Hadoop Tutorial: All you need to know about Hadoop!

Recommended by 673 users

Shubham Sinha (https://www.edureka.co/blog/author/shubham-sinha/) | Jul 23,2018 |

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

Add to Bookmark (https://www.edureka.co/blog/hadoop-tutorial/) Email this Post 87.9K (https://www.edureka.co/blog/hadoop-tutorial/#comments-wrapper) 6 (https://www.edureka.co/blog/hadoop-tutorial/#disqus_thread)

If you are looking to learn Hadoop, you have landed at the perfect place. In this Hadoop tutorial blog, you will learn from basic to advanced Hadoop concepts in very simple steps. Alternatively, you can also watch the below video from our Hadoop expert, discussing Hadoop concepts along with practical examples.

Hadoop Tutorial For Beginners | Hadoop Training | Edureka

In this Hadoop tutorial blog, we will be covering the following topics:

- How it all started
- What is Big Data
- Big Data and Hadoop: Restaurant Analogy
- What is Hadoop
- Hadoop-as-a-Solution
- Hadoop Features
- Hadoop Core Components
- Hadoop Last.fm Case Study

Hadoop Tutorial: How It All Started?

Before getting into technicalities in this Hadoop tutorial blog, let me begin with an interesting story on how Hadoop came into existence and why is it so popular in the industry nowadays. So, it all started with two people, Mike Cafarella and Doug Cutting, who were in the process of building a search engine system that can index 1 billion pages. After their research, they estimated that such a system will cost around half a million dollars in hardware, with a monthly running cost of $30,000, which is quite expensive. However, they soon realized that their architecture will not be capable enough to work around with billions of pages on the web.

They came across a paper, published in 2003, that described the architecture of Google's distributed file system, called GFS, which was being used in production at Google. Now, this paper on GFS proved to be something that they were looking for, and soon, they realized that it would solve all their problems of storing very large files that are generated as a part of the web crawl and indexing process. Later in 2004, Google published one more paper that introduced MapReduce to the world. Finally, these two papers led to the foundation of the framework called "***Hadoop***". Doug quoted on Google's contribution in the development of Hadoop framework:

"Google is living a few years in the future and sending the rest of us messages."

So, by now you would have realized how powerful Hadoop is. Now, before moving on to Hadoop, let us start the discussion with Big Data, that led to the development of Hadoop.

Hadoop Tutorial: What is Big Data?

Have you ever wondered how technologies evolve to fulfill emerging needs? For example, earlier we had landline phones, but now we have shifted to smartphones. Similarly, how many of you remember floppy drives that were extensively used back in 90's? These Floppy drives have been replaced by hard disks because these floppy

drives had very low storage capacity and transfer speed. Thus, this makes floppy drives insufficient for handling the amount of data with which we are dealing today. In fact, now we can store terabytes of data on the cloud without being bothered about size constraints.

Now, let us talk about various drivers that contribute to the generation of data.

Have you heard about ***IoT (https://www.edureka.co/blog/iot-tutorial/)***? IoT connects your physical device to the internet and makes it smarter. Nowadays, we have smart air conditioners, televisions etc. Your smart air conditioner constantly monitors your room temperature along with the outside temperature and accordingly decides what should be the temperature of the room. Now imagine how much data would be generated in a year by smart air conditioner installed in tens & thousands of houses. By this you can understand how ***IoT (https://www.edureka.co/blog/iot-tutorial/)***is contributing a major share to Big Data.

Now, let us talk about the largest contributor to the Big Data which is, nothing but, social media. Social media is one of the most important factors in the evolution of Big Data as it provides information about the people's behavior. You can look at the figure below and get an idea how much data is getting generated every minute:

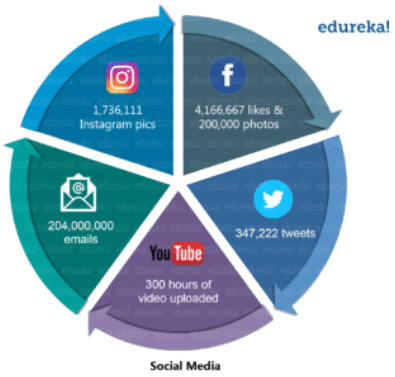

Fig: Hadoop Tutorial – Social Media Data Generation Stats

Apart from the rate at which the data is getting generated, the second factor is the lack of proper format or structure in these data sets that makes processing a challenge.

