Data Exploration with Apache Drill

Charles S. Givre
@cgivre
thedataist.com
linkedin.com/in/cgivre

Conventions for this class:

- SQL Commands and Keywords will be written in ALL CAPS
- Variable names will use underscores and be completely in lowercase
- File names will be as they are in the file system
- User provided input will be enclosed in <input>

Expectations

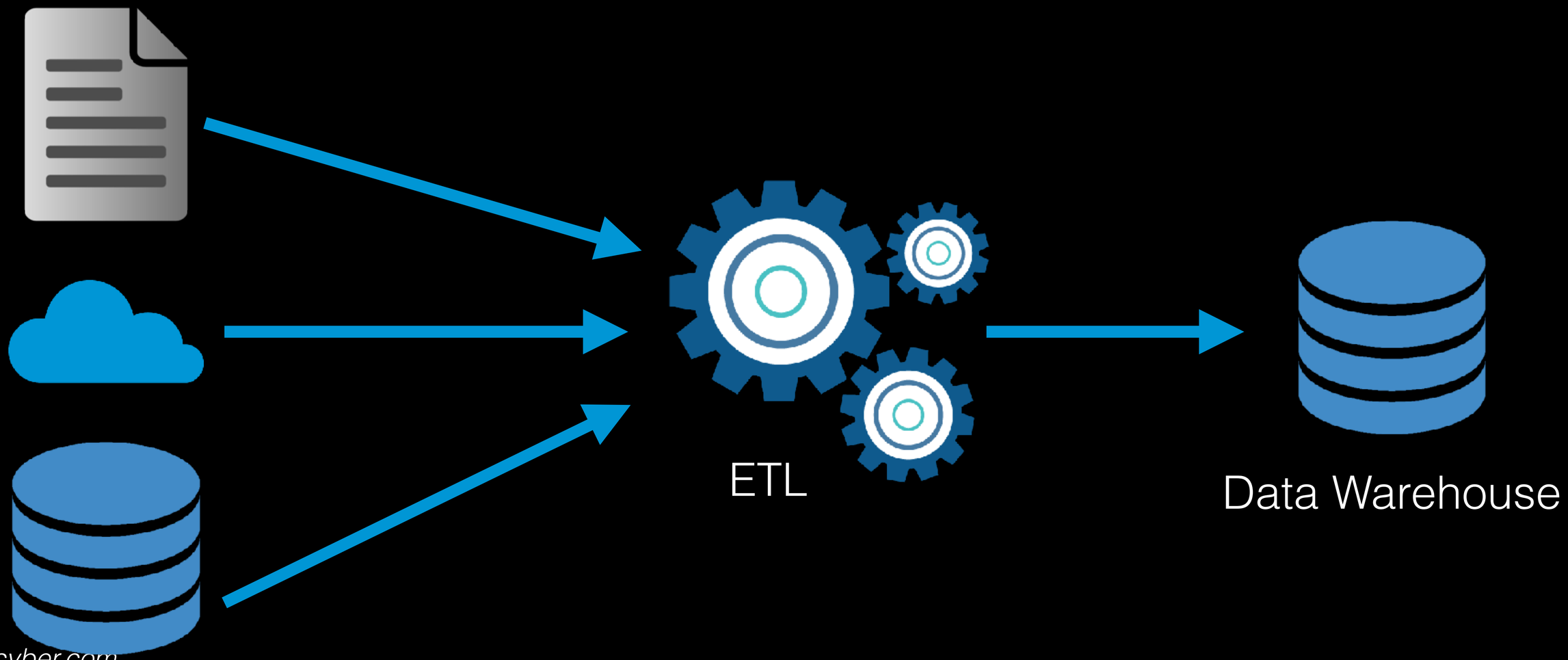
- Please participate and **ask questions.**
- Please follow along and **TRY OUT** the examples yourself during the class
- All the answers are in the slide decks or GitHub repository, but please try to complete the exercises **without looking at the answers**.
- Have fun!

The problems

We want SQL and BI support without compromising flexibility and ability of NoSchema datastores.

