pyspark-all you need Cheat Sheet

by ChesterHsieh via cheatography.com/62727/cs/29780/

ETL-1 to 1 relation

asd	asd

ETL-map/reduce

udf

```
@panda s_u df( 'long') def pandas _pl us_ one (se ries: pd.Series) -> pd.Series:
    # Simply plus one by using pandas Series.
.s how()
```

ETL - N to 1

groupby	df.groupby('color').avg()

ETL - streamming

I/O

Local_CSV	dataset = spark.read.csv('Bosto-nHousing.csv',inferSchema=True, header =True)	inferSchema=-Guess data type from csv
Local_Json		
Cloud_s3		
To SQL table	df.createOrReplaceTempView("tableA")	

Efficiency

repartition	
shuffle	
collect	

Spark All kind of handler df.sel ect (pa nda s_p lus _on e(

SparkC-ontext	Old man	
SparkS-ession	Young boy, that's only entry point got to know for late spark	
SparkConf	https://towardsdatascience.com/sparksession-vs-spar-kcontext-vs-sqlcontext-vs-hivecontext-741d50c9486a	
spark.sql	spark.sql("SELECT * FROM p left join e on p.name = e.name")	df.query() -> Dataframe
RDD		

EDA-Get the information for debugging/coding

printSchema	DataFrame.printSchema()		
columns:List-[str]	DataFrame.columns		
show()	df.show(1)	head()	A action, force the process to finish
take	df.take(1)		

Sponsored by **Readable.com**
Measure your website readability!
https://readable.com