GET HADOOP CERTIFIED TODAY

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

Hadoop Tutorial: Big Data & Hadoop – Restaurant Analogy

Let us take an analogy of a restaurant to understand the problems associated with Big Data and how Hadoop solved that problem.

Bob is a businessman who has opened a small restaurant. Initially, in his restaurant, he used to receive two orders per hour and he had one chef with one food shelf in his restaurant which was sufficient enough to handle all the orders.

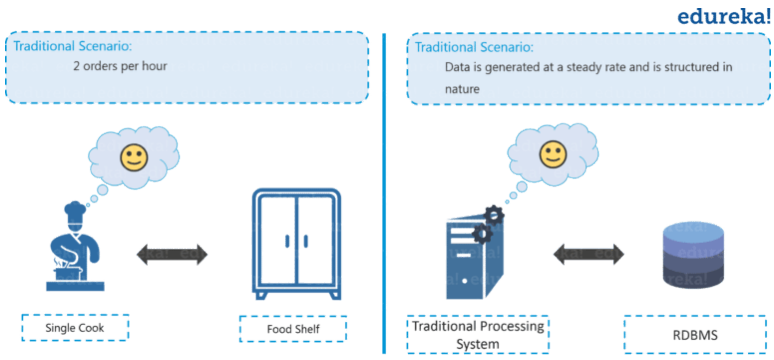

Fig: Hadoop Tutorial – Traditional Restaurant Scenario

Now let us compare the restaurant example with the traditional scenario where data was getting generated at a steady rate and our traditional systems like RDBMS is capable enough to handle it, just like Bob's chef. Here, you can relate the data storage with the restaurant's food shelf and the traditional processing unit with the chef as shown in the figure above.

Fig: Hadoop Tutorial – Traditional Scenario

After few months, Bob thought of expanding his business and therefore, he started taking online orders and added few more cuisines to the restaurant's menu in order to engage a larger audience. Because of this transition, the rate at which they were receiving orders rose to an alarming figure of 10 orders per hour and it became quite difficult for a single cook to cope up with the current situation. Aware of the situation in processing the orders, Bob started thinking about the solution.

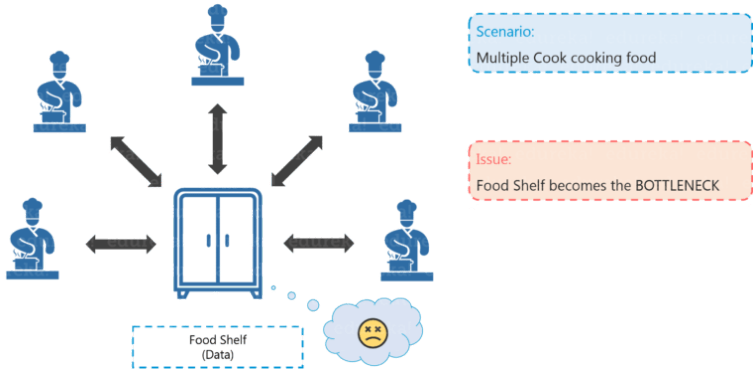

Fig: Hadoop Tutorial – Distributed Processing Scenario

Similarly, in Big Data scenario, the data started getting generated at an alarming rate because of the introduction of various data growth drivers such as social media, smartphones etc. Now, the traditional system, just like cook in Bob's restaurant, was not efficient enough to handle this sudden change. Thus, there was a need for a different kind of solutions strategy to cope up with this problem.

