Pyspark Cheat Sheet

by mitcht via cheatography.com/50563/cs/14121/

Read / Write .csv

```
df =
(sqlContext.read.format("com.databricks.spark.csv")\
       .option("header", "true")\
       .option("inferschema", "true")\
       .option("mode", "DROPMALFORMED")\
       .load("hdfs://file.csv"))
df.write.mode('overwrite').
         option("header", "true").
         csv("file://filename.csv")
```

Meta Data

```
df.printSchema()                    df.count()
len(df.columns)                     df.columns
df.dtypes
```

Arrange

```
df.withColumnRenamed("col","newcol")
df.orderBy(['var1', 'var2'], ascending = [True,
False])
df.orderBy(df.var1, df.var2.desc())
```

Filter

```
df.filter(df.var > 10000)
df.select('col1', 'col2')
df[collist] # collist = ['var1', ...]
df.head(5) / df.show(5, truncate=True) / df.take(5)
df.drop('var')
df.distinct()
df.dropDuplicates()
df.dropna(subset='var') / df.na.drop()
df.isNotNull()
df.var.isin("level1", "level2")
df.var.like("string")
df.var.startswith("m") / df.var.endswith("m")
df.sample(False with replacement, 0.5 fraction,
12345 seed)
```

Useful Functions

```
.describe('optional_var')    .count()
.show()                      .fillna(value)
.min() / .max()              .mean()
.stdev() / .variance()       .substr(1,3)
F.when(df.var > 30, "Y").otherwise("N")
df.var.alias('newvar') # used to rename something
df.cache() /                 .cast('Double')
df.unpersist()
.replace(10, 20) # ????      .na.fill(0, subset =
                             'var')
time()
from pyspark.sql import functions as F
```

Write Functions / UDF

```
from pyspark.sql.functions import udf
F1 = udf(lambda x: '-1' if condition else x,
StringType()) # NB return type
df = df.withColumn('newvar', F1(df['invar']))
```

Applying Functions

```
df.select('val').map(lambda x: x*2)
```

Summarise

```
df.crosstab('col1', 'col2') # pair-wise count
df.groupby('var').function()
df.groupby('var').agg({'val' : 'mean'})
```

Join

```
df3 = df1.join(df2, [df1.var1 == df2.var1, df1.var2 ==
df2.var2], 'left')
xtra = df1.select('var').subtract(df2.select('var')) #
anti-join
```

Method Chaining

```
df
      .select("col1","col2","col3", ...)
      .filter(df.col1 > 30 )
      .show()
```

Cheatography

Pyspark Cheat Sheet

by mitcht via cheatography.com/50563/cs/14121/

New Variable / Column

```
df.withColumn('varnew', df.var / 2.0)
```

To SQL / Pandas

```
df.registerAsTable('df_tbl')
sqlContext.sql('select var from df_tbl').show(5)
df.toPandas()
```

By **mitcht**
cheatography.com/mitcht/

Not published yet.
Last updated 29th December, 2017.