Pandas to PySpark Basic Command Conversion Cheatsheet

Action	Pandas *Assuming import of numpy as np*	PySpark *Assuming import of pyspark.sql.functions as fn*
Load a CSV	df = pd.read_csv('file.csv')	df = spark.read.options(header = True, inferSchema = True).csv('file.csv')
view entire dataframe	df	display(df)
View the head of a dataframe	df.head(5)	df.show(5) or display(df.head(5))
Graphing	df.hist() *(uses matplotlib)*	display(df) # then use graph options of display
View column names	df.columns	df.columns
View column types	df.dtypes	df.dtypes (may also want df.schema)
Get dataframe shape	print((df.count(), len(df.columns)))	df.shape()
Get summary stats	df.describe()	df.describe().show()
Change columns names	df.columns = ['a', 'b', 'c']	df.toDF('a', 'b', 'c')
Rename a column	df.rename(columns = {'old', 'new'})	df.withColumnRenamed('old', 'new')
Drop a column	df.drop('col1', axis = 1)	df.drop('col1')
Add a column	df['col_3'] = df['col_1']+ ['df.col_2']	df.withColumn('col_3', (df.col_1 + df.col_2))
Fill Nulls with zero	df.fillna(0)	df.fillna(0)
Log transform	df['log_a'] = np.log(df.a)	df.withColumns('log_a', fn.log(df.a))
Extract a subset of columns of a dataframe	df[['col1', 'col3']]	df.select('col1', 'col3')
Extract a subset of rows by condition	df.loc[df['col_1'] == some_value]	df.filter(fn.col('col_1') == some_value)
Concatenate two dataframes vertically	df = pd.concat([df1, df2, df3], axis = 0)	df = df1.unionAll(df2)
Aggregation	df.groupby(['a', 'b']).agg({'a': 'mean', 'b' = 'min'})	df.groupBy(['a', 'b']).agg({'a': 'mean', 'b' = 'min'})
Pivot data and sum (could also count etc)	pivot_table(df, values='col_3', index=['col_1'], columns=['col_2'], aggfunc=np.sum)	df.groupBy('col_1').pivot('col_2').sum('col_3')
Row conditional	df['cond'] = df.apply(lambda x: 1 if x.a > 10 else 2 if x.b < 5 else 3, axis = 1)	df.withColumn('cond', fn.when(df.a > 10, 1). when(df.b < 5, 2).otherwise(3))
Apply a function	df['squared'] = df.a.apply(lambda x: x**2)	func = fn.udf(lambda x: x**2, IntegerType()) df.withColumn('squared', func(df.a))
Join dataframes (pandas join command joins on keys which PySpark doesn't have)	left.merge(right, on='id') left.merge(right, left_on='a', right_on='b')	left.join(right, on 'id') left.join(right, left.a == right.b)
Write a CSV to file	df.to_csv('filename.csv')	df.write.csv('filename.csv')
Convert between pyspark and pandas	sqlContext.createDataFrame(pandas_df)	df.toPandas()