After a lot of research, Bob came up with a solution where he hired 4 more chefs to tackle the huge rate of orders being received. Everything was going quite well, but this solution led to one more problem. Since four chefs were sharing the same food shelf, the very food shelf was becoming the bottleneck of the whole process. Hence, the solution was not that efficient as Bob thought.

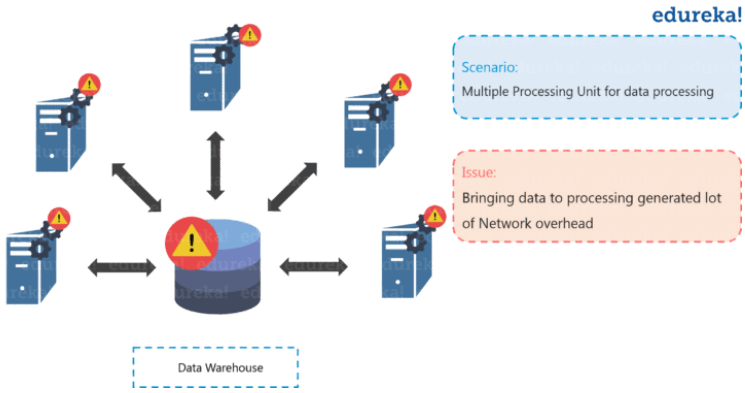

Fig: Hadoop Tutorial – Distributed Processing Scenario Failure

Similarly, to tackle the problem of processing huge datasets, multiple processing units were installed so as to process the data parallelly (just like Bob hired 4 chefs). But even in this case, bringing multiple processing units was not an effective solution because: the centralized storage unit became the bottleneck. In other words, the performance of the whole system is driven by the performance of the central storage unit. Therefore, the moment our central storage goes down, the whole system gets compromised. Hence, again there was a need to resolve this single point of failure.

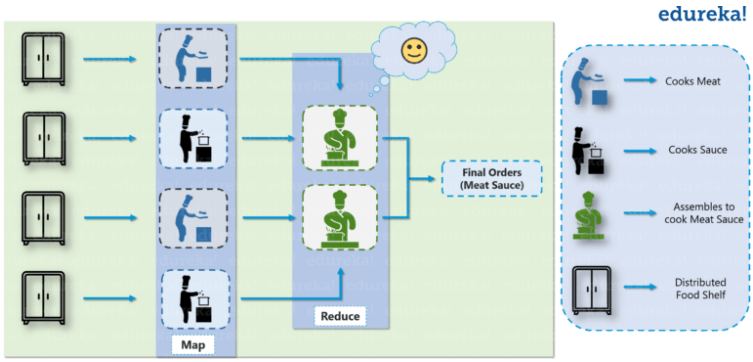

Fig: Hadoop Tutorial – Solution to Restaurant Problem

Bob came up with another efficient solution, he divided all the chefs in two hierarchies, i.e. junior and head chef and assigned each junior chef with a food shelf. Let us assume that the dish is Meat Sauce. Now, according to Bob's plan, one junior chef will prepare meat and the other junior chef will prepare the sauce. Moving ahead they will transfer both meat and sauce to the head chef, where the head chef will prepare the meat sauce after combining both the ingredients, which then will be delivered as the final order.

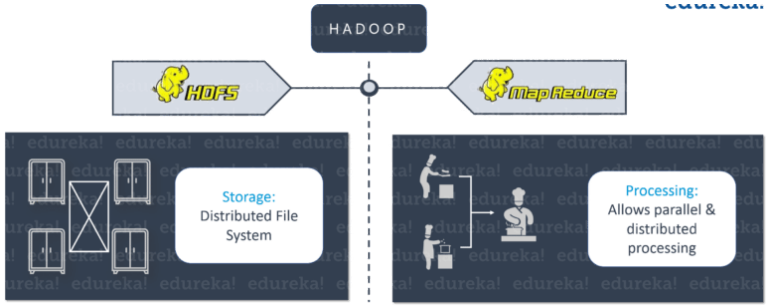

Fig: Hadoop Tutorial – Hadoop in Restaurant Analogy

Hadoop functions in a similar fashion as Bob's restaurant. As the food shelf is distributed in Bob's restaurant, similarly, in Hadoop, the data is stored in a distributed fashion with replications, to provide fault tolerance. For parallel processing, first the data is processed by the slaves where it is stored for some intermediate results and then those intermediate results are merged by master node to send the final result.

