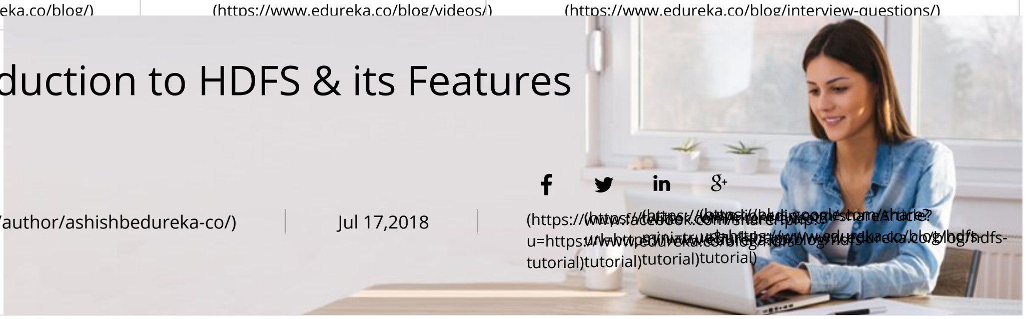


HDFS Tutorial: Introduction to HDFS & its Features

Recommended by 622 users

Ashish Bakshi (https://www.edureka.co/blog/author/ashishbedureka-co/) | Jul 17,2018

(https://www.facebook.com/sharer.php?u=https://www.edureka.co/blog/hdfs-tutorial)(https://twitter.com/share?url=https://www.edureka.co/blog/hdfs-tutorial)(https://www.linkedin.com/shareArticle?mini=true&url=https://www.edureka.co/blog/hdfs-tutorial)(https://plus.google.com/share?url=https://www.edureka.co/blog/hdfs-tutorial)

Add to Bookmark (https://www.edureka.co/blog/hdfs-tutorial) Email this Post 35.2K (https://www.edureka.co/blog/hdfs-tutorial#comments-wrapper) 13 (https://www.edureka.co/blog/hdfs-tutorial#disqus_thread)

HDFS Tutorial

Before moving ahead in this HDFS tutorial blog, let me take you through some of the insane statistics related to HDFS:

- In 2010, **Facebook** claimed to have one of the largest HDFS cluster storing **21 Petabytes** of data.
- In 2012, **Facebook** declared that they have the largest single HDFS cluster with more than **100 PB** of data**.**
- And **Yahoo**! has more than **100,000 CPU** in over **40,000 servers** running Hadoop, with its biggest Hadoop cluster running **4,500 nodes**. All told, Yahoo! stores **455 petabytes** of data in HDFS.
- In fact, by 2013, most of the big names in the Fortune 50 started using Hadoop.

Too hard to digest? Right. As discussed in ***Hadoop Tutorial (https://www.edureka.co/blog/hadoop-tutorial/)***, Hadoop has two fundamental units – ***Storage*** and ***Processing***. When I say storage part of Hadoop, I am referring to **HDFS** which stands for **Hadoop Distributed File System**. So, in this blog, I will be introducing you to **HDFS**.

Here, I will be talking about:

- What is HDFS?
- Advantages of HDFS
- Features of HDFS

Before talking about HDFS, let me tell you, what is a Distributed File System?

DFS or Distributed File System:

Distributed File System talks about **managing data**, i.e. **files or folders across multiple computers or servers.** In other words, DFS is a file system that allows us to store data over multiple nodes or machines in a cluster and allows multiple users to access data. So basically, it serves the same purpose as the file system which is available in your machine, like for windows you have NTFS (New Technology File System) or for Mac you have HFS (Hierarchical File System). The only difference is that, in case of Distributed File System, you store data in multiple machines rather than single machine. Even though the files are stored across the network, DFS organizes, and displays data in such a manner that a user sitting on a machine will feel like all the data is stored in that very machine.

