Spark
DataFrames - Introdução

2023

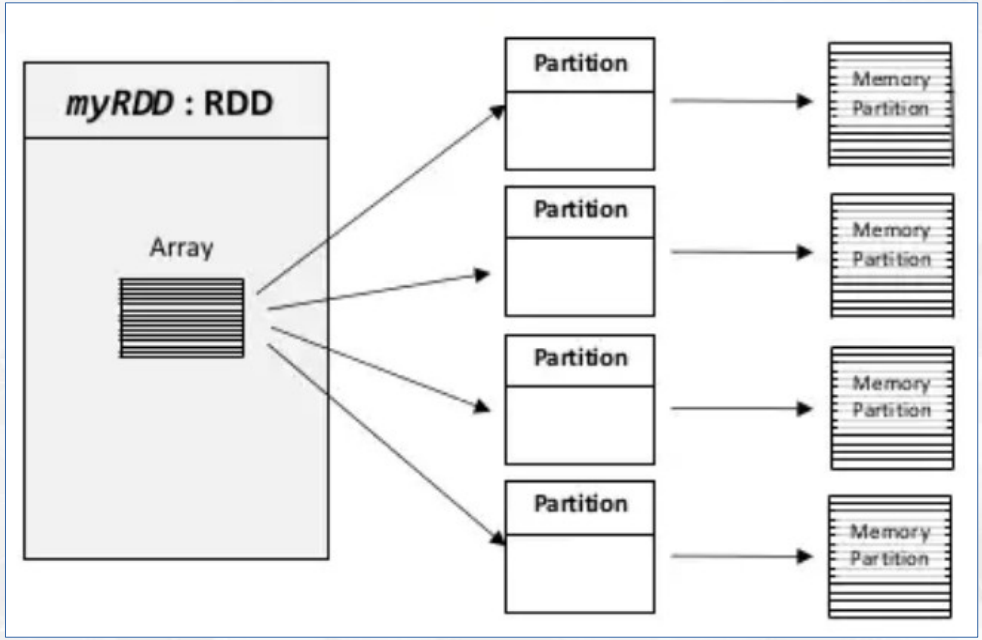
<lapti>

Agenda

- APIs Estruturadas
- Problemas com RDDs
- DataFrame API
- Tipos
- Operações comuns em DataFrames

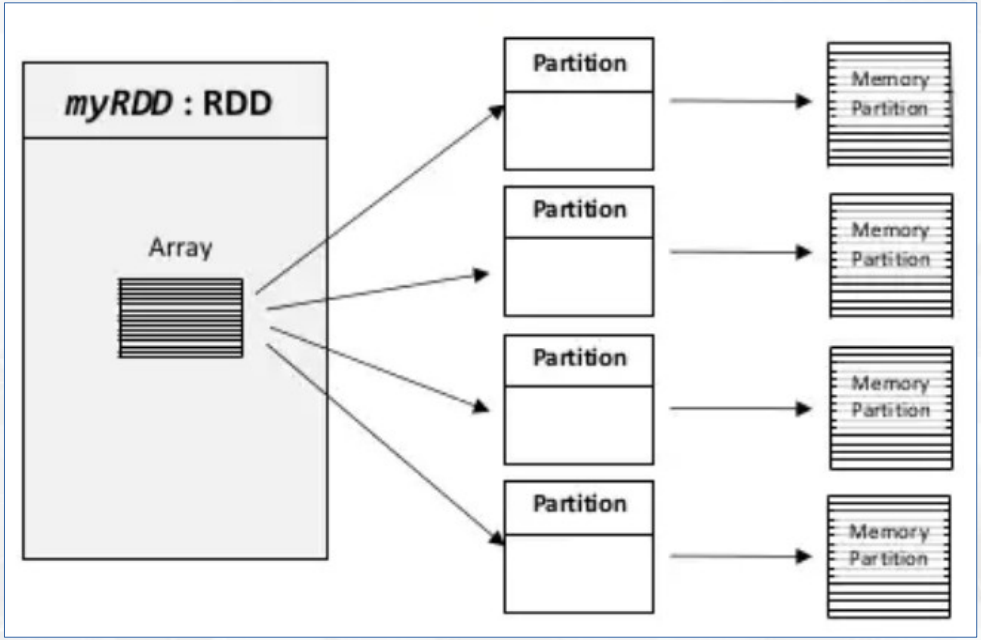
APIs de Dados Estruturados

- RDD é a abstração mais básica em Spark
  - RDDs, SchemaRDDs
- Três características principais
  - Dependências
  - Partições
    - Localidade
  - Função de execução
    - Partition => Iterator[T]



APIs de Dados Estruturados

- Problemas em RDDs…
  - Função de execução é opaca para Spark
    - Compute function
    - Spark vê apenas uma função lambda
  - Iterator[T] não tem tipos de dados expostos pelas APIs
    - Objeto genérico
      - Principalmente em Python
  - Opacidade diminui chance de otimização



