CURSO DE CIÊNCIA DE DADOS APLICADA AO PODER JUDICIÁRIO

SPARK PARA CIÊNCIA DE DADOS

Semana 5 - Apache Spark - SparkSQL

PROF. CARLOS M. D. VIEGAS

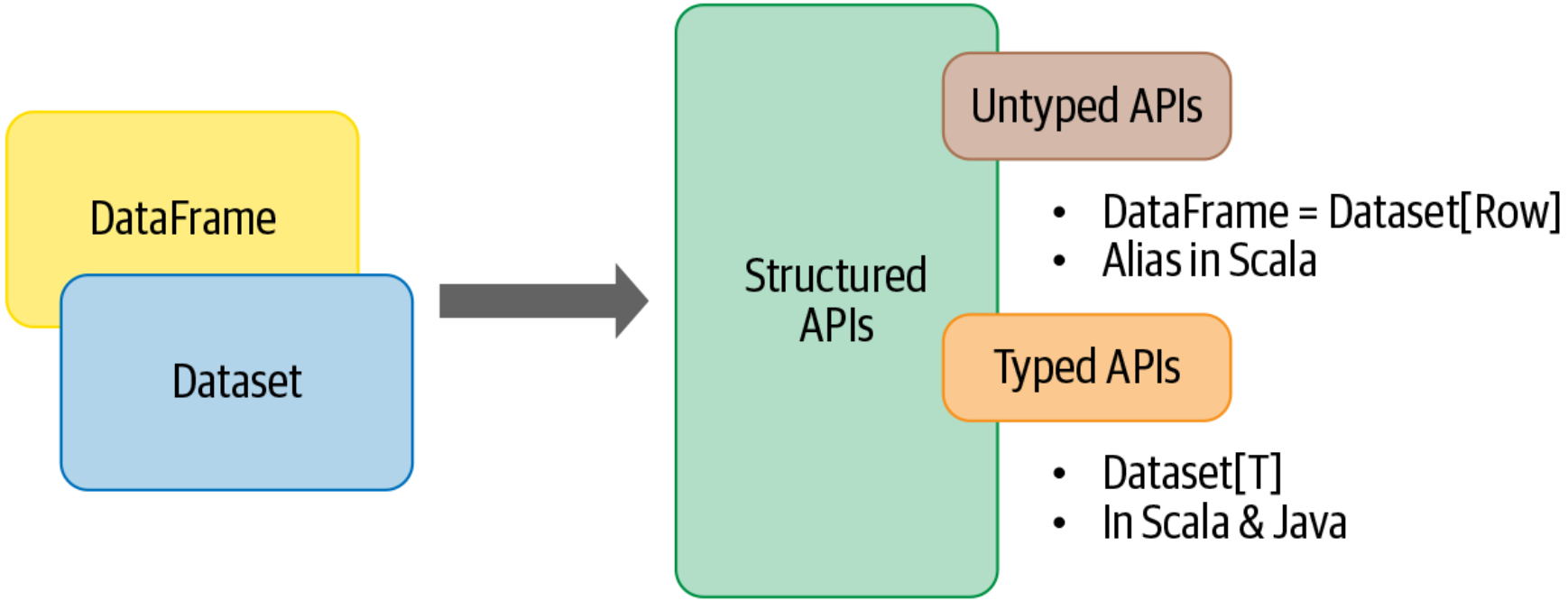
Semana 5 - Apache Spark - SparkSQL

- Conteúdo

1. Spark Dataset
2. Programação com pySpark: SparkSQL
3. Manipulação de dados com SparkSQL
4. Aplicações práticas

Spark Dataset

- Dataset
  - É uma abstração de dados que combina as características do RDD e do Dataframe
    - São coleções de dados distribuídas
    - São projetados para trabalhar com dados estruturados e semiestruturados
  - Suporta apenas as linguagens Java ou Scala



Spark Dataset

- Os dados armazenados em Datasets podem ser “tipados”
  - O tipo do objeto é definido no momento da criação do Dataset:
    - Dataset[Tipo]
  - Permite executar verificações de tipo em tempo de compilação em vez de tempo de execução, o que pode ajudar a detectar erros de programação mais cedo
  - Exemplo:

```
case class Tribunal(nome: String, id: Int, tipo: String)

val tribunais: Dataset[Tribunal] = Seq(
        Tribunal("TRE", 15, "Eleitoral"),
        Tribunal("TRT", 10, "Trabalho"),
        Tribunal("STF", 22, "Federal")
).toDS()
```

