

Spark SQL

DataFrames and Databases

DataFrame & Database table: conceptually equivalent.

Spark makes them interoperable

DataFrames & SQL

- Query existing Spark DataFrames (however created) with SQL
 - Same functionality, different interface
 - Legacy SQL

Read data and persistency

- Load data from Hive / other databases
- Save tables to Hive

Explore the Yelp dataset

```
yelp_df = sqlCtx.load(  
    source='com.databricks.spark.csv',  
    header = 'true',  
    inferSchema = 'true',  
    path =  
    'file:///usr/lib/hue/apps/search/examples/collections/solr_co  
nfigs_yelp_demo/index_data.csv')
```

Register as a SQL table

- DataFrame already has Schema
- Create a temporary table with:
`yelp_df.registerTempTable("yelp")`

Run SQL statements

```
filtered_yelp = sqlCtx.sql("SELECT * FROM yelp WHERE  
useful >= 1")
```

```
filtered_yelp
```

```
Out[]: DataFrame[business_id: string, cool: int, date: string,  
funny: int, id: string, stars: int, text: string, type: string,  
useful: int, user_id: string, name: string, full_address:  
string, latitude: double, longitude: double, neighborhoods:  
string, open: string, review_count: int, state: string]
```

Filtering

```
filtered_yelp.count()
```

```
Out[]: 601L
```

```
yelp_df.filter(yelp_df.useful >= 1).count()
```

```
Out[]: 601L
```


aggregation

```
sqlCtx.sql("SELECT MAX(useful) AS max_useful FROM  
yelp").collect()
```

```
Out[]: [Row(max_useful)=28)]
```

```
yelp_df.agg({"useful": "max"}).collect()
```

```
Out[]: [Row(MAX(useful#267)=28)]
```

Join - select - show

```
useful_perc_data.join(  
    yelp_df,  
    yelp_df.id == useful_perc_data.uid,  
    "inner"  
).select(useful_perc_data.uid, "useful_perc", "review_count")  
  
|  
  
|
```

Register as SQL table

```
useful_perc_data.registerTempTable("useful_perc_data")
```

join

```
sqlCtx.sql(  
    """SELECT useful_perc_data.uid, useful_perc,  
    review_count  
    FROM useful_perc_data  
    INNER JOIN yelp  
    ON useful_perc_data.uid=yelp.id""")  
)
```

Performance

- Either DataFrame calls or SQL
- Same under-the-hood optimizer (Catalyst)
- Creates DAG
- Parallel execution
- Creates bytecode

Spark and Hive

Spark and Hive

- copied hive-site.xml to Spark conf/
- Spark read / write to Hive

Hive table to DataFrame

- sqlCtx.sql has access to Hive tables
- Load data uploaded during the Hive class
- Result is a DataFrame

```
customers_df = sqlCtx.sql("SELECT * FROM customers")  
customers_df.show()
```


