

Big Data Libraries in Python

1. TensorFlow

It's an open-source framework highly being used by data scientists around the globe. With the help of TensorFlow, a programmer can use dataflow and alternate programming methods to perform the different tasks that are training and inference centric of deep neural networks, and moreover, it allows data scientists to develop machine learning applications with the help of various tools, and resources. It was created by Google in 2015 and currently, it holds the position of the most used library around the world. Besides this, there are certain factors to look into while going to pick TensorFlow and surely this might be helpful for you:

- It eliminates the possibilities of error by 60%
- It's highly scalable and can be easily implemented
- With the help of its data structure, it can easily identify the structure using 3 major criteria i.e. rank, type, and shape.
- In its pipelining system, multiple neural networks and the GPU can be trained, eventually creating a large-scale system.

2. Pandas

The development of panda started between 2008 and the very first version was published back in 2012 which became the most popular open-source framework introduced by Wes McKinney. The demand for Pandas has grown enormously over the past few years and even today if collective feedback will be taken then panda will be their first choice without any doubt. The name "Panda" was derived from "Panel Data" which is an econometrics term for data sets. It also allows data scientists to create tabular, multidimensional, and certain different data structures. Apart from this, there are certain other key features of the panda that makes it so popular among data scientists, have a look at them:

- Panda offers high-speed performance in data merging
- With the help of Panda, data scientists can easily align and integrate data handling of the missing one's
- Panda offers developers to create self-functions and to run them across different series of data
- Panda also contains a high level of data structure and manipulation tools

3. NumPy

Initially, when developers needed to perform numerical calculations, NumPy was introduced in Data Science. It is currently registered under the BSD (Berkeley Source Distribution) license which makes it freely open to use. Numpy allows users to perform almost any computational calculations, even Linear Algebra can be easily be achieved using NumPy. It is often called a general-purpose array processing tool and helps users in boosting sloppy performance by offering multidimensional objects (arrays and metrics) so that the operation can go smoothly. Besides this, NumPy also provides the following benefits to data scientists in different approaches, some of them are:

- Being a general-purpose arrays and metrics processing package and most importantly, the arrays in NumPy can be either one or multi-dimensional.
- It can also perform complex operations (linear algebra, Fourier transform, etc.) and for that NumPy has different modules for each set of complex functions.
- NumPy is so flexible that it can easily work with different languages by using its functions. Therefore, the functions of NumPy allow it to integrate with other languages which also include inter-platform functions.
- NumPy carries broadcasting functions which means if you're working on an array of any uneven shape, it will highlight/broadcast the shape of smaller arrays as per the larger ones.

4. Matplotlib

It is used as a 2D plotting graphic in the python programming language. Besides this, matplotlib can also be used to create histograms, power spectra, error charts, etc. Matplotlib also offers an object-oriented API that helps in embedding those plots in applications. It was introduced first in 2002 by John D. Hunter under a BSD license and was released publicly in 2003. Besides this, it also offers some extensive key features which can be looked into while choosing big data analysis:

- It helps in understanding data visualization, data analysis, and other insights of data in a better way
- The scripts of Matplotlib are already structured and the developer need not perform the entire coding and its scripts can overlap up to two APIs at a time.
- As discussed above, Matplotlib offers an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, etc.
- Matplotlib supports an extensive range of backend and output types which means that your output will not be based on what OS you're operating at that time.

5. SciPy

Abbreviated as Science Python, SciPy is a scientific computational library that generally uses NumPy. It offers more utility functions that enable better visualization, optimization, and so on. Besides this, it's an open-source platform which means anyone can use SciPy without any restrictions. Although it's written in python it holds certain elements of C Programming too. If you'll look up the trend, today it is often used by data scientists around the globe and has gained popularity by not only offering user-friendly and complex calculations but also it is one of the best choices, especially for beginners who wish to get into data science industry. However, there are some other factors to consider before diving into it:

- It's open-source under BSD license and numFORCE which means anyone can use it freely and openly.
- It can handle large data sets both as effectively and efficiently.
- NumPy carries little to envy from other specialized environments for data analysis and calculation (such as R or MATLAB).
- It helps in solving differential equations which includes linear algebra, and the Fourier transform