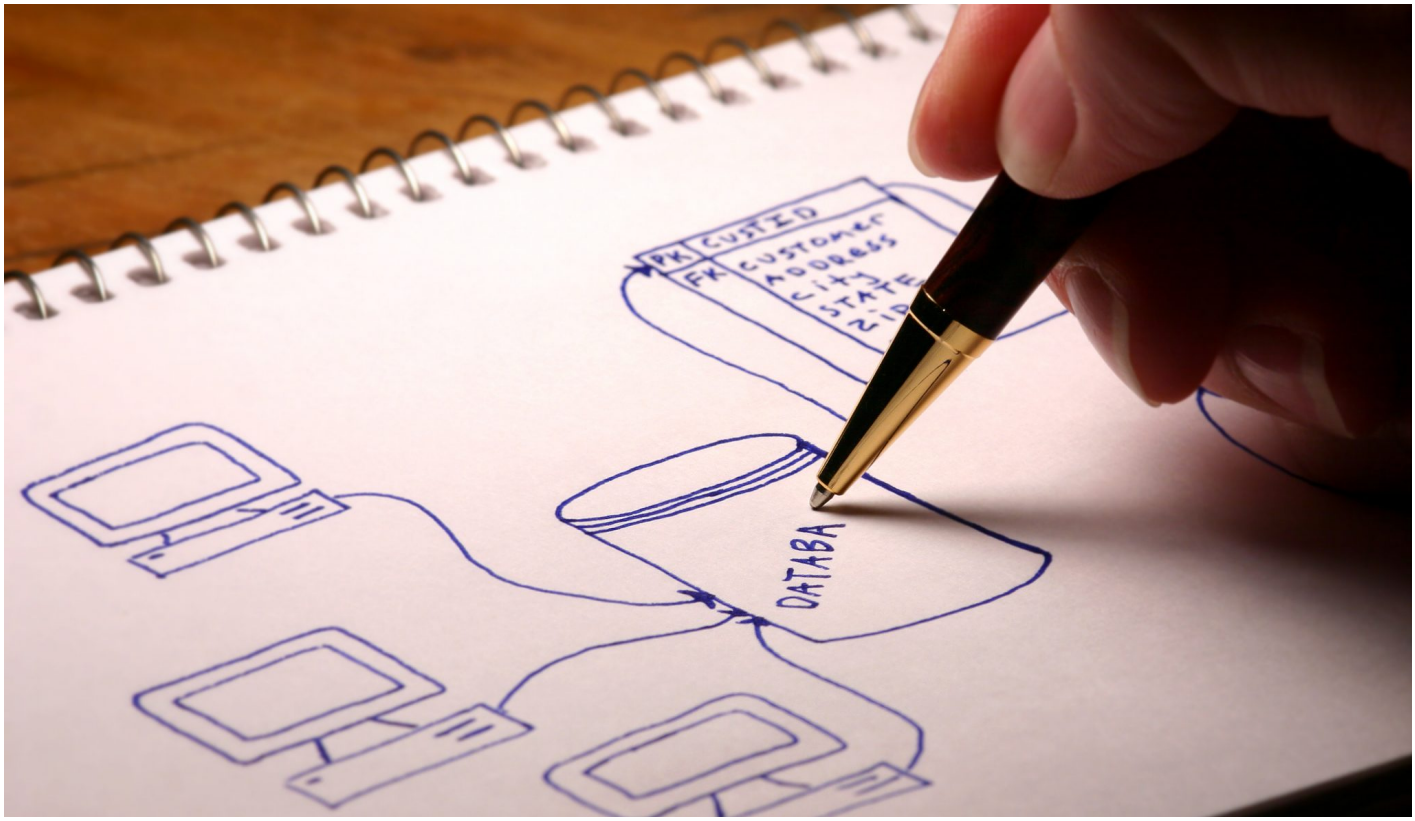
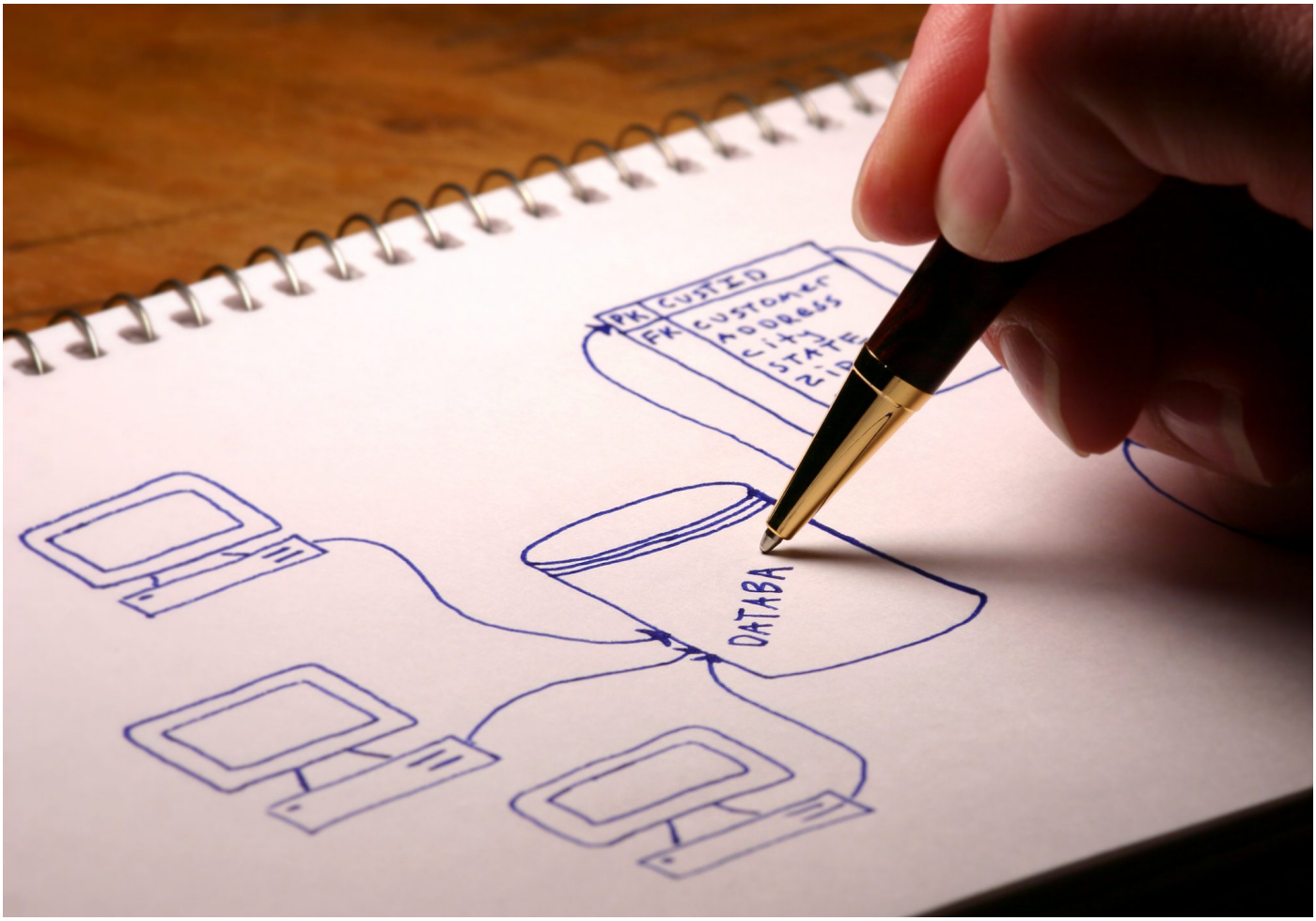
Data Engineering 101: Top Tools And Framework Resources

