Column:列操作

数学计算：四则（+-\*%）、比较、逻辑

排序：升/降序(desc,asc, 空值在前或后的desc、asc)

是否为空：isNaN、isNull

是否在集合内：isin————————（选择指定多列时有用）

like——模糊匹配

筛选——————where、otherwise(类似于case when then)

**over**(window: [WindowSpec](http://spark.apache.org/docs/latest/api/scala/org/apache/spark/sql/expressions/WindowSpec.html)):操作window函数，分组排序、取值时常用（topN,第k大）

字符型：匹配（startsWith/endsWith） 、截取（substr）、包含（contains）

反运算（取非操作）：**unary\_!、unary\_-**

**类型转换：as、cast**

**ColumnName：方便构造schema**

**除支持Column对象的各种操作：**

**构造schema:java原生类型（int,long,boolean等）**

**支持3种容器(hive中也提供的):**

**array**(dataType: [DataType](http://spark.apache.org/docs/latest/api/scala/org/apache/spark/sql/types/DataType.html)): [StructField](http://spark.apache.org/docs/latest/api/scala/org/apache/spark/sql/types/StructField.html)

**struct**(fields: [StructField](http://spark.apache.org/docs/latest/api/scala/org/apache/spark/sql/types/StructField.html)\*): [StructField](http://spark.apache.org/docs/latest/api/scala/org/apache/spark/sql/types/StructField.html)

**map**(keyType: [DataType](http://spark.apache.org/docs/latest/api/scala/org/apache/spark/sql/types/DataType.html), valueType: [DataType](http://spark.apache.org/docs/latest/api/scala/org/apache/spark/sql/types/DataType.html)): [StructField](http://spark.apache.org/docs/latest/api/scala/org/apache/spark/sql/types/StructField.html)

**DataFrame的读写**

[DataFrameReader](http://spark.apache.org/docs/latest/api/scala/org/apache/spark/sql/DataFrameReader.html)， 由SparkSession.read生成

[DataFrameWriter](http://spark.apache.org/docs/latest/api/scala/org/apache/spark/sql/DataFrameWriter.html) 由SparkSession.write生成

[DataFrameReader](http://spark.apache.org/docs/latest/api/scala/org/apache/spark/sql/DataFrameReader.html)

支持以下几种文件格式，每种基本都是多态：

jdbc——传url

传路径

csv

json

orc

parquet

text

textFile

table

也可以使用format传入“格式”，参数为枚举的String类型

多态下，也可以用option方法，传入k/v对，传值给要读的数据格式

[DataFrameWriter](http://spark.apache.org/docs/latest/api/scala/org/apache/spark/sql/DataFrameWriter.html):

除了写上述格式的文件，多出保存模式：

mode:枚举值————追加、覆盖、忽略、报错等4种

SortBy:排序

[DataFrameStatFunctions](http://spark.apache.org/docs/latest/api/scala/org/apache/spark/sql/DataFrameStatFunctions.html)————常用统计函数

一般是：取几列，计算出1个结果（如corr）或1个新的DataFrame（如crosstab）

**approxQuantile:精确分位，对多列，以给定1个的Array[概率值]，返回多列的Array[分位值]**

**bloomFilter:布隆过滤器，作用——统计某值是否在某集合内**

corr:相关系数

cov:协方差值

crosstab:二维交叉频数表——————指定两列（类似于pivot，但只返回count值，是其特例）

freqItems:频繁项集，默认minSupport为0.1

sampleBy:抽样——————————按比例随机抽样（传比例、随机种子）

**countMinSketch——————粗略估计 某元素出现的“频数”——即粗略估计“计数”**