Now, you must have got an idea why Big Data is a problem statement and how Hadoop solves it. As we just discussed above, there were three major challenges with Big Data:

- ***The first problem is storing the colossal amount of data.*** Storing huge data in a traditional system is not possible. The reason is obvious, the storage will be limited to one system and the data is increasing at a tremendous rate.
- ***The second problem is storing heterogeneous data.*** Now we know that storing is a problem, but let me tell you it is just one part of the problem. The data is not only huge, but it is also present in various formats i.e. unstructured, semi-structured and structured. So, you need to make sure that you have a system to store different types of data that is generated from various sources.
- ***Finally let's focus on the third problem, which is the processing speed***. Now the time taken to process this huge amount of data is quite high as the data to be processed is too large.

To solve the storage issue and processing issue, two core components were created in Hadoop – **HDFS** and **YARN**. HDFS solves the storage issue as it stores the data in a distributed fashion and is easily scalable. And, YARN solves the processing issue by reducing the processing time drastically. Moving ahead, let us understand what is Hadoop?

Hadoop Tutorial: What is Hadoop?

Hadoop is an open-source software framework used for storing and processing Big Data in a distributed manner on large clusters of commodity hardware. Hadoop is licensed under the Apache v2 license. Hadoop was developed, based on the paper written by Google on MapReduce system and it applies concepts of functional programming. Hadoop is written in the Java programming language and ranks among the highest-level Apache projects. Hadoop was developed by Doug Cutting and Michael J. Cafarella.

Hadoop Tutorial: Hadoop-as-a-Solution

Let's understand how Hadoop provides solution to the Big Data problems that we have discussed so far.

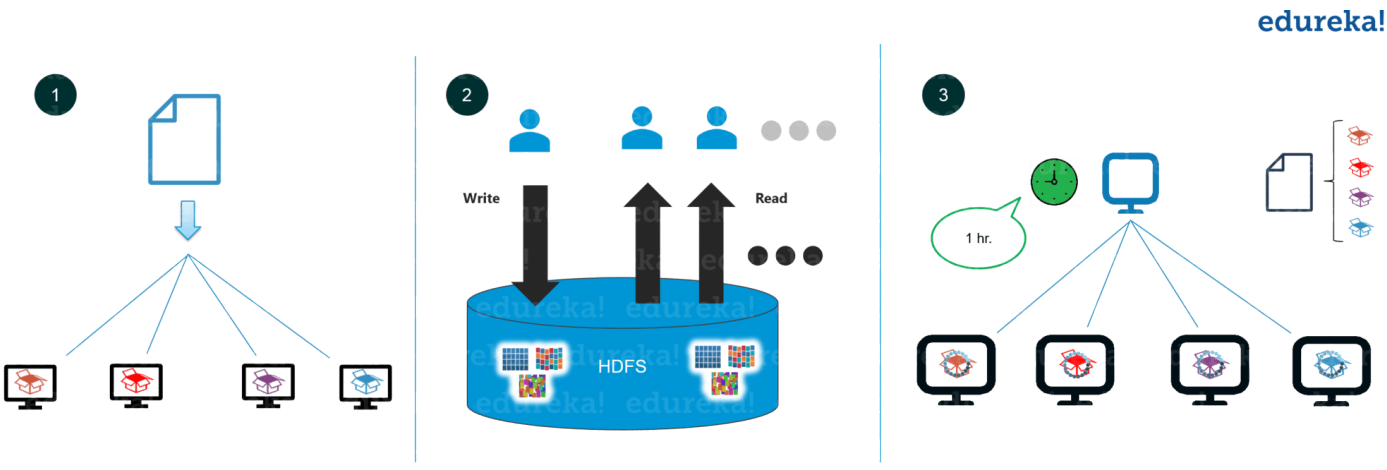

Fig: Hadoop Tutorial – Hadoop-as-a-Solution

The first problem is storing huge amount of data.