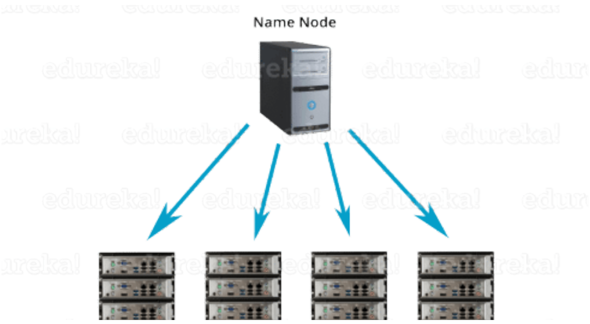
What is HDFS?

Hadoop Distributed file system or HDFS is a Java based distributed file system that allows you to store large data across multiple nodes in a Hadoop cluster. So, if you install Hadoop, you get HDFS as an underlying storage system for storing the data in the distributed environment.

Let's take an example to understand it. Imagine that you have ten machines or ten computers with a hard drive of 1 TB on each machine. Now, HDFS says that if you install Hadoop as a platform on top of these ten machines, you will get HDFS as a storage service. Hadoop Distributed File System is distributed in such a way that every machine contributes their individual storage for storing any kind of data.

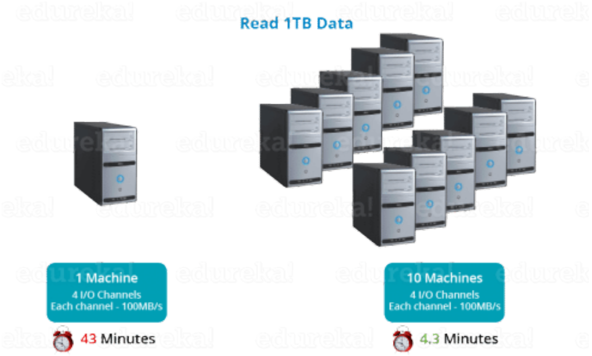
HDFS Tutorial: Advantages Of HDFS

1. Distributed Storage:



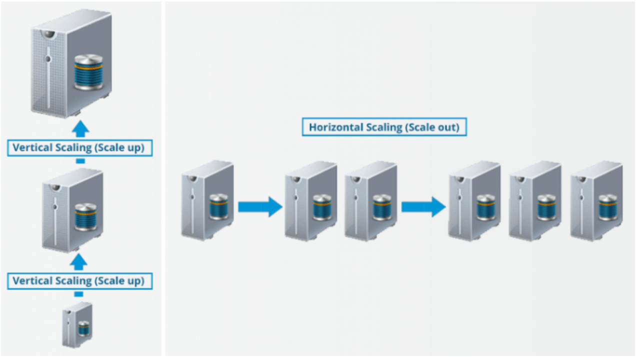
When you access Hadoop Distributed file system from any of the ten machines in the Hadoop cluster, you will feel as if you have logged into a single large machine which has a storage capacity of 10 TB (total storage over ten machines). What does it mean? It means that you can store a single large file of 10 TB which will be distributed over the ten machines (1 TB each). So, it is **not limited to the physical boundaries** of each individual machine.

2. Distributed & Parallel Computation:



Because the data is divided across the machines, it allows us to take advantage of **Distributed and Parallel Computation**. Let’s understand this concept by the above example. Suppose, it takes 43 minutes to process 1 TB file on a single machine. So, now tell me, how much time will it take to process the same 1 TB file when you have 10 machines in a Hadoop cluster with similar configuration – 43 minutes or 4.3 minutes? 4.3 minutes, Right! What happened here? Each of the nodes is working with a part of the 1 TB file in parallel. Therefore, the work which was taking 43 minutes before, gets finished in just 4.3 minutes now as the work got divided over ten machines.

