MongoDB.

Hadoop or MongoDB: What is the Difference?

Hadoop and MongoDB are great solutions to work with big data. However, they each have their forces and weaknesses.

Try MongoDB Atlas Free

MongoDB is a complete data platform that brings you more capabilities than Hadoop. However, when dealing with objects that are petabytes in size, Hadoop offers some interesting data processing capabilities.

This article will highlight the differences between Hadoop and MongoDB and explain when to use them.

What is MongoDB?

MongoDB was founded in 2007 as a NoSQL (Not Only SQL) database written in C++ to handle humongous amounts of data. Since then, it's evolved into a complete cloud-based data platform. In addition to the database that can be deployed on any major cloud provider, MongoDB Atlas also offers full-text search, advanced analytics tools, an easy-to-use query language, and much more.

What is Hadoop?

Hadoop is an open-source software suite mainly written in Java used to handle large data and computation across networks of computers. It was first released in 2006 as a

subproject of Nutch and is now part of the Apache Foundation. It is based on the MapReduce programming model first introduced in 2004 by Google.

Why should you use MongoDB?

MongoDB is a general-purpose document database that uses a flexible schema to store data. It is easy to use for software developers, and native drivers make it simple to integrate with any programming language. Because of its flexible schema, the database makes it frictionless to evolve as the application grows. It also has built-in scaling capabilities for both horizontal and vertical scaling.

When very large data sets are needed for long-running analytics, MongoDB also offers Atlas Data Federation. Atlas Data Federation can be used to query cold data, which is stored in a cost-effective manner, and hot data hosted in a MongoDB cluster using the same MongoDB Query API.

MongoDB Atlas, the cloud offering by MongoDB, provides all the tools needed for a software developer to integrate a database into their application quickly.

Why should you use Hadoop?

Hadoop is a collection of tools to process big data. Its primary use case is for the parallel processing of massive data sets. It is handy for processing unstructured data stored in data lakes. The main Hadoop modules are HDFS–the distributed file system for storage, the MapReduce module for data processing, YARN for resource management and task scheduling, and Hadoop Common for all the tooling to manage the various modules.

How MongoDB and Hadoop handle real-time data processing

When it comes to real-time data processing, MongoDB is a clear winner. While Hadoop is great at storing and processing large amounts of data, it does its processing in batches. A possible way to make this data processing faster is by using Spark. With this framework, data processing can happen in memory, increasing the speed at which data is processed.

On the other hand, MongoDB offers many built-in tools for real-time data processing. With connectors for external tools such as Kafka and Spark, MongoDB makes it easy to ingest and process data as it comes in from external sources.

MongoDB and Hadoop over relational database management systems (RDBMS)

Both MongoDB and Hadoop cover use cases that traditional relational databases simply can't help with.

Data storage

Both MongoDB and Hadoop can store unstructured data, while traditional relational databases require you to structure and normalize data that you will store. This flexibility offered by MongoDB and Hadoop is a great advantage when your data sources can take multiple forms over time.

Thanks to the distributed file system in Hadoop, it is simple to add more nodes to a cluster to increase the available storage capacity. MongoDB uses sharding to distribute data across multiple nodes to help scale horizontally.

With traditional RDBMS, scaling needs to be done by adding more disk space, which might require downtime and usually leads to higher costs.

Data processing

Traditional RDBMSs offer limited capabilities to process the data stored in the database. Data processing is typically done through another application that queries the database and then processes it.

MongoDB offers the aggregation pipeline framework that allows you to process and manipulate data to retrieve relevant information from the database. In addition to the aggregation pipelines, you can also use Atlas Search to further refine your search requests with full-text search capabilities.

Hadoop is great at processing large batches of data using the MapReduce paradigm. This algorithm is good when individual pieces of data are processed one at a time. However, when variables need to be correlated, the MapReduce algorithm might make things slower than necessary.