As you can see in the above image, HDFS provides a distributed way to store Big Data. Your data is stored in blocks in DataNodes and you specify the size of each block. Suppose you have 512MB of data and you have configured HDFS such that it will create 128 MB of data blocks. Now, HDFS will divide data into 4 blocks as 512/128=4 and stores it across different DataNodes. While storing these data blocks into DataNodes, data blocks are replicated on different DataNodes to provide fault tolerance.

Hadoop follows ***horizontal scaling*** instead of vertical scaling. In horizontal scaling, you can add new nodes to HDFS cluster on the run as per requirement, instead of increasing the hardware stack present in each node.

Next problem was storing the variety of data.

As you can see in the above image, in HDFS you can store all kinds of data whether it is structured, semi-structured or unstructured. In HDFS, there is *no pre-dumping*

schema validation. It also follows write once and read many model. Due to this, you can just write any kind of data once and you can read it multiple times for finding insights.

The third challenge was about processing the data faster.

In order to solve this, we move processing unit to data instead of moving data to processing unit. So, what does it mean by moving the computation unit to data? It means that instead of moving data from different nodes to a single master node for processing, the processing logic is sent to the nodes where data is stored so as that each node can process a part of data in parallel. Finally, all of the intermediary output produced by each node is merged together and the final response is sent back to the client.

VIEW UPCOMING HADOOP BATCHES
(https://www.edureka.co/big-data-and-hadoop)

Hadoop Tutorial: Hadoop Features

Fig: Hadoop Tutorial – Hadoop Features

Reliability:

When machines are working in tandem, if one of the machines fails, another machine will take over the responsibility and work in a reliable and fault tolerant fashion. Hadoop infrastructure has inbuilt fault tolerance features and hence, Hadoop is highly reliable.

Economical:

Hadoop uses commodity hardware (like your PC, laptop). For example, in a small Hadoop cluster, all your DataNodes can have normal configurations like 8-16 GB RAM with 5-10 TB hard disk and Xeon processors, but if I would have used hardware-based RAID with Oracle for the same purpose, I would end up spending 5x times more at least. So, the cost of ownership of a Hadoop-based project is pretty minimized. It is easier to maintain the Hadoop environment and is economical as well. Also, Hadoop is an open source software and hence there is no licensing cost.

Scalability:

Hadoop has the inbuilt capability of integrating seamlessly with cloud-based services. So, if you are installing Hadoop on a cloud, you don't need to worry about the scalability factor because you can go ahead and procure more hardware and expand your setup within minutes whenever required.

Flexibility:

Hadoop is very flexible in terms of ability to deal with all kinds of data. We discussed "Variety" in our previous blog on ***Big Data Tutorial (https://www.edureka.co/blog/big-data-tutorial)***, where data can be of any kind and Hadoop can store and process them all, whether it is structured, semi-structured or unstructured data.

These 4 characteristics make Hadoop a front-runner as a solution to Big Data challenges. Now that we know what is Hadoop, we can explore the core components of Hadoop. Let us understand, what are the core components of Hadoop.

Hadoop Tutorial: Hadoop Core Components

While setting up a Hadoop cluster, you have an option of choosing a lot of services as part of your Hadoop platform, but there are two services which are always mandatory for setting up Hadoop. One is **HDFS (storage)** and the other is **YARN (processing)**. HDFS stands for **Hadoop Distributed File System**, which is a scalable storage unit of Hadoop whereas YARN is used to process the data i.e. stored in the HDFS in a distributed and parallel fashion.

HDFS

Let us go ahead with HDFS first. The main components of **HDFS** are: **NameNode** and **DataNode**. Let us talk about the roles of these two components in detail.

