



Big Data Tutorial for Absolute Beginners

Introduction

“Entering the world of Big Data feels like being immersed in an ocean of buzz words.”

Many new to the world of data are confused between Hadoop and Spark or find themselves overwhelmed by the plethora of types of data that are either structured, semi-structured, or unstructured. A person may wonder how to make the transition from the spreadsheets of the traditional world to the management of ‘petabytes’ of data.

Our big data tutorial system demystifies the complexity, explaining big data in manageable chunks. We cover the theory, as well as the practical application, in the field of engineering.

Are you ready to learn the skills of the future? Click here to see our [big data course syllabus](#) and learn how to become a data engineer today!

Why Students or Freshers Learn Big Data?

Big Data has moved beyond being a buzz term in a company to being a defining element within a career for students and freshers. The following are reasons why you need to begin learning about Big Data today:

- **Explosive Market Growth:** The Big Data market will reach billions in value, which gives rise to a massive gap between the demand and supply of talent in this sector.
- **Industry-Wide Significance:** Be it the movie recommendation algorithms of Netflix or the fraud detection mechanisms of banks, every prominent sector is today based on data-driven decision-making, not intuition.
- **Competitive Salary Packages:** Freshers who know Big Data skills such as Hadoop, Spark, and SQL command a salary that's 30-40% more than the software development salary, the CEO said.
- **Future-Proofing Your Career:** With the mainstream adoption of AI & Machine Learning, Big Data is the "fuel" necessary to power this paradigm, making your skill set relevant for the next several decades.
- **Advanced Problem-Solving Skill:** This changes your thought processes from coding to architecture level problems, and it teaches you to design "human fault-tolerant" pipelines that process Quintillion Bytes every day.

Looking to make an impression during your next technical interview? Click here to download our [Big Data Interview Questions and Answers](#) and get your dream job today!

Take your Knowledge
Test Report

Check your Score



Step-by-Step Big Data Tutorial for Beginners

The world is producing so much data that 2.5 quintillion bytes of data are being produced daily. Traditional databases have no means of handling this data. Here comes Big Data. This big data tutorial is going to help you transition from thinking traditionally about data to building a data superpower using distributed computing.

Step 1. Laboratory Equipment: Setting up Ecosystems

To study Big Data, one would require a cluster. As we don't have ten computers with us most of the time, we run it on a single machine using the facility called Pseudo-Distributed Mode. Here, we would concentrate on the industry leader: Apache Hadoop and Apache Spark.

1.1 Install Java (The Engine)

Hadoop and Spark are Java-based. To work with them, you would need OpenJDK versions 8 or 11.

```
# Update your package list
```

```
sudo apt update
```

```
# Install OpenJDK 8
```

```
sudo apt install openjdk-8-jdk -y
```

```
# Verify the version
```

```
java -version
```

1.2: Setting Up Passwordless SSH

Hadoop employs SSH to manage its nodes, no matter whether it's one or multiple nodes.

```
sudo apt install openssh-server -y
```

```
ssh-keygen -t rsa -P "" -f ~/.ssh/id_rsa
```

```
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

```
chmod 0600 ~/.ssh/authorized_keys
```

1.3 Install and Configure Hadoop

Download the stable Hadoop binary version (for example, 3.3.6) and configure your environment variables in `~/.bashrc` files.

Environment Variables to Add:

```
export HADOOP_HOME=~/.hadoop
```

```
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
```

Step 2. Core Concept: HDFS (The Storage)

Hadoop Distributed File System (HDFS) is the core storage system for Big Data. It divides large files into “blocks” that are stored on multiple machines.

Primary Operations:

Imagine HDFS to be an off-site storage drive. Files are not opened; they are put into or retrieved from.

- **Make a new directory:** *hdfs fs -mkdir /user/data*
- **Upload a file:** *hdfs dfs -put localfile.txt /user/data*
- **List files:** Display a list of the files in those directories: *hdfs dfs -ls /user/data*

Step 3. Data Processing with Apache Spark

Even though the speed innovator was Hadoop's MapReduce, the King of Speed today is Apache Spark. Data is processed "in-memory," meaning it can be 100 times faster than MapReduce.

Practical Code: The Word Count (PySpark)

The "Hello World" of Big Data will include word-counting in a large text file. Now, here is how one can carry out the task in the Spark Python API:

```
from pyspark.sql import SparkSession

# Initialize Spark Session

spark = SparkSession.builder.appName("WordCount").getOrCreate()

# Read text file from HDFS

text_file = spark.read.text("hdfs:///user/data/large_dataset.txt")

# Transformation: Split lines into words and count them

words = text_file.rdd.flatMap(lambda line: line[0].split(" "))

word_counts = words.map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)