3. Horizontal Scalability:


Vertical Scaling (Scale up)
Vertical Scaling (Scale up)
Horizontal Scaling (Scale out)

Last but not the least, let us talk about the **horizontal scaling** or **scaling out** in Hadoop. There are two types of scaling: **vertical** and **horizontal**. In vertical scaling (scale up), you increase the hardware capacity of your system. In other words, you procure more RAM or CPU and add it to your existing system to make it more robust and powerful. But there are challenges associated with vertical scaling or scaling up:

- There is always a limit to which you can increase your hardware capacity. So, you can’t keep on increasing the RAM or CPU of the machine.
- In vertical scaling, you stop your machine first. Then you increase the RAM or CPU to make it a more robust hardware stack. After you have increased your hardware capacity, you restart the machine. This down time when you are stopping your system becomes a challenge.

In case of **horizontal scaling (scale out)**, you add more nodes to existing cluster instead of increasing the hardware capacity of individual machines. And most importantly, you can **add more machines on the go** i.e. Without stopping the system**.** Therefore, while scaling out we don’t have any down time or green zone, nothing of such sort. At the end of the day, you will have more machines working in parallel to meet your requirements.

HDFS Tutorial Video:

You may check out the video given below where all the concepts related to HDFS has been discussed in detail:

Get Started with Hadoop

(https://www.edureka.co/big-data-and-hadoop)

HDFS Tutorial: Features of HDFS

We will understand these features in detail when we will explore the HDFS Architecture in our next HDFS tutorial blog. But, for now, let's have an overview on the features of HDFS:

- ***Cost:*** The HDFS, in general, is deployed on a commodity hardware like your desktop/laptop which you use every day. So, it is very economical in terms of the cost of ownership of the project. Since, we are using low cost commodity hardware, you don't need to spend huge amount of money for scaling out your Hadoop cluster. In other words, adding more nodes to your HDFS is cost effective.
- ***Variety and Volume of Data:*** When we talk about HDFS then we talk about storing huge data i.e. Terabytes & petabytes of data and different kinds of data. So, you can store any type of data into HDFS, be it structured, unstructured or semi structured.
- ***Reliability and Fault Tolerance:*** When you store data on HDFS, it internally divides the given data into data blocks and stores it in a distributed fashion across your Hadoop cluster. The information regarding which data block is located on which of the data nodes is recorded in the metadata. **NameNode** manages the meta data and the **DataNodes** are responsible for storing the data.
  Name node also replicates the data i.e. maintains multiple copies of the data. This replication of the data makes HDFS very reliable and fault tolerant. So, even if any of the nodes fails, we can retrieve the data from the replicas residing on other data nodes. By default, the replication factor is 3. Therefore, if you store 1 GB of file in HDFS, it will finally occupy 3 GB of space. The name node periodically updates the metadata and maintains the replication factor consistent.
- ***Data Integrity:*** Data Integrity talks about whether the data stored in my HDFS are correct or not. HDFS constantly checks the integrity of data stored against its checksum. If it finds any fault, it reports to the name node about it. Then, the name node creates additional new replicas and therefore deletes the corrupted copies.
- ***High Throughput:*** Throughput is the amount of work done in a unit time. It talks about how fast you can access the data from the file system. Basically, it gives you an insight about the system performance. As you have seen in the above example where we used ten machines collectively to enhance computation. There we were able to reduce the processing time from **43 minutes** to a mere **4.3 minutes** as all the machines were working in parallel. Therefore, by processing data in parallel, we decreased the processing time tremendously and thus, achieved high throughput.
- ***Data Locality:*** Data locality talks about moving processing unit to data rather than the data to processing unit. In our traditional system, we used to bring the data to the application layer and then process it. But now, because of the architecture and huge volume of the data, bringing the data to the application layer will reduce the network performance to a noticeable extent. So, in HDFS, we bring the computation part to the data nodes where the data is residing. Hence, you are not moving the data, you are bringing the program or processing part to the data.

So now, you have a brief idea about HDFS and its features. But trust me guys, this is just the tip of the iceberg. In my next ***HDFS tutorial blog (https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/)***, I will deep dive into the *HDFS architecture* and I will be unveiling the secrets behind the success of HDFS. Together we will be answering all those questions which are pondering in your head such as:

- What happens behind the scenes when you read or write data in Hadoop Distributed File System?
- What are the algorithms like rack awareness that makes HDFS so fault tolerant?
- How Hadoop Distributed File System manages and creates replica?
- What are block operations?

