WBA0753_v1.0

Linguagens de programação para ciência de dados (*Python* com *Spark*)

Processando *Big Data* com *Apache Spark*

Analisar B*ig Data* com *Spark.* Utilizar principais bibliotecas do *Spark*. Utilizar *Python* com *Spark.*

Bloco 1

Yuri Sá

Formatos de dados em *Big Data*

- Não estruturado.
- Semiestruturado.
- Estruturado.

Características de dados em *Big Data*

- Volume.
- Velocidade.
- Variedade.
- Veracidade.
- Valor.
- Validade.
- Variabilidade.
- Distribuição.
- Vocabulário.
- Indecisão.

Componentes do *Apache Spark*

- *SparkSQL*.
- *SparkStreaming*.
- *Spark Mllib*.
- *Spark GraphX*.

Conjuntos no *Python Spark*

- *DataFrame*:
  - Escalável.
  - Múltiplos formatos.
  - Fácil integração.
  - *Python*, Java, Scala e R.

Processando *Big Data* com *Apache Spark*

Analisar *BigData* com *Spark*. Utilizar principais bibliotecas do *Spark*. Utilizar *Python* com *Spark*.

Bloco 2

Yuri Sá

Criar um *DataFrame*

- Importar as classes de *pyspark*.
  - *SparkSession*.
  - *SparkContext*.
- Importar arquivo CSV.

Spark SQL

- Criar tabelas a partir de CSV ou JSON.
- SQL *Databases*.

Teoria em Prática

Bloco 3

Yuri Sá

Reflita sobre a seguinte situação

Utilizando o *framework Apache Spark*, como você processaria esses dados, de forma a extrair informações importantes para a empresa?

Norte para a resolução...

- Verifique a estrutura de origem dos dados.
- Analise se é possível um pré-processamento dos dados.
- Selecione a estrutura dentro do *spark*, que acomoda melhor os dados lidos.
- Preste muita atenção na eficiência do código.

Dica do (a) Professor (a)

Bloco 4

Yuri Sá

Monitore sempre a performance

- Para lidar com *BigData* é crucial economizar recursos da máquina, principalmente memória. Todos os dados serão replicados milhões e bilhões de vezes.

- Escolha os módulos e bibliotecas que sirvam com precisão, economizando o máximo possível de recursos da máquina.

Referências

CHAMBERS, B.; ZAHARIA, M. Spark: **The definitive guide: Big Data Processing Made Simple**. San Francisco: O'Reilly Media, 2018.

FOSTER, J. **What is Big Data? A Beginner's Guide to the World of Big Data**. Data Driver Investor, 2019.

Bons estudos!