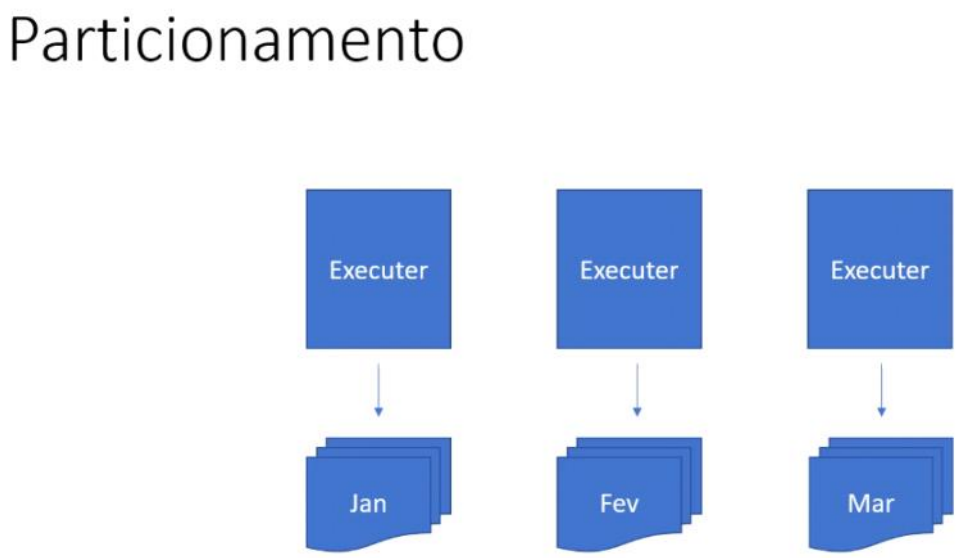
01.O que é Spark?

quinta-feira, 21 de setembro de 2023 14:06

O que é Spark?

- Ferramenta de processamento de dados (Não é um Data Storage)
- Distribuido em um cluster
- Em memória
- Veloz
- Escalável
- Dados em HDFS ou Cloud
- Particionamento

Característica útil do Spark é Replicação e Tolerância a falha onde os dados são copiados entre os nós do cluster. Isso traz o benefício de, entre outras coisas, tolerância a falhas.



Spark tem como principal objetivo:

- Você precisa processar dados
- Custo Computacional: CPU, Memória, Rede etc.
- Spark tem arquitetura voltada a processar dados
  - Melhor perfomace, porém:
  - Não substitui Python
  - Não substitui SQL ou um SGBDR

Linguagens utilizando SPARK

O Spark tem relação com o Hadoop, o Hadoop foi criado na epoca do Google File System(GFS), MapReduce(MR) e Bigtable.

Resultou em Hadoop, MapReduce, HDFS e Yarn

- Complexo
- Requer conhecimento em java
- Modelo em Batch em tarefas mapeamento e redução

Solução

- Hive criado pelo facebook - DW SQL sobre HDFS