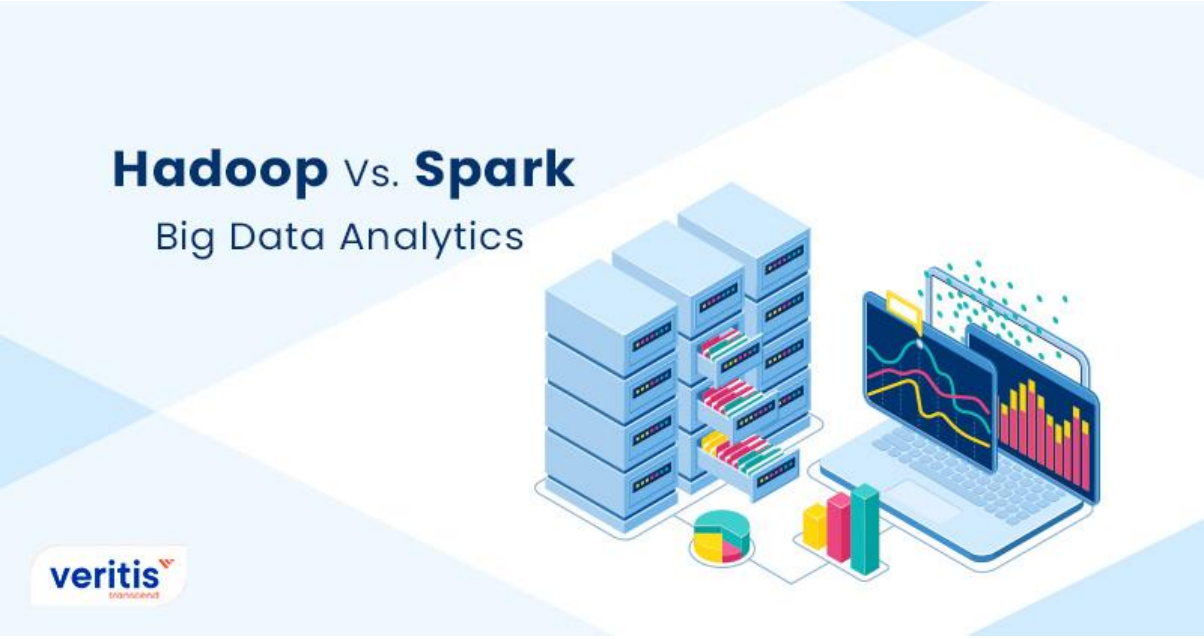
Hadoop vs Spark All You Need to Know About Big Data Analytics



Apache Hadoop and Apache Spark are dominant technologies used in big data processing frameworks for big data architectures. Both are at the epicenter of a rich ecosystem of open-source platforms that handle, manage, and analyze massive data collections. There is always a doubt in the organization's mind about which technology to opt for – Hadoop or Spark?

To add to the confusion, both these technologies frequently collaborate and handle data that is stored in the **Hadoop Distributed File System (HDFS)**. But each is a different and separate entity with its benefits and drawbacks as well as unique business applications. As a result, businesses often assess both of them for potential use in applications.

Most opinions revolve around optimizing how to use large data environments for batch processing or real-time processing center on **Hadoop and Spark**. But that oversimplifies the variations between the two frameworks. At the same time, Hadoop and some of its

components can now use for workloads involving interactive querying and real-time analytics.

There is a wide range of solutions for big data processing in the current era. Additionally, a lot of businesses provide specific enterprise features to go along with open-source platforms. Many companies run both applications for large data use cases. Initially, Hadoop was only fit for batch applications. In contrast, Spark was initially created to perform batch operations faster than Hadoop.

Additionally, Spark applications are frequently constructed on top of the HDFS and the **YARN** resource management technologies. **HDFS** is one of the leading data storage choices for Spark, which lacks a file system or repository of its own. Before delving deep into the comparison between **Hadoop and Spark**, let's know in detail about Apache Hadoop and Apache Spark.

What is Apache Hadoop?

The term Hadoop was first coined by Mike Cafarella and Doug Cutting in 2006, and they started it to process a massive amount of data. Hadoop began as a Yahoo initiative and later became a top-level Apache open-source project. The acronym stands for High Availability Distributed Object-Oriented Platform. And that's what the Hadoop technology offers developers – high availability through the simultaneous distribution of object-oriented tasks.

