SPARK – DATA FRAMES

Big Data

ESTRUTURA DE APIS DO SPARK

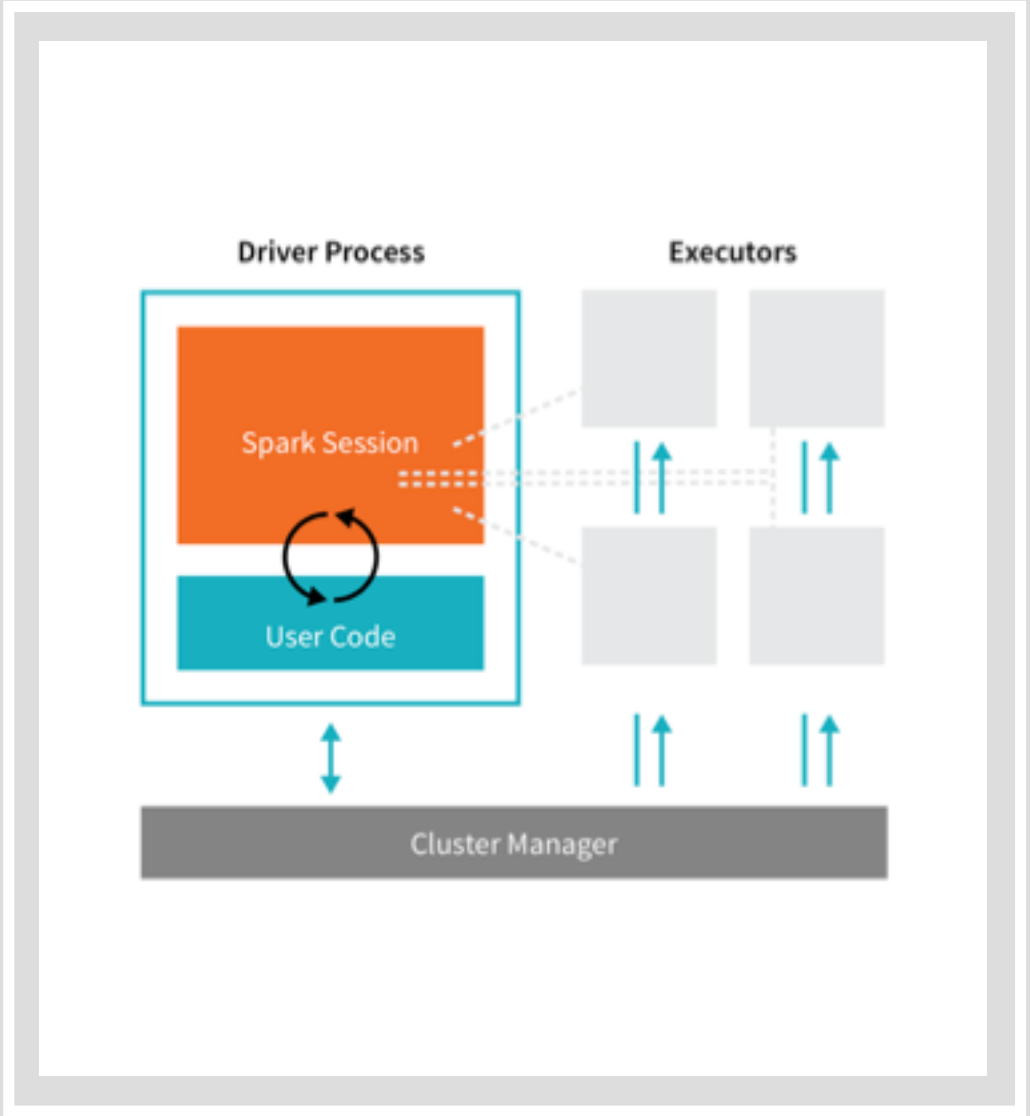
- Foco da Disciplina
  - Advanced Analytics (R, Scala, Python) utilizando DataFrames e SQL
  - Datasets – mais específico para Scala e Java

LINGUAGENS

Scala	Linguagem na qual o Spark é escrito (grande maioria) Recursos da plataforma estão sempre disponíveis nesta linguagem
Java	Apesar de Spark ser escrito em Scala, há grande preocupação nas classes funcionarem com Java
Python	Linguagem com bastante suporte, exceto pela falta de tipagem estática.
SQL	Suporte ao ANSI SQL 2003 Voltada para não programadores e analistas de dados

