WBA0753_v1.0

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

Real Time Analytics com *Python* e *Spark*

Analisar dados em tempo real com *Spark*. API *Spark Streaming*. *Spark*, *Kafka* e *Elasticsearch*.

Bloco 1

Yuri Sá

Processamento de dados em *Batch*

- Dados armazenados.
- Banco de dados.
- Menos criticidade.
- Maior detalhamento.

Processamento de dados em *Streaming*

- Dados transientes.
- Origem imediata.
- Maior criticidade.
- Menores lotes.

Apache Spark Streaming API - Fases

- *Gathering:*
  - *Basic Sources.*
  - *Advanced sources.*

- *Processing.*
- *Data Storage.*

Apache Spark Streaming API - RDD

- *In-Memory Computations.*
- *Lazy Evaluation.*
- *Fault Tolerant.*
- *Immutability.*
- *Partitioning.*
- *Persistence.*
- *Coarse-Grained Operations.*

Apache Kafka

- Processamento de mensagens:
  - Mensagens.
  - Tópicos.
  - *Producer.*
  - Leitura dos dados.

Elasticsearch

- Busca em grandes volumes.
- Tempo real.
- Alto desempenho.
- *API REST.*

Real Time Analytics com *Python* e *Spark*

Analisar dados em tempo real com *Spark*. API *Spark Streaming*. *Spark*, *Kafka* e *Elasticsearch*.

Bloco 2

Yuri Sá

Extrair e contabilizar *hashtags* do *Twitter*, em tempo real

- Conectar na API da rede social.
- Criar código base em *Python*.
- Filtrar as *hashtags*.
- Contabilizar e armazenar.

Teoria em Prática

Bloco 3

Yuri Sá

Reflita sobre a seguinte situação

- Processar arquivos de *log* em tempo real, com *PySpark*.

Norte para a resolução...

- Abrir um arquivo, manualmente, e analisar a estrutura básica.
- Separar as classes úteis para contabilizar no modelo.
- Fazer as análises.

Dica do (a) Professor (a)

Bloco 4

Yuri Sá

Mantenha um ciclo de dados saudável – *Buffer*

- Crie um *buffer* que organize a fila de dados, para que não tenha grandes saltos de dados por vez.
- Isso mantém o servidor e as conexões com dados relativamente estáveis.
- Torna a performance previsível.
- Ajuda a escalar o modelo.

Referências

HANEE’, M. **Apache Spark streaming tutorial: i**dentifying Twitter trending hashtags. Toptal Project, 2017.

Bons estudos!