BY KISHAN MALADKAR
08/08/2018





In today's fast-paced world, data can be compared to DNA — with data, it is easy to understand the past, predict the future and also replicate what it contains. Back in the early 2000s, the amount of data collected was just 5 to 10 percent of what we have collected in the last two years. Data collection, data engineering, and managing the warehouses are in high demand right now. Every company in today's world wants to hire highly-skilled professionals who can deal with massive amounts of data and draw insights from it.

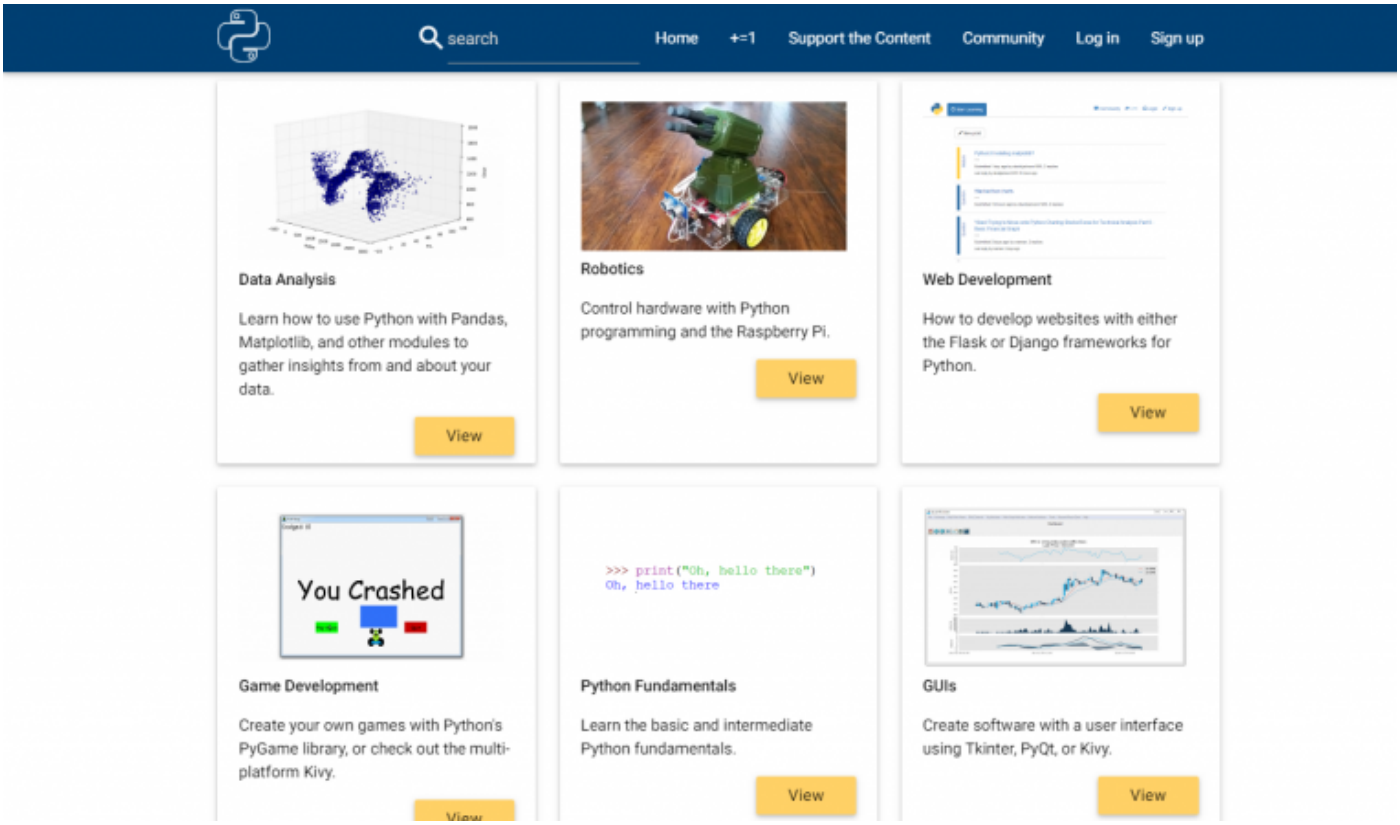
There is no formal degree to be a data engineering graduate as of now. Nonetheless, there is a huge demand for data engineers and companies are hiring engineers for analytics positions.

A recent study conducted by *Analytics India Magazine* found out that programming languages Python and R are commonly used across this domain for analysis and visualisation. Let us look at some of the MOOCs and books from which one can learn important prerequisites for data engineers — programming languages such as Python, R, and big data tools like Hadoop and Spark.

In this article, we shall look at some of the well-known resources, both paid and free, from which one can acquire the right skills for a data engineering role. *We have listed these resources according to the learning order.*

1| Python

I| Python Programming by Sentdex (MOOC)



This is an open-source educational platform built and managed by Harrison Kinsley.

One can learn Python from scratch here since it is one of the best free MOOCs out on the internet. There are advanced concepts of web developments, robotics explained using Python, which is quite fascinating. There are other interesting projects which Harrison himself has explained and built in real time.

II| How To Code In Python by Lisa Tagliaferri (eBook)

Python is one of the most versatile languages and is considered as the most widely-used language among developers in 2018. It has gained a lot of attention because it supports the scripting and object oriented programming style. This book explains how one from a non-programming background can learn and implement python for developing and various purposes. The author also explains how easy it is to learn Python because of it uses easy English words used in programming.