Data is not arranged in an optimal way for ad-hoc analysis

Data is not arranged in an optimal way for ad-hoc analysis

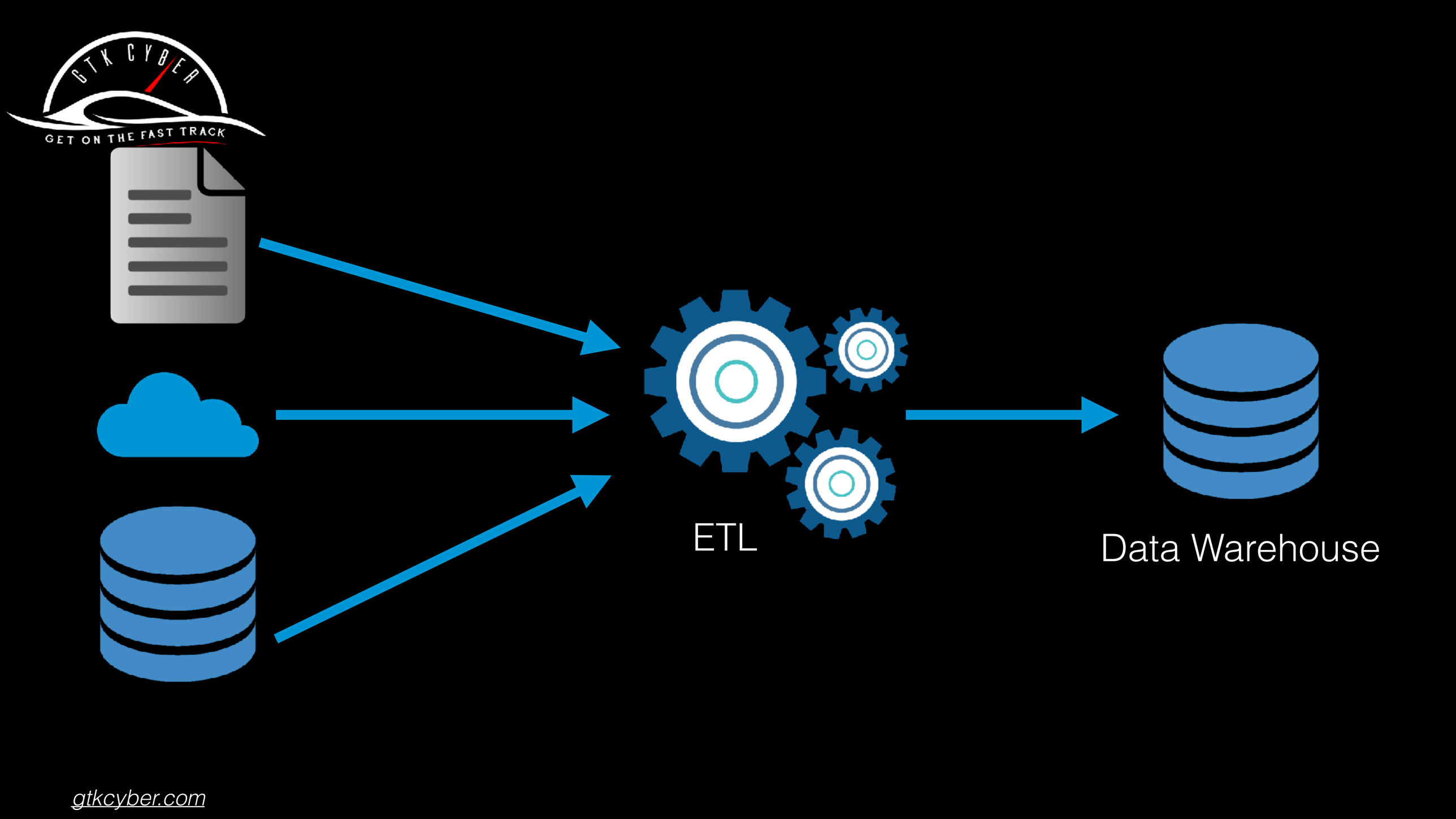


Analytics teams spend between 50%-90% of their time preparing their data.

