What is pyspark

PySpark is the Python API for Apache Spark. PySpark enables developers to write Spark applications using Python, providing access to Spark’s rich set of features and capabilities through Python language.

PySpark has become a popular choice for data engineers, data scientists, and developers working with big data and distributed computing.

Spark runs operations on billions and trillions of data on distributed clusters 100 times faster than traditional applications.

The main difference between pandas and pyspark is Pandas DataFrame is not distributed and runs on a single node.

PySpark 3.5 supports Python versions 3.8 and newer