Querying Data with DataFrames (Part 2)

Justin Pihony

@JustinPihony | justin-pihony.blogspot.com

Module Overview

DataFrames

- Windows
- More Functions
- Joins
- SQL UI

More Than a Row: Windows

Data Windows

...	...	...
1	“ABC”	25
...	...	...

Data Windows

...	...	...
9	“FOO”	53
...	...	...
1	“ABC”	25
...	...	...
7	“DEF”	12
...	...	...

Data Windows

...	...	...
9	“FOO”	53
2	“BAR”	42
1	“ABC”	25
3	“BAZ”	63
7	“DEF”	12
...	...	...

39

Data Windows

...	...	...
1	"FOO"	53
2	"BAR"	42
1	"ABC"	25
1	"BAZ"	63
7	"DEF"	12
...	...	...

47

Additional Function Overview

Math Functions

- (a)cos/(a)sin/(a)tan/...
- bin/(un)hex/toDegrees/toRadians
- abs/round/ceil/floor/shiftLeft/shiftRight/...
- cbrt/exp/factorial/pow/hypot
- log/log10/log2/...
- Column Math → `$"col" + lit(1)`
  - +/-/*/ //%/>(=)/<(=)/===/<=>/&&/||
  - plus/minus/multiply/divide/mod/gt(e)/lt(e)/equalTo/eqNullSafe/and/or

String Functions

- length
- split
- reverse
- `$"col".startsWith/.endsWith`
- substring/`$"col".substr/substring_index`
- lpad/rpad/repeat
- regexp_extract/regexp_replace/translate
- ascii/(un)base64/decode/encode

Datetime Functions

- current_date/current_timestamp/unix_timestamp
- date_add/date_sub/add_months
- date_diff/months_between
- date_format/from_unixtime/unix_timestamp/...
- to_date/to_utc_timestamp
- dayofmonth/dayofyear/minute/month/quarter/...
- last_day/next_date

Misc. Functions

- array/map/struct/... typedLit (2.2+)
- hash/sha/md5/...
- rand(n)
- monotonically_increasing_id
- `$"col".isNan/.is(Not)Null`
- greatest/least
- `$"col".(r)like`
- get_json_object/json_tuple/to_json/from_json
- input_file_name/spark_partition_id/`$"col".explain`
- ...

Putting Data Together

Smart Joining

```
val eqUDF = udf((x:Int, y:Int) => x == y)
df1.join(df2, eqUDF($“x”, $“y”))
```

