List of libraries used in spark?

Pyspark, pandas, re, math library

PySpark has several unique features that separate it from Pandas, namely its ability to run operations in parallel and in distributed environments. A Pandas Dataframe is constrained in size by the memory of the server the application is executed on, which can limit the amount of data processed. A PySpark dataframe on the other hand can also act as a distributed SQL like query engine, enabling large amounts of data to be stored and read from a single dataframe, even in a distributed environment. This is especially useful for Big-Data applications.

Read this page:

[Introduction to PySpark and PySpark ML Libraries | by Kyle Kauffman | Medium](https://kylekauffman98.medium.com/introduction-to-pyspark-and-pyspark-ml-libraries-da4400545dc5)

List of packages used in pyspark?

Diff bw pandas DF and DF?

List, tuple, dictionaries, set with examples

JSON,XML, parquet, csv file formats

Lambda functions etc

If you are working with a smaller Dataset and don’t have a Spark cluster, but still you wanted to get benefits similar to Spark DataFrame, you can use Python pandas DataFrames. The main difference is pandas DataFrame is not distributed and run on a single node.

PySpark is a Spark library written in Python to run Python applications using Apache Spark capabilities, using PySpark we can run applications parallelly on the distributed cluster (multiple nodes).

In other words, PySpark is a Python API for Apache Spark. Apache Spark is an analytical processing engine for large scale powerful distributed data processing and machine learning applications

Spark basically written in Scala and later on due to its industry adaptation it’s API PySpark released for Python using Py4J. Py4J is a Java library that is integrated within PySpark and allows python to dynamically interface with JVM objects, hence to run PySpark you also need Java to be installed along with Python, and Apache Spark.