Printout of customers_df

customer_id	customer_fname	customer_lname	customer_email	customer_password	customer_street	customer_city	customer_state	customer_zipcode
1	Richard	Hernandez	XXXXXXXXXX	XXXXXXXXXX	6303 Heather Plaza	Brownsville	TX	78521
2	Mary	Barrett	XXXXXXXXXX	XXXXXXXXXX	9526 Noble Embers...	Littleton	CO	80126
3	Ann	Smith	XXXXXXXXXX	XXXXXXXXXX	3422 Blue Pioneer...	Caguas	PR	00725
4	Mary	Jones	XXXXXXXXXX	XXXXXXXXXX	8324 Little Common	San Marcos	CA	92069
5	Robert	Hudson	XXXXXXXXXX	XXXXXXXXXX	10 Crystal River ...	Caguas	PR	00725
6	Mary	Smith	XXXXXXXXXX	XXXXXXXXXX	3151 Sleepy Quail...	Passaic	NJ	07055
7	Melissa	Wilcox	XXXXXXXXXX	XXXXXXXXXX	9453 High Concession	Caguas	PR	00725
8	Megan	Smith	XXXXXXXXXX	XXXXXXXXXX	3047 Foggy Forest...	Lawrence	MA	01841
9	Mary	Perez	XXXXXXXXXX	XXXXXXXXXX	3616 Quaking Street	Caguas	PR	00725
10	Melissa	Smith	XXXXXXXXXX	XXXXXXXXXX	8598 Harvest Beac...	Stafford	VA	22554
11	Mary	Huffman	XXXXXXXXXX	XXXXXXXXXX	3169 Stony Woods	Caguas	PR	00725
12	Christopher	Smith	XXXXXXXXXX	XXXXXXXXXX	5594 Jagged Ember...	San Antonio	TX	78227
13	Mary	Baldwin	XXXXXXXXXX	XXXXXXXXXX	7922 Iron Oak Gar...	Caguas	PR	00725
14	Katherine	Smith	XXXXXXXXXX	XXXXXXXXXX	5666 Hazy Pony Sq...	Pico Rivera	CA	90660
15	Jane	Luna	XXXXXXXXXX	XXXXXXXXXX	673 Burning Glen	Fontana	CA	92336
16	Tiffany	Smith	XXXXXXXXXX	XXXXXXXXXX	6651 Iron Port	Caguas	PR	00725
17	Mary	Robinson	XXXXXXXXXX	XXXXXXXXXX	1325 Noble Pike	Taylor	MI	48180
18	Robert	Smith	XXXXXXXXXX	XXXXXXXXXX	2734 Hazy Butterf...	Martinez	CA	94553
19	Stephanie	Mitchell	XXXXXXXXXX	XXXXXXXXXX	3543 Red Treasure...	Caguas	PR	00725
20	Mary	Ellis	XXXXXXXXXX	XXXXXXXXXX	4703 Old Route	West New York	NJ	07093

```
customers_df.printSchema()
```

```
root
```

```
|-- customer_id: integer (nullable = true)
```

```
|-- customer_fname: string (nullable = true)
```

```
|-- customer_lname: string (nullable = true)
```

```
|-- customer_email: string (nullable = true)
```

```
|-- customer_password: string (nullable = true)
```

```
|-- customer_street: string (nullable = true)
```

```
|-- customer_city: string (nullable = true)
```

```
|-- customer_state: string (nullable = true)
```

```
|-- customer_zipcode: string (nullable = true)
```

Run unmodified SQL queries

```
sqlCtx.sql("""select c.category_name,  
count(order_item_quantity) as count from order_items oi  
inner join products p on oi.order_item_product_id =  
p.product_id inner join categories c on c.category_id =  
p.product_category_id group by c.category_name  
order by count desc  
limit 10""")  
).show()
```

Most popular categories

category_name	count
Cleats	24551
Men's Footwear	22246
Women's Apparel	21035
Indoor/Outdoor Games	19298
Fishing	17325
Water Sports	15540
Camping & Hiking	13729
Cardio Equipment	12487
Shop By Sport	10984
Electronics	3156

Run unmodified SQL queries

```
sqlCtx.sql("""select p.product_id, p.product_name, r.revenue
from products p inner join
(select oi.order_item_product_id, sum(cast(oi.order_item_subtotal as float)) as
revenue from order_items oi inner join orders o on oi.order_item_order_id =
o.order_id
where o.order_status <> 'CANCELED'
and o.order_status <> 'SUSPECTED_FRAUD'
group by order_item_product_id) r
on p.product_id = r.order_item_product_id
order by r.revenue desc limit 10""")
.show()
```

Top 10 products by revenue

product_id	product_name	revenue
1004	Field & Stream Sp...	6637668.282318115
365	Perfect Fitness P...	4233794.3682899475
957	Diamondback Women...	3946837.004547119
191	Nike Men's Free 5...	3507549.2067337036
502	Nike Men's Dri-FI...	3011600.0
1073	Pelican Sunstream...	2967851.6815185547
1014	O'Brien Men's Neo...	2765543.314743042
403	Nike Men's CJ Eli...	2763977.4868011475
627	Under Armour Girl...	1214896.220287323
565	adidas Youth Germ...	63490.0