<< Previous Blog (https://www.edureka.co/blog/hadoop-ecosystem)

Next Blog >> (https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/)

*Now that you have understood HDFS and its features, check out the **Hadoop training (https://www.edureka.co/big-data-and-hadoop/)** by Edureka, a trusted online learning company with a network of more than 250,000 satisfied learners spread across the globe. The Edureka Big Data Hadoop Certification Training course helps learners become expert in HDFS, Yarn, MapReduce, Pig, Hive, HBase, Oozie, Flume and Sqoop using real-time use cases on Retail, Social Media, Aviation, Tourism, Finance domain.*

Got a question for us? Please mention it in the comments section and we will get back to you.

About Ashish Bakshi (11 Posts (https://www.edureka.co/blog/author/ashishbedureka-co/))

Share on (https://www.facebook.com/sharer.php?u=https://www.edureka.co/blog/hdfs-tutorial) (https://twitter.com/share?url=https://www.edureka.co/blog/hdfs-tutorial) (https://www.linkedin.com/shareArticle?mini=true&url=https://www.edureka.co/blog/hdfs-tutorial) (https://plus.google.com/share?url=https://www.edureka.co/blog/hdfs-tutorial)

PREVIOUS NEXT

Enter your Email Address

SUBSCRIBE

Related Posts

Apache Hadoop HDFS Architecture
👁 61.6K
(https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/)

Big Data Tutorial: All You Need To Know About Big Data!
👁 37.7K
(https://www.edureka.co/blog/big-data-tutorial)

Hadoop Tutorial: All you need to know about Hadoop!
👁 87.9K
(https://www.edureka.co/blog/hadoop-tutorial/)

Hadoop Ecosystem: Hadoop Tools for Crunching Big Data
👁 44.8K
(https://www.edureka.co/blog/hadoop-ecosystem)

Comments

13 Comments

13 Comments https://www.edureka.co/blog/ Rajiv Chaudhuri

♡ Recommend 3 Share Sort by Best

Join the discussion…

Neelesh • a year ago
Nice blog
2 ^ | v • Reply • Share ›

EurekaSupport Mod ➦ Neelesh • a year ago
Hey Neelesh, thanks for checking out our blog. We're glad you found it useful.
You might also like our tutorials here: https://www.youtube.com/edu...
Do subscribe to stay posted on upcoming blogs and videos. Cheers!
^ | v • Reply • Share ›

Abhishek Agarwal • 2 years ago
This blog gives a nice and precise introduction to HDFS.
1 ^ | v • Reply • Share ›

EdurekaSupport Mod ➦ Abhishek Agarwal • 2 years ago
Hey Abhishek, thanks for checking out the blog! We're glad you liked it. Do spread the word and subscribe to stay posted on upcoming blogs. Cheers!
^ | v • Reply • Share ›

Rishav Kumar • 2 years ago
Nice pick up for beginners
1 ^ | v • Reply • Share ›

EdurekaSupport Mod ➦ Rishav Kumar • 2 years ago
We're glad you found the blog useful, Rishav! Do subscribe to our blog to stay posted on upcoming blogs. Cheers!
^ | v • Reply • Share ›

SacTiw • 8 months ago
Is it good with storage of binary data as well e.g. large executable, compressed files, VDI etc?
By good I meant performance wise w.r.t get/put operations, block storage, replication etc?
^ | v • Reply • Share ›

Rahul Joshi • a year ago
It has Awesome Explaination in a precise way. Pls do suggest How to start Big data Thing .
Thank You so Much .. :)
^ | v • Reply • Share ›

EdurekaSupport Mod ➦ Rahul Joshi • 7 months ago
Hey **@Rahul Joshi**, check out our Big Data course here: https://www.edureka.co/big-...
This course will help you truly master the Big Data Hadoop technology in and out. Do check it out. Cheers :)
^ | v • Reply • Share ›

malli arjun • a year ago
Edureka..! is giving such a visualization where reading article in hadoop makes u like talking with different persons as a part of concept ..finally it makes u to feel as BIG

DATA NOT A BIG DEAL.
Thanks a lot .