Apache Hadoop is an open-source platform that stores and processes a vast amount of data applications. It offers highly reliable, scalable, and distributed processing of big data applications.

This Java-based software can scale from a single server to thousands of devices, each providing storage and local computing. It can offer the building blocks on which can develop different applications and services. Hadoop is developed on clusters of commodity computers, offering a cost-effective solution for storing and processing a large volume of organized, semi-structured, and unstructured data with no format restrictions. **Hadoop** is primarily built in Java and supports numerous languages such as Perl, Ruby, Python, PHP, R, C++, and Groovy.

Useful link: ITIL vs DevOps: Can Both Concepts Work Together?

Apache Hadoop involves four main modules, and they are:

Headquarters: Veritis Group, Inc , 1231 Greenway Drive, Suite 1040, Irving, TX 75038

Phone: 972-753-0022 | **Email:** connect@veritis.com

1) HDFS

Hadoop Distributed File System (HDFS) controls how big data sets are stored within a Hadoop cluster. It can even generate both structured and unstructured data. In addition, it offers high fault tolerance and high throughput data access.

2) YARN

YARN stands for Yet Another Resource Negotiator. YARN is Hadoop's cluster resource manager that schedules tasks and distributes resources (such as CPU and memory) to applications using a cluster resource manager.

3) Hadoop MapReduce

Hadoop MapReduce divides large data processing projects into smaller ones, distributes the smaller tasks over various nodes, and then executes each task individually.

4) Hadoop Common (Hadoop Core)

Hadoop commonly refers to a group of standard tools and libraries that guide support to other modules, such as Apache Hadoop Framework, HDFS, YARN, and Hadoop MapReduce. Hadoop Core is often referred to as Hadoop Common.

Useful link: Understanding the Shift Left DevOps Approach

Benefits of Hadoop

Data define how businesses can improve their operations. Many industries revolve around data, and a lot of data is collected and analyzed through multiple methods and technologies. Hadoop is one of the popular tools to extract information from data, and it has its advantages in dealing with big data. Let's have a look at the most common benefits of Hadoop.

Cost

This technology is very economical, and anyone can access its source code. It can modify source code as per business needs. Hadoop offers cost-effective commodity hardware to create a cost-efficient model. Unlike RDBMS, which requires costly hardware and high-end processors to handle extensive data. The issue with RDBMS is that storing extensive data is not cost-effective. As a result, the organization has begun to delete the raw data.

Scalable

Hadoop is a highly scalable tool that stores vast data from a single server to thousands of machines. Without any downtime, users can expand the cluster's size by adding new nodes per requirement. Unlike RDBMS, which can't scale to handle the massive amount of data. Hadoop has no limit restrictions on the storage system.

Speed

Hadoop operates HDFS to handle its storage that maps data to any location on a cluster. When handling a massive amount of unstructured data, speed is a crucial factor. With Hadoop, it is possible to access terabytes of data in minutes and petabytes in hours.

Flexible

Hadoop is designed to access different datasets, such as structured, semi-structured, and unstructured data, to generate value from those datasets. This means enterprises can use Hadoop software to extract business insights from data sources such as email and social media conversations.

Availability

The nature of Hadoop makes it available to everyone who requires it. The enormous open-source community cleared the way for big data processing to be accessible.

Low Network Traffic

This application divides each task into multiple smaller sub-tasks in the Hadoop cluster, which are then assigned to each available data node. Each data node processes a little bit of data, leading to minimum traffic in a Hadoop cluster.

Useful link: Understanding the Differences Between Deep Learning and Machine Learning

---

What is Apache Spark?

Apache Spark is an open-source platform for data processing framework that can quickly execute data science, data engineering, and machine learning operations on single-node clusters. The Apache Software Foundation released Spark software to speed up the Hadoop computational computing software process. Spark uses Hadoop for processing and storage. Since Spark manages clusters independently, Spark uses Hadoop for only storage purposes.