Save DataFrames to Hive

`registerTempTable` only gives temporary SQL-like access to DataFrames

Store permanently to Hive with:

```
yelp_df.saveAsTable("yelp_reviews")
```

Check persistency

- Restart PySpark
- Run: `sqlCtx.sql("SELECT * FROM yelp").show()`
- Fails with "Table not found"

Check persistency

- Restart PySpark
- Run: `sqlCtx.sql("SELECT * FROM yelp_reviews").show()`
- Restores from Hive

Loaded Yelp DataFrame

business_id	cool	date	funny_id	stars	text	type	useful	user_id	name	full_address	latitude	longitude	neighborhood
hoods open review_count state													
9yKzy9PApe1P0UJE... 2	True 116	2011-01-26 0	fWkVx83p0-ka4JS3d...	4	My wife took me h...	business	5	rLt18ZkDX5vH5nAx9...	Morning Glory Cafe	6106 S 32nd St Ph...	33.3907928467	-112.012504578	[]
ZRjwVlyzEJq1VAihD... 0	True 182	2011-07-27 0	IjZ33sJrzXqU-0X6U...	4	I have no idea wh...	business	0	0a2KyEL0d3Yb1V6ai...	Spinato's Pizzeria	4848 E Chandler B...	33.305606842	-111.978759766	[]
6oRAC4uyJCsj1lX0W... 0	True 265	2012-06-14 0	IESLBzqUCLdSzSqm0...	4	love the gyro pla...	business	1	0hT2KtflLiobPvh6cD...	Haji-Baba	1513 E Apache BL...	33.4143447876	-111.913032532	[]
_10QZuf4zZ0yFcvXc... 1	True 88	2010-05-27 0	G-WvGaISbqqqMHLNn...	4	Rosie, Dakota, an...	business	2	uZet19T0NcR0G0yFf...	Chaparral Dog Park	5401 N Hayden Rd ...	33.5229454041	-111.90788269	[]
6ozycU1RpktNG2-1B... 0	True 5	2012-01-05 0	luJFq2r5qfJG_6ExM...	4	General Manager S...	business	0	vYmM4KTsC8zf0Bg-j...	Discount Tire	1357 S Power Road...	33.3910255432	-111.68447876	[]
-yxfBYGB6SEsqzmxJ... 4	True 109	2007-12-13 1	m2CKSsepBCoRYWxiR...	3	Quiescence is, si...	business	3	sqYN3lNgvPbPCTRsM...	Quiescence Restau...	6106 S 32nd St Ph...	33.3907928467	-112.012504578	[]
zp713qNhx8d9KCJjN... 7	True 307	2010-02-12 4	riFQ3vxNpP4rWlK_C...	4	Drop what you're ...	business	7	wFweIWhv2fREZV_dY...	La Condesa Gourme...	1919 N 16th St Ph...	33.4691314697	-112.04750824	[]
hW0Ne HTHEAgGF1rA... 0	True 862	2012-07-12 0	JL7GXJ9u4YMx7Rzs0...	3	Luckily, I didn't...	business	1	lieuYcKS7zeAv_U15...	Phoenix Sky Harbo...	3400 E Sky Harbor...	33.4347496033	-112.006439209	[]
wNUea3IXZWd63bb0Q... 0	True 163	2012-08-17 0	XtnfnYmnJy17yIuG...	3	Definitely come f...	business	0	Vh_Dl1zgGhSqHq4qf...	Stingray Sushi	2574 E Camelback ...	33.5096054077	-112.025741577	[]
nMHhuYan8e3c0No3P... 0	True 189	2010-08-11 0	jJAIXA46pU1swYyRC...	4	Nobuo shows his u...	business	1	sUNkXg8-KftCMQDV6...	Nobuo At Teeter H...	622 E Adams St Ph...	33.4495391846	-112.065666199	[]
AsSCv0q_BWqIe3mX2... 1	False 74	2010-06-16 1	E11jzpKz9Kw5K7fuA...	4	The oldish man wh...	business	3	-0M1S6yWkYjVldNhC...	Cookiez On Mill	514 S Mill Ave St...	33.4248809814	-111.940200806	[]
e9nN4XxjdHj4qtKC0... 1	True 192	2011-10-21 0	3rPt0Lx7rgmEurzn...	4	Wonderful Vietnam...	business	1	ClRhP3dmePnea7Xio...	Lee's Sandwiches	1901 W Warner Rd ...	33.3347129822	-111.874786377	[]