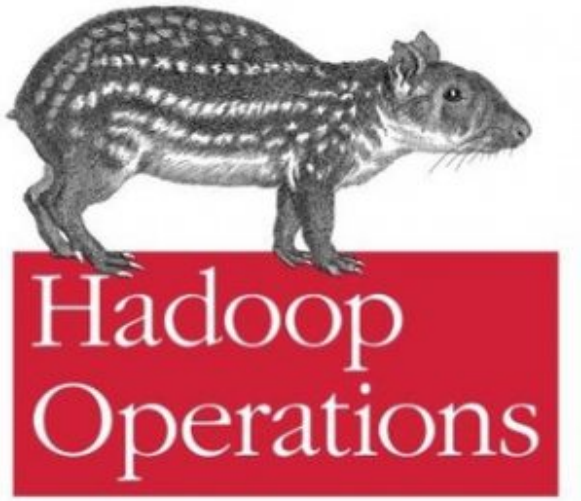
2| R

I| Introduction to R – DataCamp (MOOC)

This course is focused towards statistical modelling and analysis using R language. As many companies ask for R skills during hiring, this course comes in very handy. If one knows how to handle data then the company expects you to understand it too.

3| Apache Hadoop

I| Hadoop Operations by Eric Sammer (Book)



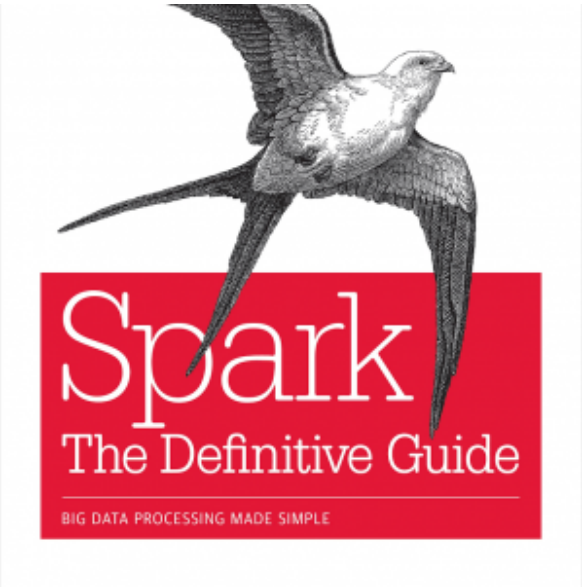
Eric Sammer has explained how one can start with Hadoop, from installing it on your system to architecture construction. It also explains clustering data with huge samples. An overview of HDFS and MapReduce has also been explained — why they are implemented and how they help in streaming the data. This is great to cluster and run a production environment.

II| Hadoop Platform and Application Framework by Coursera (MOOC)

This course focuses on teaching Hadoop frameworks for big data analysis. Also teaches the MapReduce techniques various other Hadoop-related content.

4| Apache Spark

1| Spark: The Definitive Guide: Big Data Processing Made Simple (eBook) by Bill Chambers and Matei Zaharia



SEE ALSO

A Hands-on Guide To Hybrid Ensemble Learning Models, With Python Code

This book talks about how one can deal with query languages like SQL, learn about data frames and also make use of Spark's API. Spark also includes clustering and monitoring, where one can process the data and execute them in real time. It also includes how one can make use of MLlib of Spark for data modelling and machine learning applications.

II| Introduction To Apache Spark and AWS – Coursera (MOOC)

An end-to-end applications of Spark is explained in this Coursera course. Spark is 100 times faster than Hadoop MapReduce and 5 times faster on the disk. It also has real-time batch processing which is unavailable on Hadoop. This MOOC also gives you a grading system where one can have a hands on experience for better understanding.

5| Apache Kafka

I| Kafka: The Definitive Guide: Real-Time Data and Stream Processing at Scale (eBook) by Gwen Shapira, Neha Narkhede, and Todd Palino

Kafka: The Definitive Guide: Real-Time Data and Stream Processing at Scale

Streaming of data refers to controlling the data flow. With the help of computer programming, one can build a stream processing program which efficiently uses the concept of parallel programming for computations. This book gives you a quick understanding of the NoSQL databases and MongoDB. It also gives you insights on how relational databases are different from document-oriented databases.

II| Tutorials Point – Apache Kafka Tutorial (Open Source Tutorial)

This tutorial will explore the principles of Kafka, from installation to operations and then it will walk one through with the deployment of Kafka cluster. It is concluded with real-time applications, hands-on and integration with Big Data Technologies.

Conclusion

This article summarises several unique resources for learning and implementing data engineering concepts in the industry. One can make use of these to understand how to deal with and process huge databases. These are resources of 5 most common skills. The requirements for data engineer roles might vary depending on companies and they might ask for skills such as Java, C++, SQL, Scala, etc.