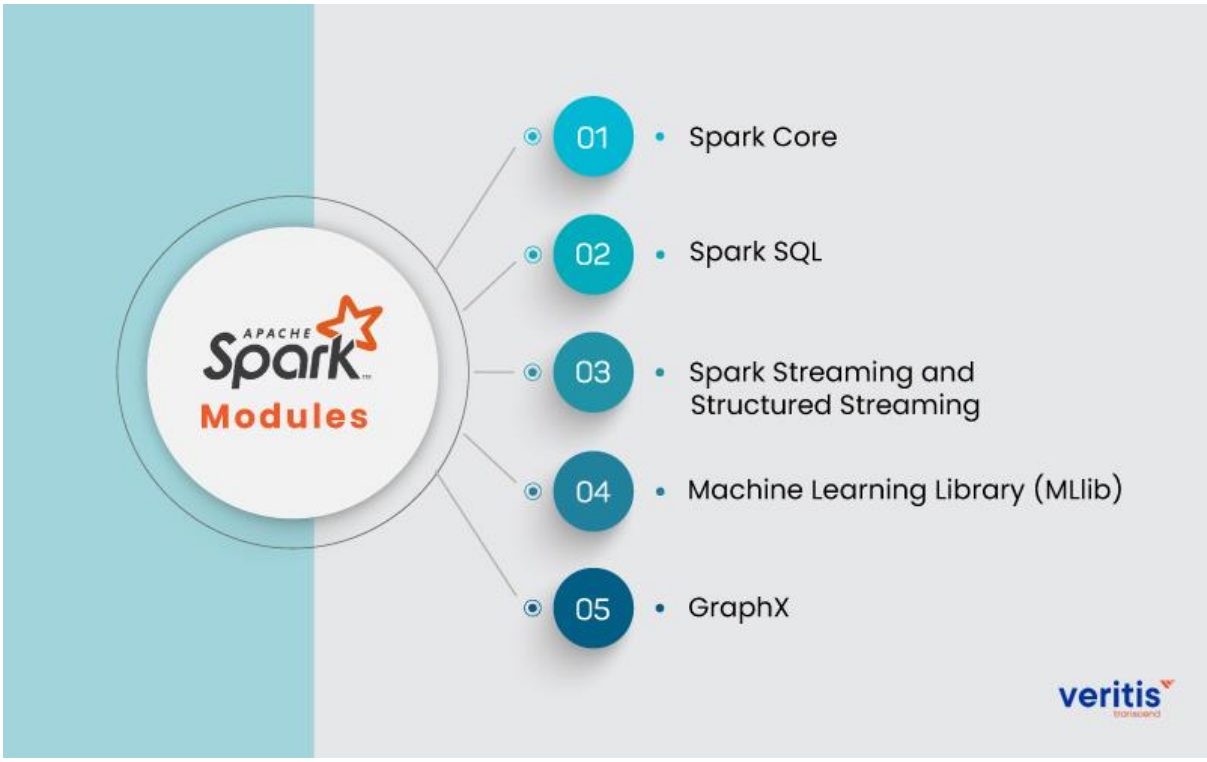
Apache Spark supports numerous programming languages, such as Java, R, Scala, and Python. It includes libraries for a wide range of tasks such as SQL, machine learning, and streaming as well as it can be used anywhere from a laptop to a cluster of hundreds of servers. However, it typically runs quicker than Hadoop and processes data

using **random access memory (RAM)** rather than a file system. Moreover, Spark can now handle use cases that Hadoop can't perform.

Apache Spark is the only processing framework that involves **artificial intelligence (AI)** and data. It is the most significant open-source project in the data processing. This allows users to execute cutting-edge **machine learning (ML)** and **artificial intelligence (AI)** algorithms after performing extensive data transformations and analysis.

Useful link: Comparison on AWS Vs Azure Vs GCP

Apache Spark involves five main modules, and they are:



1) Spark Core

Spark Core underlays an execution engine that coordinates input and output (I/O) activities, schedules, and dispatches tasks.

Headquarters: Veritis Group, Inc , 1231 Greenway Drive, Suite 1040, Irving, TX 75038

Phone: 972-753-0022 | **Email:** connect@veritis.com

2) Spark SQL

Spark SQL collects structured data information so users can permit to improve structured data processing.

