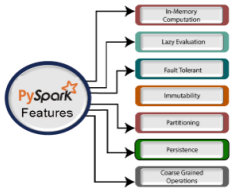
```
Start coding or generate with AI.
```

What is PySpark?

PySpark is a Python API to support Python with Apache Spark. PySpark provides Py4j library, with the help of this library, Python can be easily integrated with Apache Spark. PySpark plays an essential role when it needs to work with a vast dataset or analyze them. This feature of PySpark makes it a very demanding tool among data engineers.

Key features of PySpark

There are various features of the PySpark which are given below:



Real-time Computation

PySpark provides real-time computation on a large amount of data because it focuses on in-memory processing. It shows the low latency.

Support Multiple Language

PySpark framework is suited with various programming languages like Scala, Java, Python, and R. Its compatibility makes it the preferable frameworks for processing huge datasets.

Caching and disk constancy

PySpark framework provides powerful caching and good disk constancy.

Swift Processing

PySpark allows us to achieve a high data processing speed, which is about 100 times faster in memory and 10 times faster on the disk.

Works well with RDD

Python programming language is dynamically typed, which helps when working with RDD. We will learn more about RDD using Python in the further tutorial.

What is Apache Spark?

Apache Spark is an open-source distributed cluster-computing framework introduced by Apache Software Foundation. It is a general engine for big data analysis, processing and computation. It is built for high speed, ease of use, offers simplicity, stream analysis and run virtually anywhere. It can analyze data in real-time. It provides fast computation over the big data.

The fast computation means that it's faster than previous approaches to work with Big Data such as MapReduce. The main feature of Apache Spark is its in-memory cluster computing that enhances the processing speed of an application.

It can be used for multiple things like running distributed SQL, creating data pipelines, ingesting data into a database, running Machine Learning algorithms, working with graphs or data streams, and many more.

Why PySpark?

A large amount of data is generated offline and online. These data contain the hidden patterns, unknown correction, market trends, customer preference and other useful business information. It is necessary to extract valuable information from the raw data.

We require a more efficient tool to perform different types of operations on the big data. There are various tools to perform the multiple tasks on the huge dataset but these tools are not so appealing anymore. It is needed some scalable and flexible tools to crack big data and gain benefit from it.

Difference between Scala and PySpark

Apache Spark is officially written in the Scala programming language. Let's have a look at the essential difference between Python and Scala.

Sr.	Python	Scala
1.	Python is an interpreted, dynamic programming language.	Scala is a statically typed language.
2.	Python is Object Oriented Programming language.	In Scala, we need to specify the type of variable and objects.
3.	Python is easy to learn and use.	Scala is slightly difficult to learn than Python.
4.	Python is slower than Scala because it is an interpreted language.	Scala is 10 times faster than Python.
5.	Python is an Open-Source language and has a huge community to make it better.	Scala also has an excellent community but lesser than Python.
6.	Python contains a vast number of libraries and the perfect tool for data science and machine learning.	Scala has no such tool.

One of the most amazing tools that helps handle big data is Apache Spark. As we are familiar that Python is one of the most widely used programming languages among data scientist, data analytics and in various fields. Because of its simplicity and interactive interface, it is trusted by the data scientist folks to perform data analysis, machine learning and many more tasks on big data using Python.

So, the combination of the Python and Spark would be the very efficient for the world of big data. That's why Apache Spark Community came up with a tool called PySpark that is a Python API for Apache Spark.

Real-life usage of PySpark

Data is an essential thing for every industry. Most of the industries works on big data and hires analysts to extract useful information from the raw data. Let's have a look at the impact of the PySpark on several industries.

1. Entertainment Industry

The entertainment industry is one of the largest sectors which is growing towards online streaming. The popular online entertainment platform Netflix uses the Apache spark for real-time processing to personalized online movies or web series to its customers. It processes approx. 450 billion events per day that are streamed on server-side application.

2. Commercial Sector

The commercial sector also uses Apache Spark's Real-time processing system. Banks and other financial fields are using Spark to retrieve the customer's social media profile and analyze to gain useful insights which can help to make the right decision.

The extracted information is used for the credit risk assessment, targeted ads, and customer segmentation.

Spark plays a significant role in Fraud Detection and widely used in machine learning tasks.

3. Healthcare

Apache Spark is used to analyze the patient records along with the previous medical reports data to identify which patient is probable to face health issues after being discharged from the clinic.

4. Trades and E-commerce

The leading e-commerce websites like Flipkart, Amazon, etc, use Apache Spark for targeted advertising. The other websites such as Alibaba provides targeted offers, enhanced customer experience and optimizes overall performance.

5. Tourism Industry

The tourism industry widely uses Apache Spark to provide advice to millions of travelers by comparing hundreds of tourism websites.

Prerequisites

Before learning PySpark, you must have a basic idea of a programming language and a framework. It will be very beneficial if you have a good knowledge of Apache Spark, Hadoop, Scala programming language, Hadoop Distribution File System (HDFS), and Python.