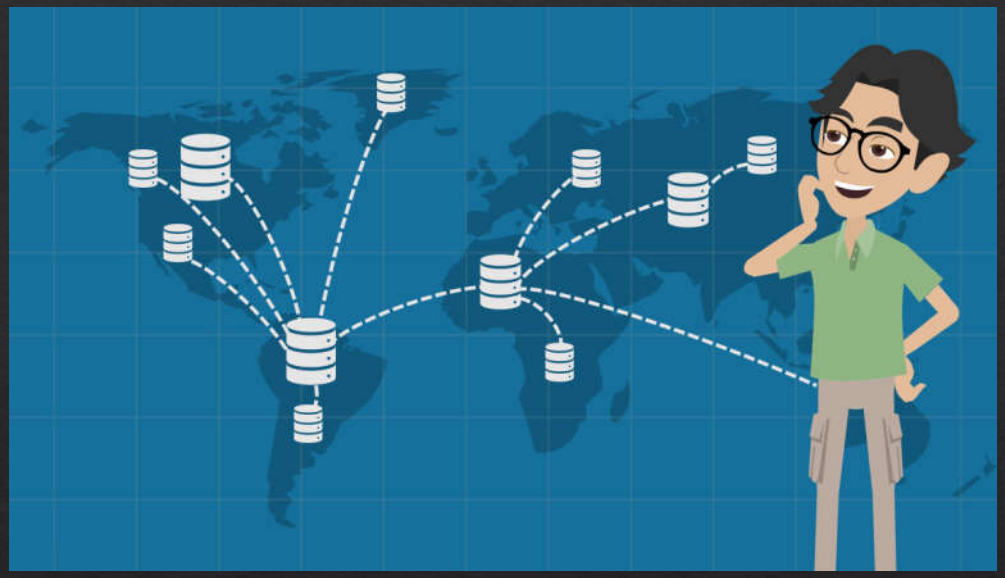

Formação Engenheiro de Dados

Spark Streaming

PROF. FERNANDO AMARAL
www.datascientist.com.br

O que é Streaming

- Dados processados de forma contínua
- Em tempo real, ou próximo ao tempo real
- Exemplos:
  - Dados de sensores
  - Logs de acesso (detecção de invasão)
  - Transações financeiras (busca de fraudes)

Spark Streaming

DStream

Representação interna do stream

Composto por RDDs contínuos

Cada RDD possui dados de um intervalo

Fontes de Stream

- Básicas: Disponíveis diretamente (sem precisar importar bibliotecas). Pastas e Conexão TCP
- Avançadas: Precisam de bibliotecas extras: Ex: Flume.

Processo Básico

- Define o contexto da aplicação, com threads e nome:

```
sc = SparkContext("local[2]", "AppTeste")
```

- Contexto de Streaming e Intervalo do batch

```
ssc = StreamingContext(sc, 1)
```

- Define a fonte de dados

```
streamingContext.textFileStream(diretorio)
```

- Aplica transformações
- Inicia o processamento

```
streamingContext.start()
```

- Aguarda Encerramento

```
streamingConext.awaiTermination()
```

textFileStream

O arquivo é lido do diretório

Um mesmo arquivo não é lido novamente, mesmo que modificado

Deve ser incluído novo arquivo ou renomeado