# Action: Save the result back to HDFS

word_counts.saveAsTextFile("hdfs:///user/data/output_counts")
```

Step 4. Understanding The 5 V's

To truly understand Big Data, you need to expand beyond just the term "Big." Here is a compiled list of things that need

- **Volume:** The Volume refers to the magnitude of the data that ranges from terabytes to
- **Velocity:** The speed at which data is produced (real-time data streams).
- **Variety:** Multiple types of data, e.g., Images, Logs, Tweets,
- **Veracity:** It refers to data accuracy, cleanliness, and integrity.
- **Value:** The end game—converting raw bits into business insights.

Step 5. Modern Data Pipelines: The ETL Process

People working in Big Data do 80% of their work on ETL.

- **Extract:** Extracting information from social media APIs, IoT devices, or SQL databases.
- **Transform:** Null value cleaning, format change, and grouping.
- **Load:** It involves transferring the refined data from a Data Lake (such as Amazon S3) to a Data Warehouse (like Snowflake).

Big Data isn't just about familiarity with tools; it's about familiarity with the tool to solve the problem at the right time. But whether you opt for the reliability of Hadoop, the speed of Spark, or the capabilities of real-time processing of Kafka, the principles of distributed computing are the same.

Prepared to apply these skills to solve the architectural mysteries that the real world presents? Take the challenge by working through these tough scenarios, from managing skewed data to optimizing Spark partitions. Click here to access [Big Data Challenges and Solutions](#).

Real Time Examples for Big Data Tutorial for Learners

But in order to fully grasp the true power of Big Data, it is necessary to strip away the code and observe how it processes the unprecedented amounts of high-speed data that characterize contemporary existence. Now, here are some case studies of the use of Big Data ecosystems:

Streaming Services (Netflix, Spotify)

- Each time you pause, skip, or re-watch a scene, it provides data for Netflix. Having more than 200 million subscribers, it gives them the “Volume” and “Velocity” issue.
- Apache Spark is utilized here; it processes billions of events every day, making it generate “Variety” in their recommendation profiles, displaying their “Veracity” in determining the exact series or program that will retain subscribers.

Smart City and IoT (Traffic Control)

- Thousands of sensors are used in today’s cities, implanted in roads, traffic lights, and GPS systems in buses.
- It’s a Big Data problem in Variety, as structured data is extracted from GPS systems and unstructured data from cameras.
- The data is processed in real-time by Apache Kafka, and traffic light times are changed in real-time by smart cities to avoid traffic congestion and greenhouse gas emissions in the Atmosphere.

Financial Fraud Detection

- Credit card companies handle thousands of transactions every second. Big Data systems assess your past purchase behavior (Data at Rest) and then compare it to your present transaction (Data in Motion) in a fraction of a second.
- If you unexpectedly purchase a high-end laptop in a region you never previously visited in your whole life, Machine Learning programs processing a Big Data system classify it as fraud before it is even confirmed.

With knowledge of these concepts, you can move from a simple programmer to a Data Architect proficient in developing systems to cater to millions of users at a time.

Ready to develop your own data-driven application? It's time to turn theory into reality, and this is where all your newfound knowledge truly begins. Click here to explore our [Big Data Project Ideas](#) along with Real-World Datasets.

FAQs About Big Data Tutorial for Beginners

1. What do you mean by big data?

Big data is characterized as large amounts of data that are too large for typical database management techniques. It is also described as big data analytics, referring to those techniques used to capture, store, and analyze data that is so large that data is measured in terms of petabytes or exabytes.

2. What are the 4 types of big data?

Data is generally categorized by its format:

Structured: Very organized (SQL tables).

Unstructured: Unorganized, no defined structure (Videos, PDFs).

Semi-Structured Formats: This format is semi-structured (JSON, XML).

Quasi-Structured: Unorganized formats (search result strings).

3. How big data differs from normal data?

Big Data differs from normal data processing in the sense that it requires distributed computing because of the aforementioned 4 elements.

4. What are the 5 big data?

It pertains to the 5 V's which identify the challenges involved in big data:

Volume: Size of the data.

Velocity: The rate of creation.

Variety: This refers to the types of data.

Veracity: The quality or accuracy.

Value: The findings obtained.

5. What is the basic concept of data?

Data is an assembly of unprocessed facts, figures, and observations that include numbers, words, or measurements converted into a form that a computer can process.

In their unprocessed form, data is devoid of any meaning, and after it has been processed, it attains importance in the form of information. Explore [big data salary for freshers](#).

6. What tools are used for big data?

It requires distributed technologies in Hadoop, such as HDFS for data storage in Hadoop, Apache Spark, Kafka libraries (for real-time data transfer), NoSQL databases (for storage), or BigQuery (or Amazon EMR, used with Google Big Query).

7.What are the benefits of big data?

Big data allows capabilities in the following areas:

Predictive Analytics: This has the ability to analyze forecasting trends.

Personalized Marketing: This has the capacity to provide recommendation engines.

8. Is Excel a big data?

No, it's not a big data tool. It can process approximately 1 million rows of data per sheet and can process data in the memory of a single computer. Big data can process data in the size of billions of rows of data.

9.What is the basic concept of big data?

The fundamental idea is Distributed Computing. Rather than using one massive and costly computer to analyze a gigantic file, big data involves splitting that file into smaller files and analyzing those files in a “cluster” of many regular computers working in parallel.

10.Which is an example of big data?

The example of Big Data is The New York Stock Exchange, which produces one terabyte of new trade data each day. Other examples of Big Data include the number of social media posts on the social networking site, X (twitter), and the sensor data produced by jet engines during each flight.

Conclusion

The skill of handling Big Data is more about transitioning from information management to insight engineering. You have just begun your journey through how systems, like

Hadoop and Spark, turn noise into signal. As the need for real-time analytics and AI convergence increases across industries, the need for these skills will continue to remain at the forefront of the tech arena. The transition from thinking like a single machine to thinking like a cluster is your first lesson in becoming a master Data Engineer. Enroll in our [Big Data Architect Certification Course in Chennai](#) to learn how to create production-level pipelines.