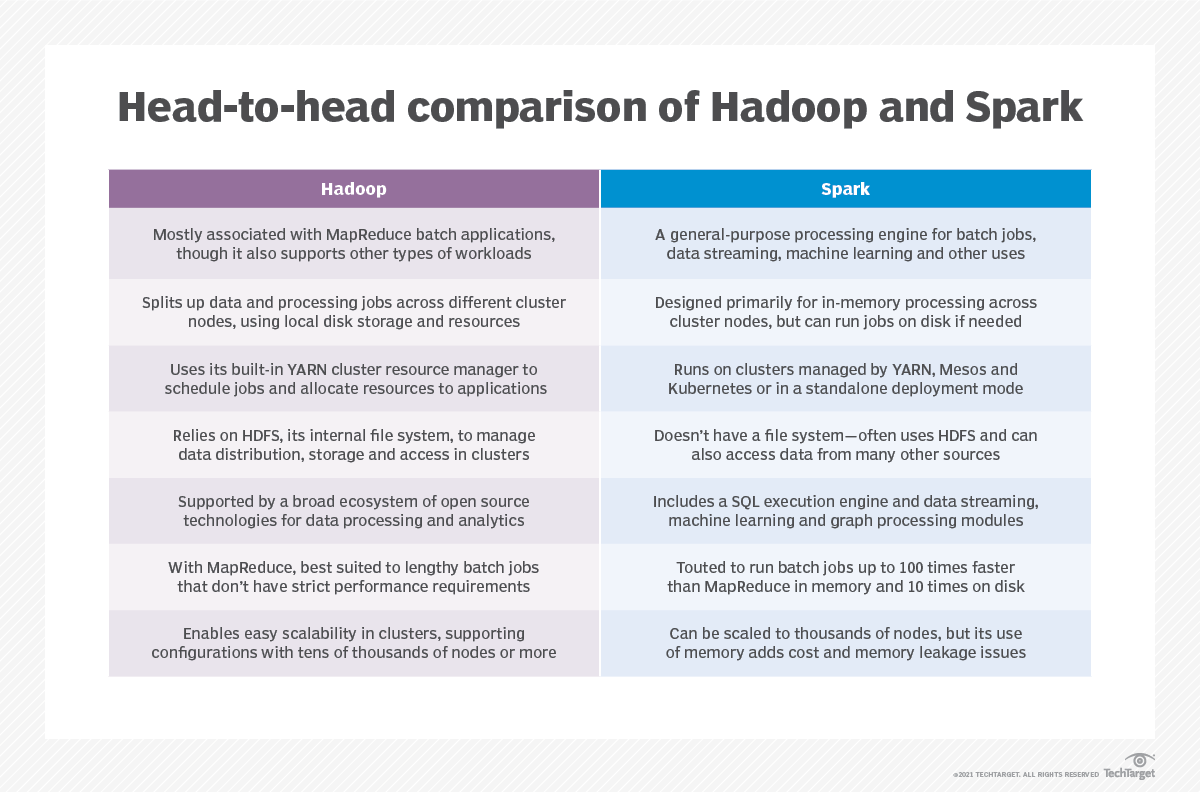


Figure : Comparison between Spark and Hadoop

# **5.References**

GFG.*Difference between Hadoop and Spark*. GeeksforGeeks. (2022, September 20). Retrieved December 11, 2022, from https://www.geeksforgeeks.org/difference-between-hadoop-and-spark/

IBM.*Hadoop vs. spark: What's the difference?* IBM. (n.d.). Retrieved December 11, 2022, from https://www.ibm.com/cloud/blog/hadoop-vs-spark

Figure 1.Hari, S. (n.d.). *Hadoop vs spark: A head to head comparison in 2022* . Hackr.io. Retrieved December 11, 2022, from https://hackr.io/blog/hadoop-vs-spark

Figure 2. Hari, S. (n.d.). *Hadoop vs spark: A head to head comparison in 2022* . Hackr.io. Retrieved December 11, 2022, from https://hackr.io/blog/hadoop-vs-spark

Figure 3. Lawton, G. (2022, February 17). *Hadoop vs. spark: An in-depth Big Data Framework Comparison: TechTarget*. Data Management. Retrieved December 11, 2022, from https://www.techtarget.com/searchdatamanagement/feature/Hadoop-vs-Spark-Comparing-the-two-big-data-frameworks