O que é o PySpark?

É uma ferramenta que podemos utilizar para **processar grandes quantidades de dados**. Ela tem esta capacidade porque permite paralelização.

A programação paralela é uma área da computação que nos permite dividir nosso processador em "outros pequenos processadores".

Assim, ao invés de executarmos linha a linha dos códigos que desenvolvemos, podemos executar de duas em duas, de quatro em quatro, de oito em oito, de acordo com a quantidade de pequenos processadores que definimos.

Para trabalhar com o PySpark **você não precisa entender de programação paralela**, a ferramenta já faz isso para você!

O que preciso saber para usar o PySpark?

Você precisa saber programar em Python de acordo com tudo que aprendeu durante o curso. Então se você chegou até aqui isso não será um problema.

Além disso, **é fundamental conhecer as funcionalidades que o PySpark nos possibilita**, por isso estamos disponibilizando aqui a documentação da fermenta. Divirta-se!

- **Documentação do PySpark:** https://spark.apache.org/docs/latest/api/python/user_guide/index.html
- **Site oficial do PySpark para implementações em Python:** https://spark.apache.org/docs/latest/api/python/