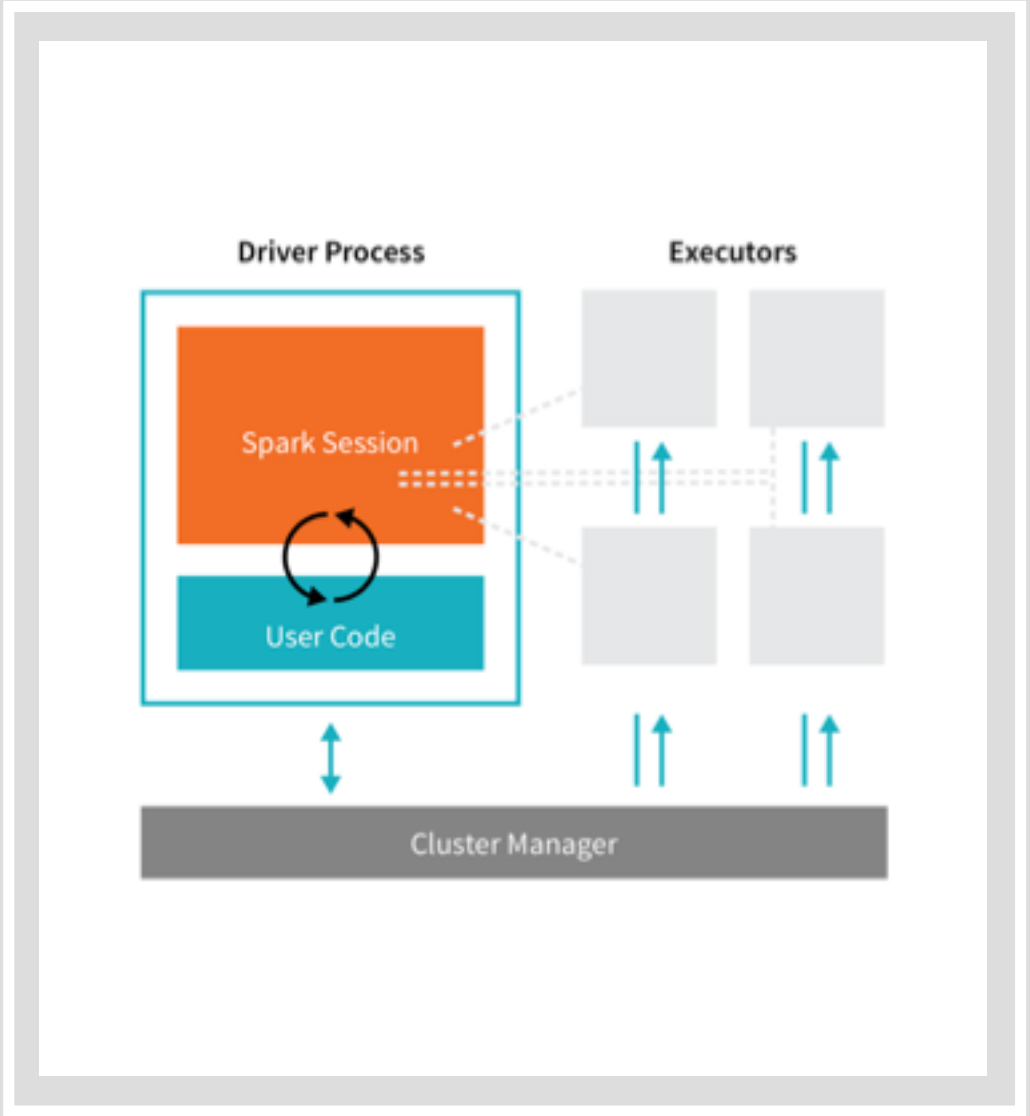
ARQUITETURA

- Spark Applications
  - Processo *driver* e conjunto de processos *executor*
- *Driver Process* – rotina principal
  - *script, main method*
  - Suporta múltiplas linguagens
  - Mantém informações sobre o Spark Application
  - Responde às entradas ou operações dos usuários
  - Analisa, distribui e agenda cargas de trabalho nos *executors*



