Data Processing with PySpark

Presenter

Sanket Sharma

Agenda

1. What is PySpark?
2. Why PySpark?
3. Hands-on

1. What is PySpark?

Source:
https://databricks.com/glossary/pyspark

To be continued..

PySpark Overview

- Apache Spark was written in Scala
- PySpark was released in order to support the Apache Spark's capabilities using python
- In other words playing with ***big data*** at scale using the ***simplicity of python***

2. Why PySpark?

"A picture is worth a thousand words."

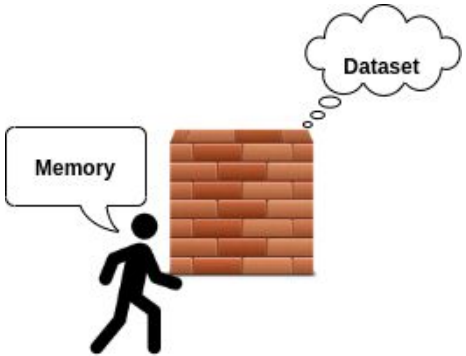

Fig 1. Loading dataset using *pandas*

Fig 2. Loading dataset using *PySpark*

Source: Created by *presenter*

Advantages of PySpark

- Distributed Computing
- Applications running on it are 100x times faster
- Supports SQL like queries

3. Hands-on

Colab Notebook

https://colab.research.google.com/drive/12rVmOKTzP67Ayu-ky0G8atUDseNnw2Wu?usp=sharing

Git Repo: https://github.com/xsansha/PySpark-Workshop

References

[1] What is PySpark? - Databricks (accessed Nov. 8, 2021)

[2] PySpark Tutorial for Beginners (accessed Nov. 8, 2021)