h53YuC1IDfEF5JC0p... 1	True 36	2010-01-11 0	cGnKNX3I9rthE0-TH...	4	They have a limit...	business	2	UPtysDF6cUDUxq2KY...	Jason's Deli	1065 E Baseline R...	33.3796195984	-111.809425354	[]
WGN1YMeXPyowav1AP... 1	True 25	2011-12-23 0	FvEEw1_OsrYdvwLV5...	4	Good tattoo shop...	business	2	Xm8HXE1JHqsXe5BK...	The Lady Luck Tat...	961 E Guadalupe R...	33.3637619019	-111.9272995	[]
yc5AH9H71xJidA_J2... 1	True 151	2010-05-20 0	pfUwBKYYmUXeiwrhD...	4	I'm 2 weeks new t...	business	1	J0G-4G4e8ae3lx_sz...	Rosie McCaffrey's	906 E Camelback R...	33.5095176697	-112.061569214	[]
Vb9FPCeL6Ly24PNxL... 0	False 28	2011-03-20 0	HvqmdqWcerVW03Gs6...	4	Was it worth the ...	business	2	ylW0J2y7TV2e3yYeW...	Z Pizza	13637 N Tatum Blv...	33.6101531982	-111.976852417	[]
sup1gcPN09IKo6ola... 3	True 86	2008-10-12 2	HXP_0UL-FCmA4f-k9...	3	We went here on a...	business	4	SBbftLzfYYKItOMFw...	1130 The Restaurant	455 N 3rd St Ste ...	33.452796936	-112.069320679	[]
0510Re68m0y9dU490... 0	False 39	2010-05-03 0	j4SIzrIy0WrmW4yr4...	3	okay this is the ...	business	0	u1KwcbPmXFEEYkZ2...	Oakville Grocery	15015 N Scottsdal...	33.6246795654	-111.924377441	[]
b5cEoKR81Qliq-yT2... 5	True 262	2009-03-06 4	v0cTd3PNpYCKtYGks...	4	I met a friend fo...	business	6	UsULgP4bKA8Rmzs8d...	Carlsbad Tavern	3313 N Hayden Rd ...	33.4869194031	-111.908737183	[]
4JzbSbK9wm10BJZW... 1	True 13	2011-11-17 1	a0lCu-j2Sk_kHQsZi...	2	They've gotten be...	business	1	nDBly08j5URmrHQ2J...	Frontier Airlines	Phoenix Sky Harbo...	33.4396476746	-112.026153564	[]

quickstart.cloudera:8888/metastore/tables

The screenshot shows the Cloudera Hue web interface. The browser address bar displays `10.0.2.15:8888/metastore/tables/`. The top navigation bar includes links for Cloudera, Hue, and various data tools like Hadoop, HBase, Impala, Spark, Solr, and Oozie. Below this, a secondary navigation bar features 'HUE' and menu items for Query Editors, Data Browsers, Workflows, Search, and Security.

The main content area is titled 'Metastore Manager'. On the left, there's a sidebar with a 'DATABASE' section containing a dropdown menu set to 'default', and an 'ACTIONS' section with two options: 'Create a new table from a file' and 'Create a new table manually'.

The main panel on the right is titled 'Databases > default'. It contains a search bar labeled 'Search for table name'. Below the search bar is a list of tables, each with a checkbox and a diamond icon, followed by the table name:

- ☐ ☐ Table Name
- ☐ categories
- ☐ customers
- ☐ departments
- ☐ order_items
- ☐ orders
- ☐ products
- ☐ yelp_reviews

Conclusion

- Analytics with DataFrames, filtering, aggregation, joins, grouping
- How to add new packages to Spark and modify Hadoop configuration
- Operate on DataFrames with SQL
- Persist DataFrames as Hive tables