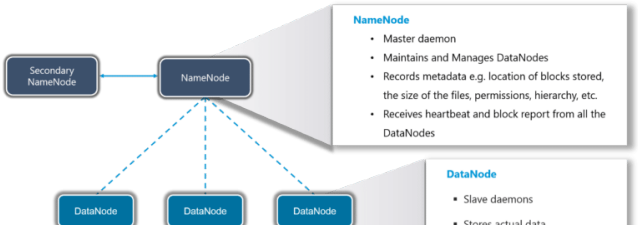

Fig: Hadoop Tutorial – HDFS

NameNode

- It is the master daemon that maintains and manages the DataNodes (slave nodes)
- It records the metadata of all the blocks stored in the cluster, e.g. location of blocks stored, size of the files, permissions, hierarchy, etc.
- It records each and every change that takes place to the file system metadata
- If a file is deleted in HDFS, the NameNode will immediately record this in the EditLog
- It regularly receives a Heartbeat and a block report from all the DataNodes in the cluster to ensure that the DataNodes are live
- It keeps a record of all the blocks in the HDFS and DataNode in which they are stored
- It has high availability and federation features which I will discuss in ***HDFS architecture (https://www.edureka.co/blog/hdfs-tutorial)*** in detail

DataNode

- It is the slave daemon which run on each slave machine
- The actual data is stored on DataNodes
- It is responsible for serving read and write requests from the clients
- It is also responsible for creating blocks, deleting blocks and replicating the same based on the decisions taken by the NameNode
- It sends heartbeats to the NameNode periodically to report the overall health of HDFS, by default, this frequency is set to 3 seconds

So, this was all about HDFS in nutshell. Now, let move ahead to our second fundamental unit of Hadoop i.e. YARN.

YARN

YARN comprises of two major component: **ResourceManager** and **NodeManager**.

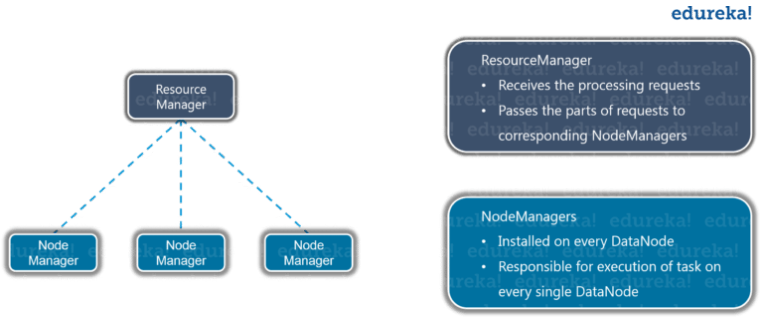

Fig: Hadoop Tutorial – YARN

ResourceManager

- It is a cluster level (one for each cluster) component and runs on the master machine
- It manages resources and schedule applications running on top of YARN
- It has two components: Scheduler & ApplicationManager
- The Scheduler is responsible for allocating resources to the various running applications
- The ApplicationManager is responsible for accepting job submissions and negotiating the first container for executing the application
- It keeps a track of the heartbeats from the Node Manager

NodeManager

- It is a node level component (one on each node) and runs on each slave machine
- It is responsible for managing containers and monitoring resource utilization in each container
- It also keeps track of node health and log management
- It continuously communicates with ResourceManager to remain up-to-date

Hadoop Tutorial: Hadoop Ecosystem

So far you would have figured out that Hadoop is neither a programming language nor a service, it is a platform or framework which solves Big Data problems. You can consider it as a suite which encompasses a number of services for ingesting, storing and analyzing huge data sets along with tools for configuration management.