3) Spark Streaming and Structured Streaming

Spark Streaming and Structured Streaming can increase the capacity for stream processing. Spark Streaming gathers information from several streaming sources and splits it into micro-batches for a continuous stream. Structured Streaming developed on Spark SQL decreases latency and makes programming easy.

4) Machine Learning Library (MLlib)

A group of scalable machine learning algorithms as well as tools for choosing features and constructing ML pipelines. The main API for MLlib is data frames, which offers consistency across numerous programming languages such as Python, Scala, and Java.

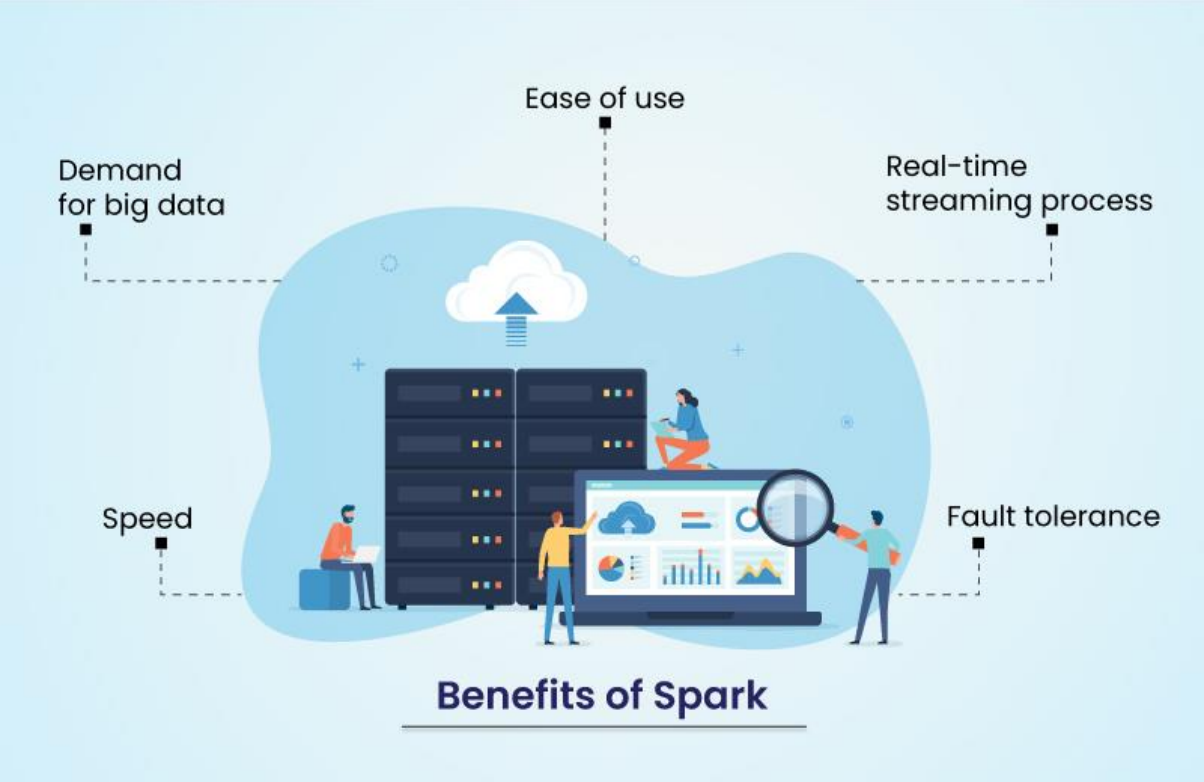
5) GraphX

GraphX is a user-friendly computation engine that allows the interactive construction, editing, and analysis of graph-structured data and is scalable.

Useful link: Kubernetes Adoption: The Prime Drivers and Challenges

---

Benefits of Spark



Apache Spark can advance big data-related business across industries. Apache Spark has numerous benefits for dealing with big data, and let's have a look at the most common benefits of Spark.

Speed

Processing speed is always vital for big data. Because of its speed, Apache Spark is incredibly popular among data scientists. Spark is 100 times quicker than Hadoop for processing massive amounts of data. It runs in memory (RAM) computing system, while Hadoop runs local memory space to store data. Spark can process clustered data with more than 8000 nodes and many petabytes at once.