Spark Dataset

- Ou os dados armazenados em Datasets podem ser não-tipados
  - Representa uma tabela distribuída de dados com colunas nomeadas
    - É semelhante a um DataFrame em Python
    - Dataset[Row]
  - Exemplo:

```
val schema = StructType(
  List(
    StructField("nome", StringType, true),
    StructField("id", IntegerType, true),
    StructField("tipo", StringType, true)
  )
)
val rows = Seq(
  Row("TRE", 15, "Eleitoral"),
  Row("TRT", 10, "Trabalho"),
  Row("STF", 22, "Federal")
)
val ds = spark.createDataFrame(spark.sparkContext.parallelize(rows), schema)
```

Spark Dataset

- Dataset[T] vs Dataset[Row]
  - A principal vantagem de se usar Dataset[T] é a tipagem forte dos dados
  - Dataset[T] fornece uma API mais rica e intuitiva para manipulação de dados em comparação com Dataset[Row]

Spark Dataset

- Programação com Datasets
  - Vamos praticar consultando o arquivo:
    - `1 Primeiros exemplos - Datasets.txt`
  - Sugestão de documentação para consulta:
    - https://spark.apache.org/docs/latest/quick-start.html
    - https://spark.apache.org/docs/latest/sql-getting-started.html
    - Livro:

      Learning Spark: Lightning-Fast Data Analytics, 2nd edition de Jules Damji, Brooke Wenig, Denny Lee, Tathagata Das

