



Dataframe

Tabelas com linhas e colunas

Imutáveis

Com schema conhecido

Linhagem preservada

Colunas podem ter tipos diferentes

Existem análises comuns: Agrupar, ordenar, filtrar

Spark pode otimizar estas análises através de planos de execução



Lazy Evaluation

O processamento de transformação de fato só ocorre quando há uma Ação: Lazy Evaluation

Tipos de Dados



Tipo
ByteType
ShortType
IntegerType
LongType
FloatType
DoubleType
DecimalType
StringType
BinaryType
BooleanType
TimestampType
DateType
ArrayType
MapType
StructType
StructField

Schema

- Você pode deixar para o Spark inferir a partir de parte dos dados ou
- Você pode definir o schema
- Definir tem vantagens:
 - Tipo correto
 - Sem overhead