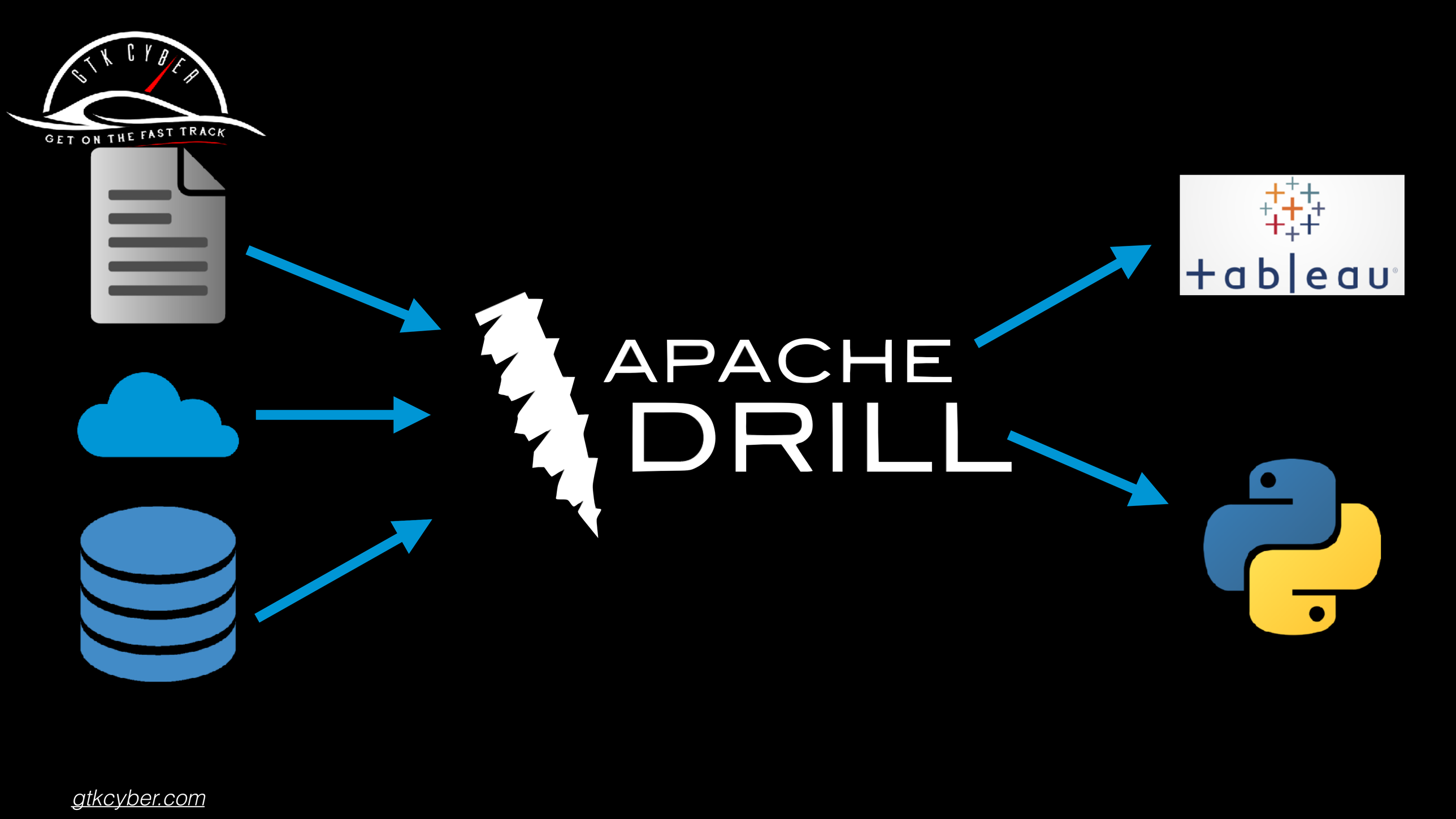
76% of Data Scientist say this is the least enjoyable part of their job.

http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

The ETL Process **consumes the most time** and **contributes almost no value** to the end product.



gtkcyber.com


GTK CYBER
GET ON THE FAST TRACK
APACHE
DRILL
+ableau
gtkcyber.com

You just query the data...
no schema

Drill is NOT just SQL on Hadoop

Drill scales

Drill is open source

Download Drill at: drill.apache.org

Why should you use Drill?

Why should you use Drill?

Drill is **easy to use**

Drill is **easy to use**

Drill uses **standard ANSI SQL**

Drill is **FAST!!**

CASE-1: Aggregation Query: Total number of records that have a rating of "3" or above	
Spark	spark-submit --class AggQuery --conf "spark.driver.memory=3g" sql1/target/scala-2.11/sparksqldemo1-assembly-1.0.jar
	real 0m52.631s user 1m31.097s sys 0m1.002s
Drill	drill-embedded -f sql1.sql (ALTER SYSTEM SET `store.json.read_numbers_as_double` = true; SELECT COUNT(MOVIE) FROM dfs.`ratings.json` WHERE RATE > 3.0;)
	real 0m23.917s user 0m26.819s sys 0m0.703s

https://www.mapr.com/blog/comparing-sql-functions-and-performance-apache-spark-and-apache-drill