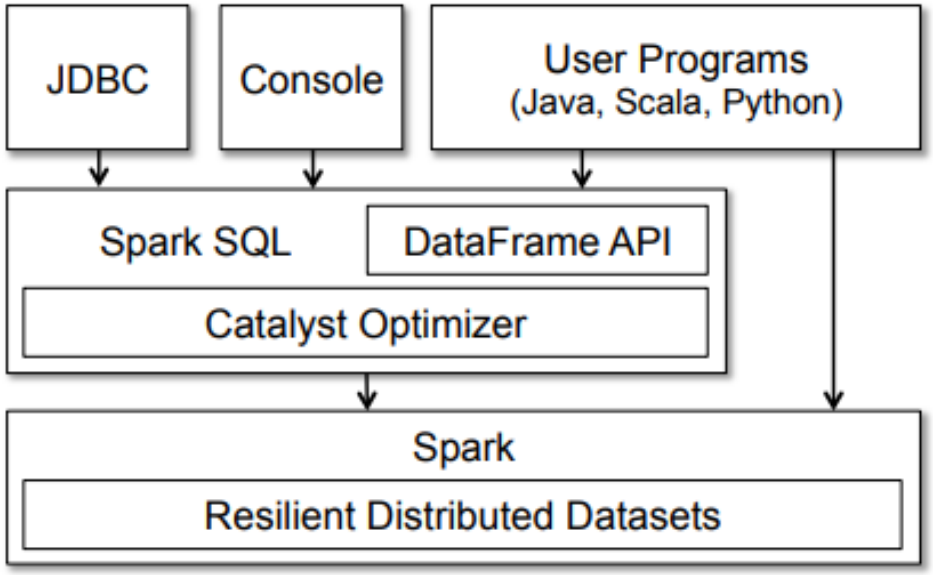
Semana 5 - Apache Spark - SparkSQL

- Conteúdo

1. Spark Dataset
2. Programação com pySpark: SparkSQL
3. Manipulação de dados com SparkSQL
4. Aplicações práticas

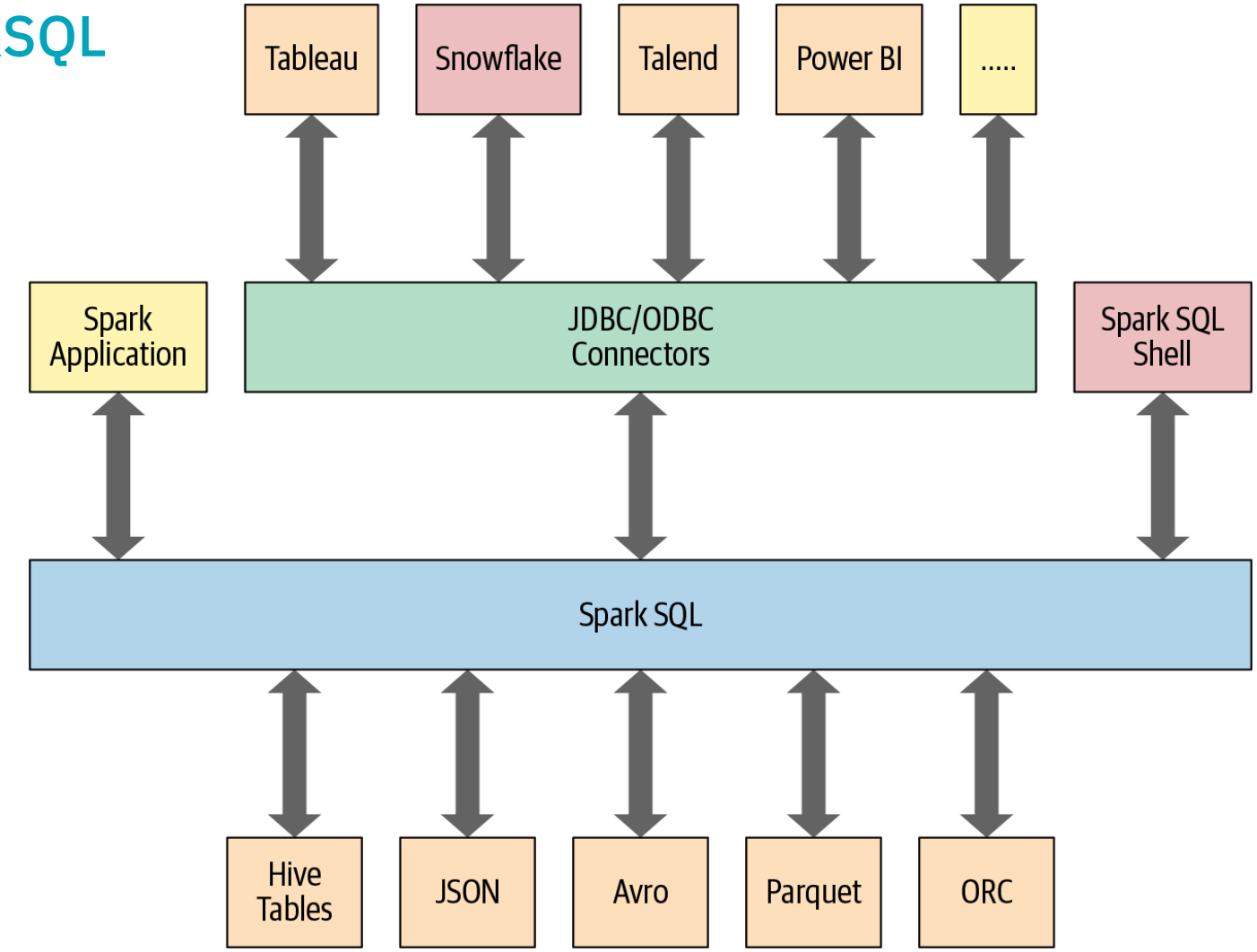
SparkSQL

- É um módulo do Apache Spark que permite trabalhar com dados estruturados usando SQL
  - Fornece uma interface programática para interagir com dados estruturados
  - Suporta consultas SQL, incluindo joins entre tabelas, agregações e subconsultas
  - Integra-se a outros módulos do Spark, tais como Spark Streaming e Mllib
  - Suporta diversas fontes de dados, incluindo CSV, JSON, Parquet e bancos de dados relacionais



SparkSQL

- Conectores SparkSQL



SparkSQL

- Tabelas vs Dataframes
  - Tabelas persistem
  - Objeto tabular reside em um banco de dados
  - Pode ser gerenciado e consultado utilizando linguagem SQL
  - Interoperável com DataFrame
    - É possível transformar um DataFrame em Tabela e vice-versa

- Tabelas podem ser:
  - Gerenciadas
    - Spark gerencia os dados e os metadados por meio do Apache Hive (ou outra solução de Data Warehouse)
    - Os dados ficam armazenados no cluster
  - Não gerenciadas
    - Spark gerencia apenas os metadados
    - Os dados são armazenados em um banco de dados externo ao cluster

SparkSQL

- Programação com SparkSQL
  - Vamos praticar consultando o arquivo:
    - `2 Primeiros exemplos - SparkSQL.txt`
  - Sugestão de documentação para consulta:
    - https://spark.apache.org/docs/latest/sql-ref.html
    - Livro:

      Learning Spark: Lightning-Fast Data Analytics, 2nd edition de Jules Damji, Brooke Wenig, Denny Lee, Tathagata Das
    - Livro:

      Data Algorithms with Spark: Recipes and Design Patterns for Scaling Up Using Pyspark, 1st edition de Mahmoud Parsian

OBRIGADO

CONTATO: viegas@dca.ufrn.br

CURSO DE CIÊNCIA DE DADOS APLICADA AO PODER JUDICIÁRIO

JUSTIÇA 4.0

PNUD

CNJ CONSELHO NACIONAL DE JUSTIÇA