- **Cartesian w/ Filter**

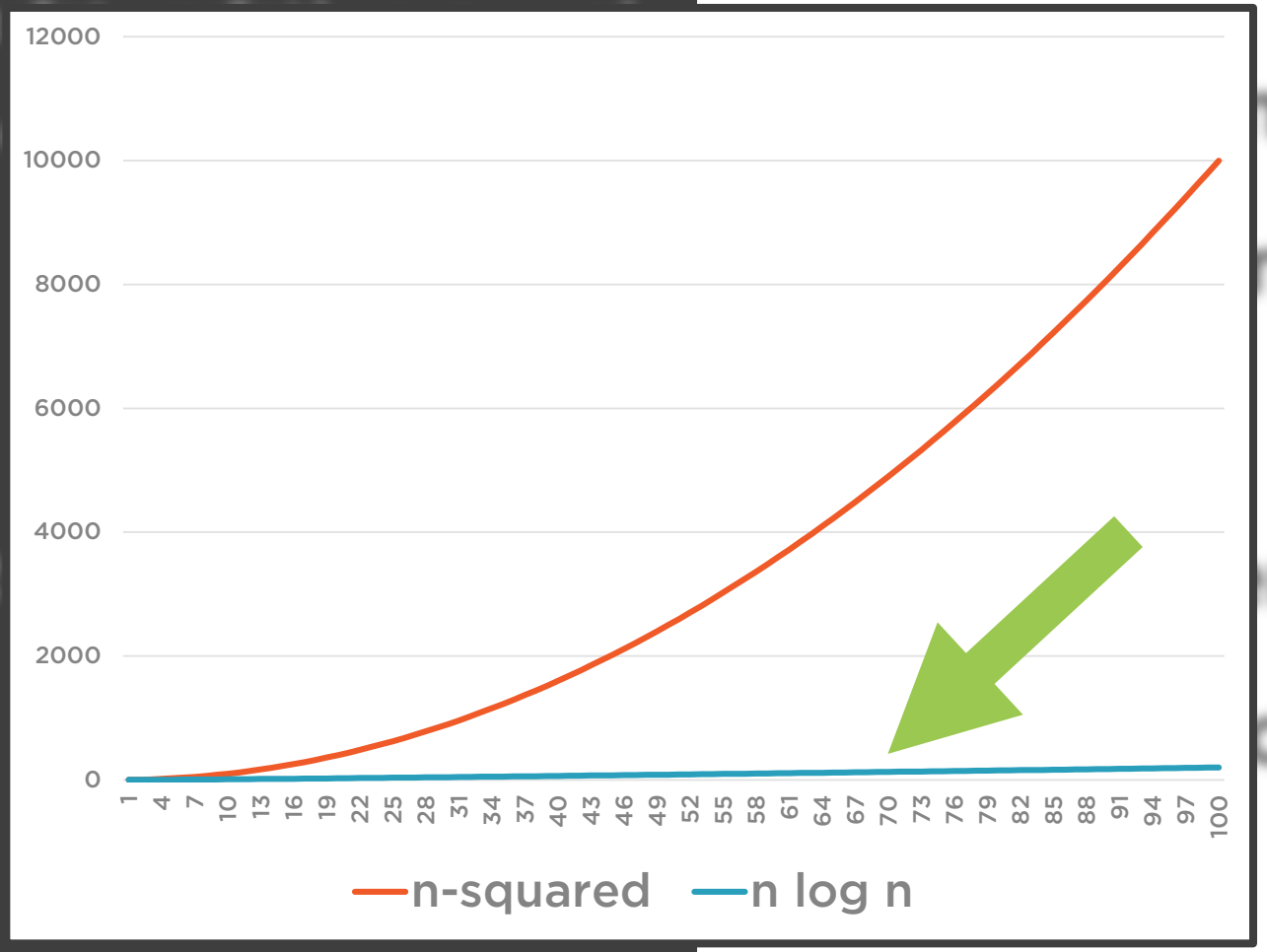
  n^2

```
df1.join(df2, $“x” === $“y”)
```

- **SortMergeJoin**

  n log n

Smart Joining



Smart Joining

```
val eqUDF = udf((x:Int, y:Int) => x == y)
df1.join(df2, eqUDF($"x", $"y"))
```

◀ Cartesian w/ Filter

n^2

```
df1.join(df2, $"x" === $"y")
```

◀ SortMergeJoin

n log n

Joins

Person

Id	FirstName	LastName
1	Justin	Pihony
2	Jane	Doe
3	John	Smith
4	Melissa	Jackson

Role

Id	JobRole
1	Programmer
3	Designer
1	Manager
5	CEO
4	Programmer
4	CTO

```
personDF.join(roleDF, $"col1" === $"col2", JOIN_TYPE)
```

```
personDF("Id") === roleDF("Id")
```

Joins

Person

Id	FirstName	LastName
1	Justin	Pihony
2	Jane	Doe
3	John	Smith
4	Melissa	Jackson

Role

Id	JobRole
1	Programmer
3	Designer
1	Manager
5	CEO
4	Programmer
4	CTO

```
personDF.join(roleDF,"Id")
personDF.join(roleDF,Seq("Id"))
personDF.join(roleDF,Seq("Id"),"inner")
```

Id	FirstName	LastName	JobRole
1	Justin	Pihony	Programmer
1	Justin	Pihony	Manager
3	John	Smith	Designer
4	Melissa	Jackson	Programmer
4	Melissa	Jackson	CTO

Joins

Person

Id	FirstName	LastName
1	Justin	Pihony
2	Jane	Doe
3	John	Smith
4	Melissa	Jackson

Role

Id	JobRole
1	Programmer
3	Designer
1	Manager
5	CEO
4	Programmer
4	CTO

"left"

```
personDF.join(roleDF,Seq("Id"),"left_outer")
```

Id	FirstName	LastName	JobRole
1	Justin	Pihony	Programmer
1	Justin	Pihony	Manager
2	Jane	Doe	null
3	John	Smith	Designer
4	Melissa	Jackson	CTO
4	Melissa	Jackson	Programmer

