Data Services

PySpark 3.1.1 Cheat Sheet

A curated list of DataFrame transformations to reduce your StackOverflow time and get you familiarized with the Apache Spark DataFrame Python API.

Legend:

D = prefix for DataFrame functions
F = prefix for pyspark.sql.functions
C = equivalent to F.col(a)
T = prefix for pyspark.sql.types

Σ = groupby function
λ = lambda function
Γ = set of options

a,b,c = column names

s = string parameter
i = integer parameter
d = datetime parameter
p = any parameter

r = regular expression
* = collection | array

Types | Casting

```
C.astype(T)
C.cast(T)

F.rint

F.from_csv
F.from_json
F.from_timestamp
F.from_utc_timestamp

F.to_csv
F.to_date
F.to_json
F.to_timestamp
F.to_utc_timestamp
```

Nulls

```
C.isNull
C.isNotNull
C.eqNullSafe(b)

F.isnan
F.isnull
F.nanvl(i, i)

F.coalesce(a, b)

D.dropna
```

Strings

```
C.startswith
C.endswith

C.contains
C.isin(*)
C.substr(a, i)

F.lower
F.upper
F.initcap

F.length

F.trim
F.ltrim
F.rtrim

F.split(a, s, i)
F.substring(a, s)
F.substring_index(a, s, i)

F.instr(a, s)
F.locate(a, s, i)

F.concat
F.concat_ws(s, *)

F.repeat
F.reverse
F.translate(a, s, s)

F.ascii
F.encode(a, Γ)
F.decode(a, Γ)

F.overlay(a, b, i)
F.levenshtein(a, b)
F.soundex
```

Dates | Time | Timezones

```
C.between(d, d)

F.from_unixtime
F.from_utc_timestamp(a, Γ)

F.timestamp_seconds
F.unix_timestamp(a, Γ)

F.current_date
F.current_timestamp
F.next_day(a, Γ)

F.date_add(d, i)
F.date_format(a, Γ)
F.date_sub(d, i)
F.date_trunc(Γ, Γ)

F.dayofweek
F.dayofmonth
F.dayofyear

F.second
F.minute
F.hour
F.days
F.last_day
F.week_of_year
F.month
F.quarter
F.year

F.add_months(a, i)
F.months_between(a, b)
F.datediff(a, b)
```

Arrays | Lists

```
C.getItem(i)

F.size

F.array

F.array_contains(a, p)
F.array_distinct
F.array_except(a, b)
F.array_intersect(q, b)
F.array_join(a, s)
F.array_union(a, b)
F.slice(a, i, i)

F.array_sort
F.sort_array

F.array_position(a, p)
F.array_remove(a, p)
F.array_repeat(a, i)
F.array_overlap(a, b)
F.array_zip(*a)
F.zip_with(a, b, λ)


F.array_max
F.array_min

F.element_at(a, i)

F.explode
F.explode_outer
F.posexplode
F.posexplode_outer
F.flatten

F.forall(a, λ)
```

Maps | Dicts

```
C.getField(s)
C.dropFields(*a)
C.withField(s, a)

F.size
F.struct(*a)
F.create_map(*a)
F.json_tuple(a, *s)

F.map_concat(*a)

F.map_entries
F.map_filter(a, λ)
F.map_from_arrays(a, b)
F.map_from_entries(a)

F.map_keys
F.map_values
F.map_zip_with(a, b, λ)
```

High Order Functions on Arrays | Maps

```
F.filter(a, λ)
F.exists(a, λ)

F.aggregate(a, p, λ)
F.transform(a, λ)
F.transform_keys(a, λ)
F.transform_values(a,  λ)
```

Aggregations

```
Σ.count()
Σ.sum()
Σ.max()
Σ.min()
Σ.mean()
Σ.avg()
Σ.pivot(a)
Σ.agg(*λ)

D.crossJoin(D)
D.crosstab(a, b)
D.cube
```

Regular Expressions

```
C.rlike(r)

F.regexp_extract(a, r, i)
F.regexp_replace(a, r, s)

D.colRegex(r)
```