Memory handling

MongoDB optimizes its memory usage to enable quick delivery of data. It keeps indexes and some data in memory to ensure predictable latency.

Hadoop, on the other hand, focuses on disk storage. It is better at optimizing disk space but will be slower to deliver query results since it needs to read from the disk.

Traditional relational databases will use a mix of both disk and memory to deliver the results as fast as possible. However, because of the need for joins, a lot of memory usage is dedicated to joining tables and might cause latency.

Advantages of using MongoDB

There are many advantages of using MongoDB.

- It's a general-purpose database. MongoDB can query data lakes and unstructured data, just like Hadoop. In addition to that, it has a rich query language and can replace any RDBMS, while Hadoop needs to be used in conjunction with a database. MongoDB also supports more use cases natively, including the processing and storage of time series and geo data.
- It's fast. Thanks to the aggregation pipeline framework, MongoDB can process and return results much quicker than Hadoop could.
- It's scalable. MongoDB can easily be scaled vertically and horizontally, as needed. This scalability ensures that you can optimize the costs of using MongoDB for your current needs. With the new serverless databases in Atlas, you can ensure that you only pay for what you use.
- It's easy to maintain. MongoDB Atlas, the cloud data platform by MongoDB, is a fully managed database that can accommodate all of your data needs. No need for expensive configuration, updates, or maintenance.
- It's easy to use. Data stored in S3 buckets with data lakes can still be accessed and queried using the MongoDB Query API, just like any other data in MongoDB.

Advantages of using Hadoop

Hadoop shines when it comes to enormous data sets (petabytes).

- It handles unstructured data. In Hadoop, you can store any data type, including binary files. No data transformation is required before storage.
- It has a distributed file system. One of the significant advantages of Hadoop is its distributed file system. This makes it easy to add more nodes to a cluster. This also provides Hadoop with the necessary computing power for fast data computation.

- It's cost-effective. Hadoop runs on commodity servers, making it cheap to add more nodes when needed.

Why you should use MongoDB

Both Hadoop and MongoDB have many benefits over traditional databases when it comes to handling big data. However, only MongoDB can act as a complete replacement for a traditional database. With its flexible schema, MongoDB makes it easy to store information in a format that doesn't require many transformations ahead of time. Its query language makes it possible to efficiently access and even process data on the fly.

There are still times when you might want to use Hadoop, though. With its distributed file system, Hadoop can come in handy when dealing with massive objects. In these cases, it is possible to use Hadoop to complement MongoDB to leverage the power of both into a single cohesive architecture.

FAQ

Is MongoDB faster than Hadoop?

When it comes to real-time data processing, MongoDB is much faster than Hadoop. The latter will work in batches and doesn't optimize its memory management.

Is MongoDB used for big data?

MongoDB is a general-purpose database that can also process big data. Its flexible schema enables it to accept any form or volume of data.

Can MongoDB replace Hadoop?

In many cases, MongoDB can replace Hadoop for storage of big data and
data processing. However, there are use cases where you might still need
Hadoop. In those cases, you can use Hadoop to complement MongoDB.

✕

Does MongoDB use Hadoop?

No. MongoDB and Hadoop are two separate products.

✕

Is Hadoop good for big data?

Hadoop is a collection of software used to store and process big data. It
works well for unstructured data.

✕

Which is better, Hadoop or MongoDB?

Hadoop and MongoDB both have specific use cases. However, in many
scenarios, you can use MongoDB to replace Hadoop. For the occasional use
cases where Hadoop is needed, you can use it with MongoDB. For this
reason, MongoDB is more flexible than Hadoop.

✕

🍃 MongoDB®

🌐 English ⌄

About

Careers						Investor Relations

Legal Notices

Privacy Notices

Security Information

Trust Center

Support

Contact Us

Customer Portal

Atlas Status

Customer Support

Manage Cookies

Social

GitHub

Stack Overflow

LinkedIn

YouTube

X

Twitch

Facebook

© 2024 MongoDB, Inc.