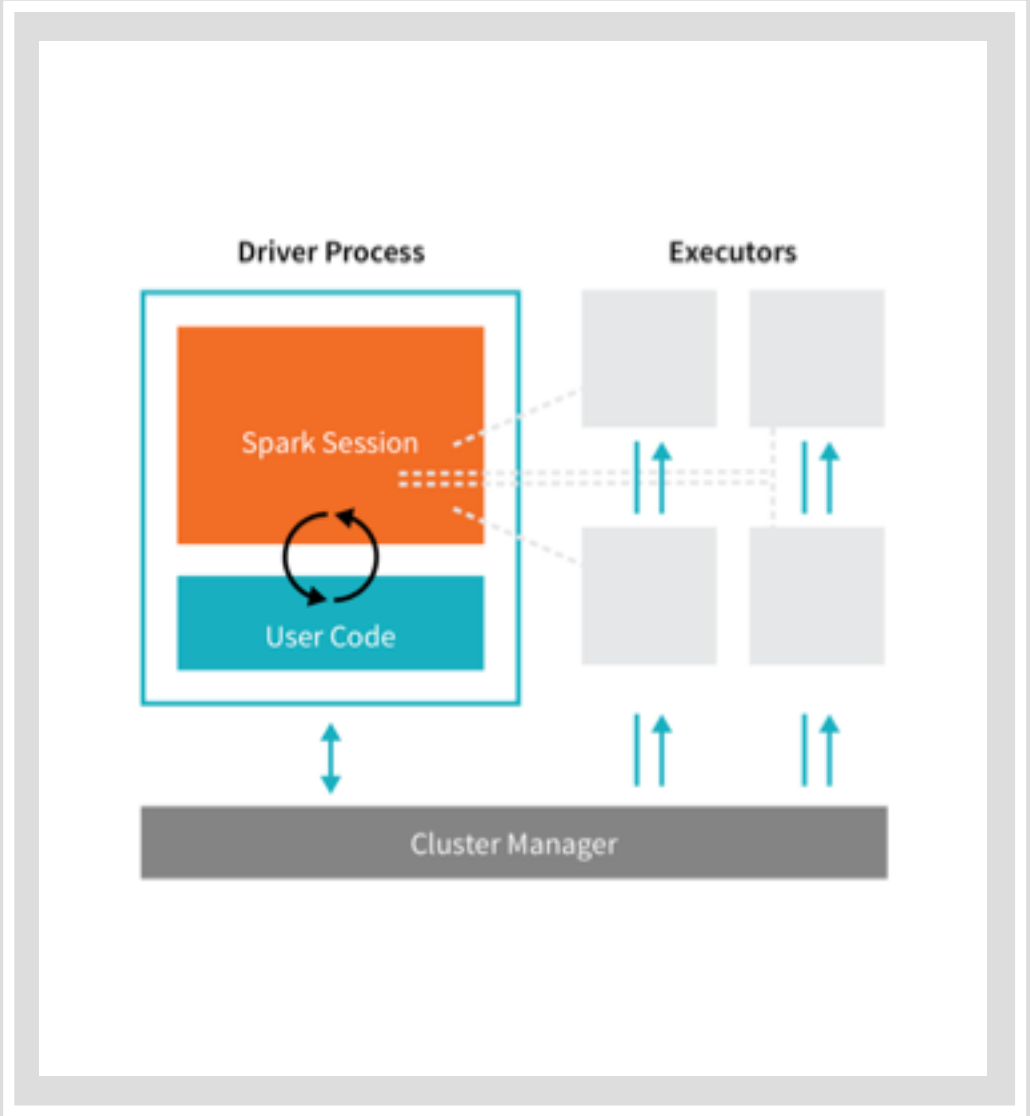
ARQUITETURA

- *Executors*
  - Executam códigos Spark que desempenham as cargas atribuídas pelo *driver*
- Cada *executor:*
  - Executa o código distribuído pelo *driver*
  - Reporta o estado da processamento para o *driver*



ARQUITETURA

- *Cluster Manager*
  - Controla os servidores e aloca os recursos para processamento das *Spark Applications*
- Alternativas
  - Spark Standalone Cluster Manager
  - YARN
  - Mesos
- Um *cluster* suporta múltiplas aplicações Spark



SPARK SHELL

Python	pyspark
Scala	spark-shell
SQL	spark-sql

SPARK SESSION

- SparkSession é a interface entre os usuários e o processo *Driver*
- É por onde o Spark executa no cluster as operações definidas pelo usuário
- Cada *Spark Application* possui um SparkSession, que é único para cada *Application*
- Scala/Python
  - Objeto chamado *spark*

SPARK SESSION

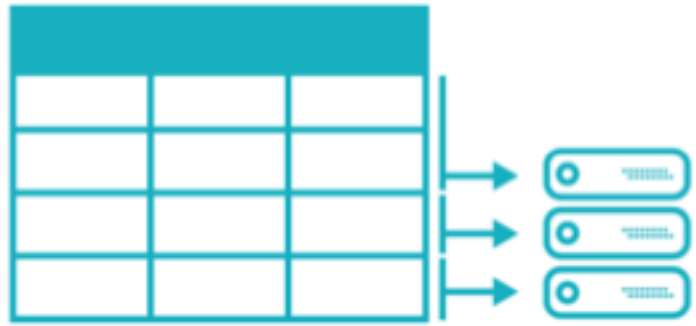
- Abrir o spark-shell
- Escrever
  - spark
- Escrever
  - val myRange = spark.range(1000).toDF("number")