Estruturando o Spark

- Spark 2.x
  - Operações de alto nível
    - Filtragem, seleção, contagem, agregação, etc
  - Operadores e funções comuns em uma DSL
    - Domain Specific Language
    - Exposta via APIs
  - Schemas e dados estruturados
    - Formato tabular
  - Mais desempenho, melhor eficiência em otimização

DataFrame API

- Inspirada por API semelhante em Pandas
  - Tabelas distribuídas in-memory
    - Colunas nomeadas e schemas
    - Colunas podem ter tipos específicos
      - Integer, string, array, map, real, date, timestamp, etc
- DataFrames são imutáveis
  - Torna distribuição e desempenho melhores
- Tipos de colunas podem ser deduzidos ou declarados
  - Inferência de schema ou declaração

Tipos básicos

Data type	Value assigned in Scala	API to instantiate
ByteType	Byte	DataTypes.ByteType
ShortType	Short	DataTypes.ShortType
IntegerType	Int	DataTypes.IntegerType
LongType	Long	DataTypes.LongType
FloatType	Float	DataTypes.FloatType
DoubleType	Double	DataTypes.DoubleType
StringType	String	DataTypes.StringType
BooleanType	Boolean	DataTypes.BooleanType
DecimalType	java.math.BigDecimal	DecimalType

Data type	Value assigned in Python	API to instantiate
ByteType	int	DataTypes.ByteType
ShortType	int	DataTypes.ShortType
IntegerType	int	DataTypes.IntegerType
LongType	int	DataTypes.LongType
FloatType	float	DataTypes.FloatType
DoubleType	float	DataTypes.DoubleType
StringType	str	DataTypes.StringType
BooleanType	bool	DataTypes.BooleanType
DecimalType	decimal.Decimal	DecimalType

Tipos estruturados

Data type	Value assigned in Scala	API to instantiate
BinaryType	Array[Byte]	DataTypes.BinaryType
TimestampType	java.sql.Timestamp	DataTypes.TimestampType
DateType	java.sql.Date	DataTypes.DateType
ArrayType	scala.collection.Seq	DataTypes.createArrayType(ElementType)
MapType	scala.collection.Map	DataTypes.createMapType(keyType, valueType)
StructType	org.apache.spark.sql.Row	StructType(ArrayType[fieldTypes])
StructField	A value type corresponding to the type of this field	StructField(name, dataType, [nullable])

Tipos estruturados

Data type	Value assigned in Python	API to instantiate
`BinaryType`	`bytearray`	`BinaryType()`
`TimestampType`	`datetime.datetime`	`TimestampType()`
`DateType`	`datetime.date`	`DateType()`
`ArrayType`	List, tuple, or array	`ArrayType(dataType, [nullable])`
`MapType`	`dict`	`MapType(keyType, valueType, [nullable])`
`StructType`	List or tuple	`StructType([fields])`
`StructField`	A value type corresponding to the type of this field	`StructField(name, dataType, [nullable])`

Operações comuns em DataFrames

- Leitura e carregamento de DataFrames
  - Interface DataFrameReader
    - Exposta por diferentes métodos nas APIs específicas
  - Vários formatos de entrada
    - JSON, CSV, Parquet, Text, Avro, ORC, etc
- Gravação de DataFrames
  - Exportação de um DataFrame em diferentes formatos
  - Interface DataFrameWriter

Operações comuns em DataFrames

- Leitura e carregamento de DataFrames
  - Interface DataFrameReader
    - Exposta por diferentes métodos nas APIs específicas
  - Vários formatos de entrada
    - JSON, CSV, Parquet, Text, Avro, ORC, etc
- Gravação de DataFrames
  - Exportação de um DataFrame em diferentes formatos
  - Interface DataFrameWriter

```
df = spark.read.csv('dados.csv', header = True, inferSchema = True)

df.write.format('parquet').saveAsTable('tabela.parquet')
```

Operações comuns em DataFrames

- Transformações e Ações
  - Projeções e filtros
    - select()
    - filter() ou where()
    - Modificadores como distinct(), limit()
  - Renomear, excluir e adicionar colunas
    - drop(), withColumnRenamed(), withColumn()
    - Operações para criação de dados em select()

Operações comuns em DataFrames

- Transformações e Ações
  - Manipulação de datas
    - Definir campos data no schema
    - to_timestamp(), to_date()
    - month(), year(), day()
  - Agregações
    - groupBy, orderBy, count()
  - Funções
    - min(), max(), sum(), avg()

Funções em
spark.sql.functions

Operações comuns em DataFrames

- Junção de DataFrames
  - join()
- Funções estatísticas
  - stat(), describe(), correlation(), covariance(), sampleBy(), approxQuantile(), frequentItems(), etc
- Muitas das funções já existiam em RDDs
  - Mas devido a opacidade dos dados internos, não era possível otimizar a sua execução

Obrigado

leandro@utfpr.edu.br

<lapti>