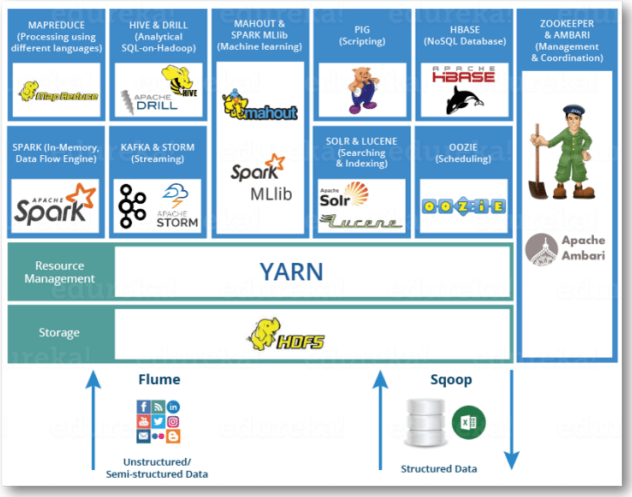

Fig: Hadoop Tutorial – Hadoop Ecosystem

We have discussed Hadoop Ecosystem and their components in detail in our ***Hadoop Ecosystem blog (https://www.edureka.co/blog/hadoop-ecosystem)***. Now in this Hadoop Tutorial, let us know how **Last.fm used Hadoop as a part of their solution strategy**.

Hadoop Tutorial: Last.fm Case Study

Last.fm is internet radio and community-driven music discovery service founded in 2002. Users transmit information to Last.fm servers indicating which songs they are listening to. The received data is processed and stored so that, the user can access it in the form of charts. Thus, Last.fm can make intelligent taste and compatible decisions for generating recommendations. The data is obtained from one of the two sources stated below:

- **scrobble:** When a user plays a track of his or her own choice and sends the information to Last.fm through a client application.
- **radio listen:** When the user tunes into a Last.fm radio station and streams a song.

Last.fm applications allow users to love, skip or ban each track they listen to. This track listening data is also transmitted to the server.

- Over 40M unique visitors and 500M page views each month
- Scrobble stats:
  - Up to 800 scrobbles per second
  - More than 40 million scrobbles per day
  - Over 75 billion scrobbles so far
- Radio stats:
  - Over 10 million streaming hours per month
  - Over 400 thousand unique stations per day
- Each scrobble and radio listen generates at least one log line

Hadoop at Last.FM

- 100 Nodes
- 8 cores per node (dual quad-core)
- 24GB memory per node
- 8TB (4 disks of 2TB each)
- Hive integration to run optimized SQL queries for analysis

Last.FM started using Hadoop in 2006 because of the growth in users from thousands to millions. With the help of Hadoop they processed hundreds of daily, monthly, and weekly jobs including website stats and metrics, chart generation (i.e. track statistics), metadata corrections (e.g. misspellings of artists), indexing for search, combining/formatting data for recommendations, data insights, evaluations & reporting. This helped Last.FM to grow tremendously and figure out the taste of their users, based on which they started recommending musics.

I hope this blog was informative and added value to your knowledge. In our next blog on ***Hadoop Ecosystem*** (https://www.edureka.co/blog/hadoop-ecosystem), we will discuss different tools present in Hadoop Ecosystem in detail.

MASTER HADOOP WITH EDUREKA

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

*Now that you have understood Hadoop 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 Shubham Sinha (25 Posts (https://www.edureka.co/blog/author/shubham-sinha/))

Shubham Sinha is a Big Data and Hadoop expert working as a Research Analyst at Edureka. He is keen to work with Big Data related technologies such as Hadoop, Spark, Flink and Storm and web development technologies including Angular, Node.js & PHP.

Share on (https://www.facebook.com/sharer.php?u=https://www.edureka.co/blog/hadoop-tutorial/) (https://twitter.com/intent/tweet?text=Hadoop%20Tutorial%20|%20Getting%20Started%20With%20Big%20Data%20And%20Hadoop%20|%20Edureka&url=https://www.edureka.co/blog/hadoop-tutorial/) (https://www.linkedin.com/shareArticle?mini=true&url=https://www.edureka.co/blog/hadoop-tutorial/&title=Hadoop%20Tutorial%20|%20Getting%20Started%20With%20Big%20Data%20And%20Hadoop%20|%20Edureka&summary=Hadoop%20Tutorial%20|%20Getting%20Started%20With%20Big%20Data%20And%20Hadoop%20|%20Edureka&source=https://www.edureka.co/blog/hadoop-tutorial/) (https://plus.google.com/share?url=https://www.edureka.co/blog/hadoop-tutorial/)