SPARK SESSION

- Abrir o pyspark
- Escrever
  - spark
- Escrever
  - myRange = spark.range(1000).toDF("number")

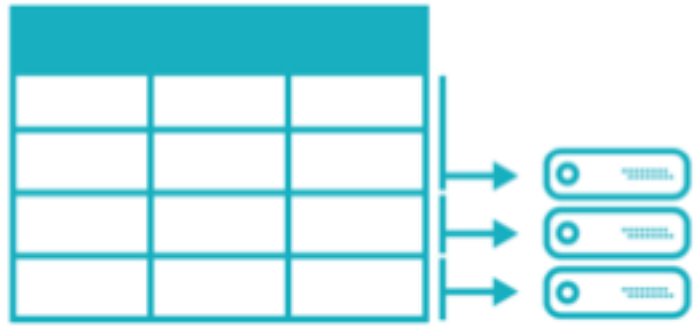
DATA FRAMES

- Forma mais comum da Structured API
- Representa tabela com linhas e colunas
- *Schema:* lista de nomes das colunas e seus tipos
- DataFrame é particionado e dividido pelos servidores que compõe o *cluster*
- Razões para esta distribuição:
  - Não caber em uma única máquina
  - Processamento demoraria demais em uma única máquina



PARTIÇÕES

- Divisão dos dados do *DataFrame* em blocos
- Partição é uma coleção de linhas (registros) que estão armazenadas em um nodo do cluster.
  - Não ocorre particionamento por colunas
- *DataFrame* possui uma lista que representa como as partições estão distribuídas no cluster
- Quantidade de partições define o grau de paralelismo durante a execução
- Nas Structured APIs não é comum manipular as partições diretamente

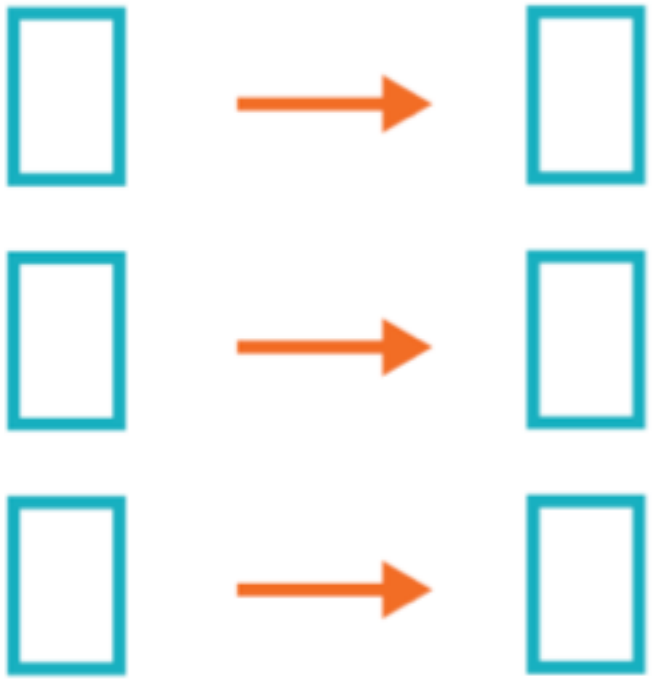


TRANSFORMATIONS

- Data Frames são imutáveis
- Transformações são funções que retornam um novo Data Frame resultante de uma modificação
  - Exemplos: Filtros (*where, filter*), Projeções (*select*), Modificações e criações de colunas (*select, map*), Joins, Agrupamentos (*groupBy*), etc
  - *Scala:* val divisBy2 = myRange.where("number % 2 = 0")
  - *Python:* divisBy2 = myRange.where("number % 2 = 0")
- *Lazy Evaluation:* Transformações somente são efetivadas quando uma *action* ocorrer (em breve)

NARROW TRANSFORMATIONS

Narrow Transformations
1 to 1



WIDE TRANSFORMATIONS

Wide Transformations (shuffles)
1 to N

ACTIONS

- Spark constrói plano lógico de execução a partir da combinação das diferentes transformações
- *Actions* determinam a computação (processamento) do resultado a partir da sequência de transformações
- Existem 3 tipos de ações
  - Visualizar dados na console (*show*)
  - Coletar os dados do resultado em estruturas da linguagem de programação do driver (*collect, take*)
  - Persistir os dados (*write*)
- Scala / Python: divisBy2.count()
  - A contagem de registros ao final de uma série de transformações também é uma action.
  - Que tipo de transformações são processadas nesta action?

EXEMPLOS