Headquarters: Veritis Group, Inc , 1231 Greenway Drive, Suite 1040, Irving, TX 75038

Phone: 972-753-0022 | **Email:** connect@veritis.com

Ease of use

This open-source application provides easy-to-use APIs for working with big data sets. It provides 80 high-level operators that make it simple to create similar apps. We can reprocess the Spark code for joining streams with historical data, operate ad hoc stream state queries, and do batch processing.

Demand for Big Data

A recent survey by IBM announced that it would train more than 1 million data scientists and engineers in **Apache Spark**. This is because it offers numerous opportunities for big data and has so much demand for developers.

Fault Tolerance

Through Spark abstraction-RDD, Apache Spark offers fault tolerance. Apache Spark RDDs are created to manage the failure of any cluster worker node. As a result, it ensures that data loss decreases to zero.

Real-time Streaming Process

Apache Spark includes a feature for real-time streaming processes. The issue with Hadoop MapReduce is that it can only manage existing data but not real-time data. However, we can resolve this issue with Spark Streaming.

Useful link: Hadoop Vs Kubernetes: Is K8s invading Hadoop Turf?

Comparison between Apache Hadoop and Apache Spark

Let’s look at the different parameters between Apache Hadoop and Apache Spark.

Parameters	Apache Hadoop	Apache Spark
Cost	Hadoop runs at low cost	Spark runs at high cost

Performance	Hadoop is relatively slow because it stores data from numerous sources and uses MapReduce to process it in batches	Spark is faster as it uses RAM
Data processing	It is ideal for linear data and batch processing	It is ideal for live unstructured data streams processing and real-time processing
Security	It is more secure. Hadoop runs various access control and authentication methods	It is less secure. Spark improves security with shared secret authentication or event logging
Efficiency	It is built to manage batch processing efficiently	It is built to manage real-time data efficiently
Fault tolerance	It is a highly fault tolerance system. It uses the data that is replicated among the nodes in the event of a problem.	When a partition fails, it can recreate a dataset by tracking the construction of RDD blocks. In order to reconstruct data across nodes, Spark can also use a DAG.
Scalability	It is simple to scale by adding nodes and disks for storage	It is hard enough to scale because it depends on Ram for computations

Supports programming languages	Java, Perl, Ruby, Python, PHP, R, C++, and Groovy	Java, R, Scala, and Python
Machine Learning	It is relatively slow	It is faster with in-memory processing
Category	It is the data processing engine	It is the data analytics engine
Latency	It has high latency computing	It has low latency computing
Scheduler	It requires an external job scheduler	It doesn’t require an external scheduler
Open source	Yes	Yes
Data integration	Yes	Yes
Speed	Low performance	High performance (100x faster)
Developer community support	Yes	Yes
Memory consumption	It depends on disk	It depends on RAM

Final Thoughts on Hadoop and Spark

Hadoop is excellent for processing multiple sets of massive amounts of data in parallel. Apache Hadoop can store unlimited amounts of data in its cluster. It involves analytical tools such as HBase, MongoDB, Apache Mahout, Pentaho, and R Python.

Spark is suitable for analyzing real-time data from multiple sources such as sensors, the **Internet of Things (IoT)**, and financial systems. In addition, analytics can be utilized to target particular groups for machine learning and media campaigns. Without modifying code, Spark has been tested to be 100 times quicker than Hadoop Hive.

Both Apache Hadoop and Apache Spark have prominent features in the area of analytics and big data processing. With 2,000 developers from 20,000 organizations, including 80% of the Fortune 500, Apache Spark has a thriving and active community.

At the same time, Hadoop technology is implementing in multiple industries such as healthcare, education, government, banking, communication, and entertainment. As a result, there is a clear enough for both to grow and numerous use cases for each of these open-source technologies.

However, adopting both Hadoop and Spark technologies is a laborious process; this is why where companies seek Vertis's services. **Veritis, the Stevie Awards winner**, is an IT consulting services provider that has been partnering with small to large companies, including Fortune 500 firms, for over a decade. We offer the best solutions for customers with world-class experiences and cost-effective solutions.

Services

Headquarters: Veritis Group, Inc , 1231 Greenway Drive, Suite 1040, Irving, TX 75038

Phone: 972-753-0022 | **Email:** connect@veritis.com