PREVIOUS NEXT

Got your brain cells running?
Stay tuned to latest technology updates

Enter your Email Address

SUBSCRIBE

Related Posts

10 Reasons Why Big Data Analytics is the Best Career Move
251.4K
(https://www.edureka.co/blog/10-reasons-why-big-data-analytics-is-the-best-career-move)

HDFS Tutorial: Introduction to HDFS & its Features
35.2K
(https://www.edureka.co/blog/hdfs-tutorial)

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

Hadoop Certification – Become a Certified Big Data Hadoop Professional
13.3K
(https://www.edureka.co/blog/hadoop-certification/)

Comments

6 Comments

6 Comments https://www.edureka.co/blog/ Rajiv Chaudhuri
Recommend 4 Share Sort by Best
Join the discussion…

ruchi chouhan • 5 months ago
Big data is basically indicating large amount of data. Now a day data is increasing day by day ,so handle this large amount of data Big Data term is came. Hadoop is open source ,distributed java based programming framework that was launched as an Apache open source project in2006.MapReduce algorithm is used for run the Hadoop application ,where the data is processed in parallel on different CPU nodes.
1 • Reply • Share ›

harshita chawhan • 7 days ago
I really like this post
www.bisptrainings.com
• Reply • Share ›

Kokila • a year ago
Good blog. Thanks for sharing this information. keep sharing about hadoop tutorial. Data Analytics Training Bangalore
• Reply • Share ›

Bhaskar Das • a year ago
Is it possible to create an Encryption Zone in the HDFS or Hive Warehouse, when we will put or load any data or table into encryption zone location then it will get encrypted automatically?
• Reply • Share ›

EdurekaSupport Mod ➚ Bhaskar Das • a year ago
Hey Bhaskar, thanks for checking out our blog.
Yes, it is possible to create zones and encrypt it using Hadoop provided APIs .You can refer the link for reference https://docs.hortonworks.co...
Hope this helps. Cheers!
• Reply • Share ›

Shaik • 3 years ago
Hi
• Reply • Share ›

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

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

Spring Boot Microservices: Building Microservices Application Using Spring Boot
60 comments • 3 months ago
Gunasekaran — Thanks!!!

Java Thread Tutorial: Creating Threads and Multithreading in Java
4 comments • 7 months ago
Himanshu Garg — we can extend only one class in java...if we extend thread then we can not extend another class whereas we can implement multiple interfaces

Capsule Neural Networks – Set of Nested Neural Layers
1 comment • 8 months ago
Saurabh Jain — Hi Saurabh ,Would like to know about ..How to compare two images using the Neural net ?? Like i have few brain MRI ...in that case ..how can i compare the same in R ...Do …

Subscribe Add Disqus to your siteAdd DisqusAdd Disqus' Privacy PolicyPrivacy PolicyPrivacy

Subscribe
to our newsletter

Enter your Email Address

SUBSCRIBE

Related Blogs

10 Reasons Why Big Data Analytics is the Best Career... (https://www.edureka.co/blog/10-reasons-why-big-data-analytics-is-the-best-career-move)

(https://www.edureka.co/blog/10-reasons-why-big-data-analytics-is-the-best-career-move)

HDFS Tutorial: Introduction to HDFS & its Features (https://www.edureka.co/blog/hdfs-tutorial)

(https://www.edureka.co/blog/hdfs-tutorial)

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

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

Hadoop Certification – Become a Certified Big Data Hadoop Professional (https://www.edureka.co/blog/hadoop-certification/)

(https://www.edureka.co/blog/hadoop-certification/)

© 2014 Brain4ce Education Solutions Pvt. Ltd. All rights Reserved.

"PMP®","PMI®", "PMI-ACP®" and "PMBOK®" are registered marks of the Project Management Institute, Inc.
MongoDB®, Mongo and the leaf logo are the registered trademarks of MongoDB, Inc.