Joins

Person

Id	FirstName	LastName
1	Justin	Pihony
2	Jane	Doe
3	John	Smith
4	Melissa	Jackson

Role

Id	JobRole
1	Programmer
3	Designer
1	Manager
5	CEO
4	Programmer
4	CTO

"right"

```
personDF.join(roleDF,Seq("Id"),"right_outer")
```

Id	FirstName	LastName	JobRole
1	Justin	Pihony	Programmer
3	John	Smith	Designer
1	Justin	Pihony	Manager
5	null	null	CEO
4	Melissa	Jackson	Programmer
4	Melissa	Jackson	CTO

Joins

Person

Id	FirstName	LastName
1	Justin	Pihony
2	Jane	Doe
3	John	Smith
4	Melissa	Jackson

Role

Id	JobRole
1	Programmer
3	Designer
1	Manager
5	CEO
4	Programmer
4	CTO

"outer"

"full"

```
personDF.join(roleDF,Seq("Id"),"full_outer")
```

Id	FirstName	LastName	JobRole
1	Justin	Pihony	Programmer
1	Justin	Pihony	Manager
3	John	Smith	Designer
5	null	null	CEO
4	Melissa	Jackson	Programmer
4	Melissa	Jackson	CTO
2	Jane	Doe	null

Joins

Person

Id	FirstName	LastName
1	Justin	Pihony
2	Jane	Doe
3	John	Smith
4	Melissa	Jackson

Role

Id	JobRole
1	Programmer
3	Designer
1	Manager
5	CEO
4	Programmer
4	CTO

```
personDF.join(roleDF,Seq("Id"),"left_semi")
```

Id	FirstName	LastName
1	Justin	Pihony
3	John	Smith
4	Melissa	Jackson

Joins

Person

Id	FirstName	LastName
1	Justin	Pihony
2	Jane	Doe
3	John	Smith
4	Melissa	Jackson

Role

Id	JobRole
1	Programmer
3	Designer
1	Manager
5	CEO
4	Programmer
4	CTO

Spark 2.0

```
personDF.join(roleDF,Seq("Id"),"left_anti")
```

Id	FirstName	LastName
2	Jane	Doe

Joins

Spark 2.1

Person

Id	FirstName	LastName
1	Justin	Pihony
2	Jane	Doe
3	John	Smith
4	Melissa	Jackson

Role

Id	JobRole
1	Programmer
3	Designer
1	Manager
5	CEO
4	Programmer
4	CTO

```
personDF.crossJoin(roleDF)
personDF.join(roleDF,Seq("Id"),"cross")
```

~~`personDF.join(roleDF)`~~ → `spark.sql.crossJoin.enabled`

Id	FirstName	LastName	Id	JobRole
1	Justin	Pihony	1	Programmer
1	Justin	Pihony	3	Designer
1	Justin	Pihony	1	Manager
1	Justin	Pihony	5	CEO
1	Justin	Pihony	4	Programmer
1	Justin	Pihony	4	CTO
2	Jane	Doe	1	Programmer
2	Jane	Doe	3	Designer
...	...	...	...	...

Resources

- Understanding Windows and More
  - T-SQL Window Functions: Kathi Kellenberger
    - app.pluralsight.com/courses/tsql-window-functions
  - SQL Window Functions: Mode Analytics
    - community.modeanalytics.com/sql/tutorial/sql-window-functions
  - Introducing Window Functions in Spark SQL: Databricks
    - databricks.com/blog/2015/07/15/introducing-window-functions-in-spark-sql
  - Reshaping Data with Pivot in Apache Spark: Andrew Ray
    - databricks.com/blog/2016/02/09/reshaping-data-with-pivot-in-apache-spark
- User-defined Aggregate Functions
  - Apache Spark 1.5...UDAFs: Databricks
    - databricks.com/blog/2015/09/16/apache-spark-1-5-dataframe-api-highlights
  - How to define and use a User-Defined Aggregate Function in Spark SQL?
    - stackoverflow.com/a/32101530/779513
- Spark SQL Reference
  - docs.databricks.com/spark/latest/spark-sql/index

Summary

Windows

More Functions

User Defined Functions

Joins

SQL UI