• Reply • Share ›

EdurekaSupport Mod ➦ malli arjun • a year ago

Hey Malli Arjun, thanks for checking out our blog. We're glad you found it useful.
You might also like our tutorials here: https://www.youtube.com/edu.... You can also check out our complete training here: https://www.edureka.co/big-....
Do subscribe to stay posted on upcoming blogs and videos. Cheers!

• Reply • Share ›

Velimir Milovanoski • 2 years ago

Every part is explained well, but for better following of your speech should to be attached an English subtitle in blog's YT video.

• Reply • Share ›

EdurekaSupport Mod ➦ Velimir Milovanoski • 2 years ago

Hey Velimir, thanks for checking out our blog. We're glad you found it useful. We have noted your feedback and passed it on to the relevant team. In fact, we're working on including sub titles in videos. Do follow our blog to stay updated on upcoming posts. Cheers!

• Reply • Share ›

ALSO ON HTTPS://WWW.EDUREKA.CO/BLOG/

Git vs Github – Demystifying The Differences
2 comments • 8 months ago
Donald Peoples — this was very helpful! Thanks.

Top 55 Blockchain Interview Questions You Must Prepare In 2018
4 comments • 5 months ago
Maximilian Fischer — A blockchain (database) only contains a record of all changes that were ever made to it. Value is ascribed to the fact of actually being able to make such changes. Value in …

Top 10 Trending Technologies To Master In 2018
2 comments • 8 months ago
Mohammed Innat — Cool

What is Robotic Process Automation? – An Introduction to RPA
2 comments • 5 months ago
Lucas Ramalho Salata — Hello, you. Very cool!Could you please correct the text below "COMPLIANCE"? It's wrong, I guess. It's the same text as "PRODUCTIVITY" (I'm talking about the "Benefits of …

✉ Subscribe Ⓓ Add Disqus to your siteAdd DisqusAdd 🔒 Disqus' Privacy PolicyPrivacy PolicyPrivacy

Subscribe
to our newsletter

Enter your Email Address

SUBSCRIBE

Related Blogs

(https://www.
hadoop-hdfs-
architecture/)

Apache Hadoop HDFS Architecture (https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/)

(https://www.
data-tutorial)

Big Data Tutorial: All You Need To Know About Big... (https://www.edureka.co/blog/big-data-tutorial)

(https://www.
tutorial/)

Hadoop Tutorial: All you need to know about Hadoop! (https://www.edureka.co/blog/hadoop-tutorial/)

(https://www.
questions/hac
interview-
questions-
hdfs-2/)

Top Hadoop Interview Questions To Prepare In 2018 – HDFS (https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-hdfs-2/)

Edureka

About us (https://www.edureka.co/about-us)

News & Media (https://www.edureka.co/allmedia)

Contact us (https://www.edureka.co/contact-us)

Careers (https://www.edureka.co/careers)

Blog (https://www.edureka.co/blog/all)

Reviews (https://www.edureka.co/reviews)

Terms & conditions (https://www.edureka.co/terms-and-conditions)

Privacy policy (https://www.edureka.co/privacy-policy)

Sitemap (https://www.edureka.co/sitemap)

Work with us

Become an Instructor (https://www.edureka.co/instructors/add)

Hire from Edureka (https://www.edureka.co/hire-from-edureka)

Follow us on

(https:[illegible]

Learn on the GO!

(https://itunes.apple.com/in/app/edureka/id1033145415?mt=8)

(https://play.google.com/store/apps/details?id=co.edureka.app)

edureka! (https://www.edureka.co)