CASE-2: Join Query: The user and movie names were joined, and the top 10 female users who most rated movies were extracted	
Spark	spark-submit --class topTenWomen --conf "spark.driver.memory=3g" sql2-1/target/scala-2.11/sparksqldemo2-1-assembly-1.0.jar
	real 0m57.671s user 1m40.031s sys 0m1.152s
Drill	drill-embedded -f sql2-1.sql (ALTER SYSTEM SET `store.json.read_numbers_as_double` = true; SELECT RATtbl.UID, COUNT(RATtbl.UID) as NUMEVALS FROM dfs.`ratings.json` as RATtbl JOIN dfs.`users.json` as USRtbl ON RATtbl.UID = USRtbl.UID WHERE USRtbl.GENDER = 'F' GROUP BY RATtbl.UID ORDER BY COUNT(RATtbl.UID) DESC LIMIT 10;)
	real 0m27.576s user 0m28.370s sys 0m0.803s

https://www.mapr.com/blog/comparing-sql-functions-and-performance-apache-spark-and-apache-drill

CASE-3: Join Query: The user and movie names were joined, and the top 10 names of highly rated movies were extracted	
Spark	spark-submit --class topTenMoviename --conf "spark.driver.memory=3g" sql2-2/target/scala-2.11/sparksqldemo2-2-assembly-1.0.jar
	real 0m57.982s user 1m41.480s sys 0m1.246s
Drill	drill-embedded -f sql2-2.sql (ALTER SYSTEM SET `store.json.read_numbers_as_double` = true; SELECT TMP2.MOVIE, TMP2.NUMRAT, TITLE FROM (SELECT MOVIE, COUNT(MOVIE) as NUMRAT FROM dfs.`ratings.json` GROUP BY MOVIE) TMP2 JOIN dfs.`movies.json` AS MOVtbl ON TMP2.MOVIE = MOVtbl.MOVIE ORDER BY TMP2.NUMRAT DESC LIMIT 10;)
	real 0m28.217s user 0m29.263s sys 0m0.658s

https://www.mapr.com/blog/comparing-sql-functions-and-performance-apache-spark-and-apache-drill

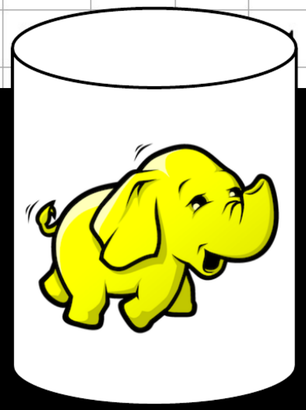
Quick Demo

Thank you Jair Aguirre!!

Quick Demo

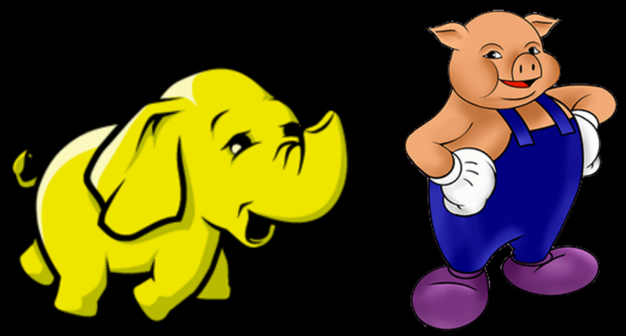
yearID	lgID	teamID	franchID	divID	Rank	G	Ghome	W	L	DivWin	WCWin	LgWin	WSWin	R	AB	H	2B	3B	HR	E
1871	NA	BS1	BNA		3	31		20	10			N		401	1372	426	70	37	3	
1871	NA	CH1	CNA		2	28		19	9			N		302	1196	323	52	21	10	
1871	NA	CL1	CFC		8	29		10	19			N		249	1186	328	35	40	7	
1871	NA	FW1	KEK		7	19		7	12			N		137	746	178	19	8	2	
1871	NA	NY2	NNA		5	33		16	17			N		302					1	

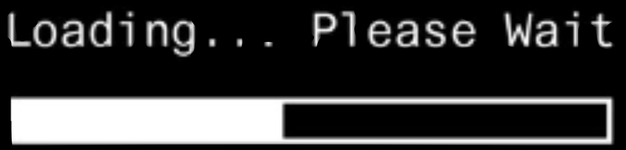
anlahman.com/baseball-archive/statistics



Quick Demo

```
data = load '/user/cloudera/data/baseball_csv/Teams.csv' using PigStorage(',');
filtered = filter data by ($0 == '1988');
tm_hr = foreach filtered generate (chararray) $40 as team, (int) $19 as hrs;
ordered = order tm_hr by hrs desc;
dump ordered;
```





Execution Time:
1 minute, 38 seconds

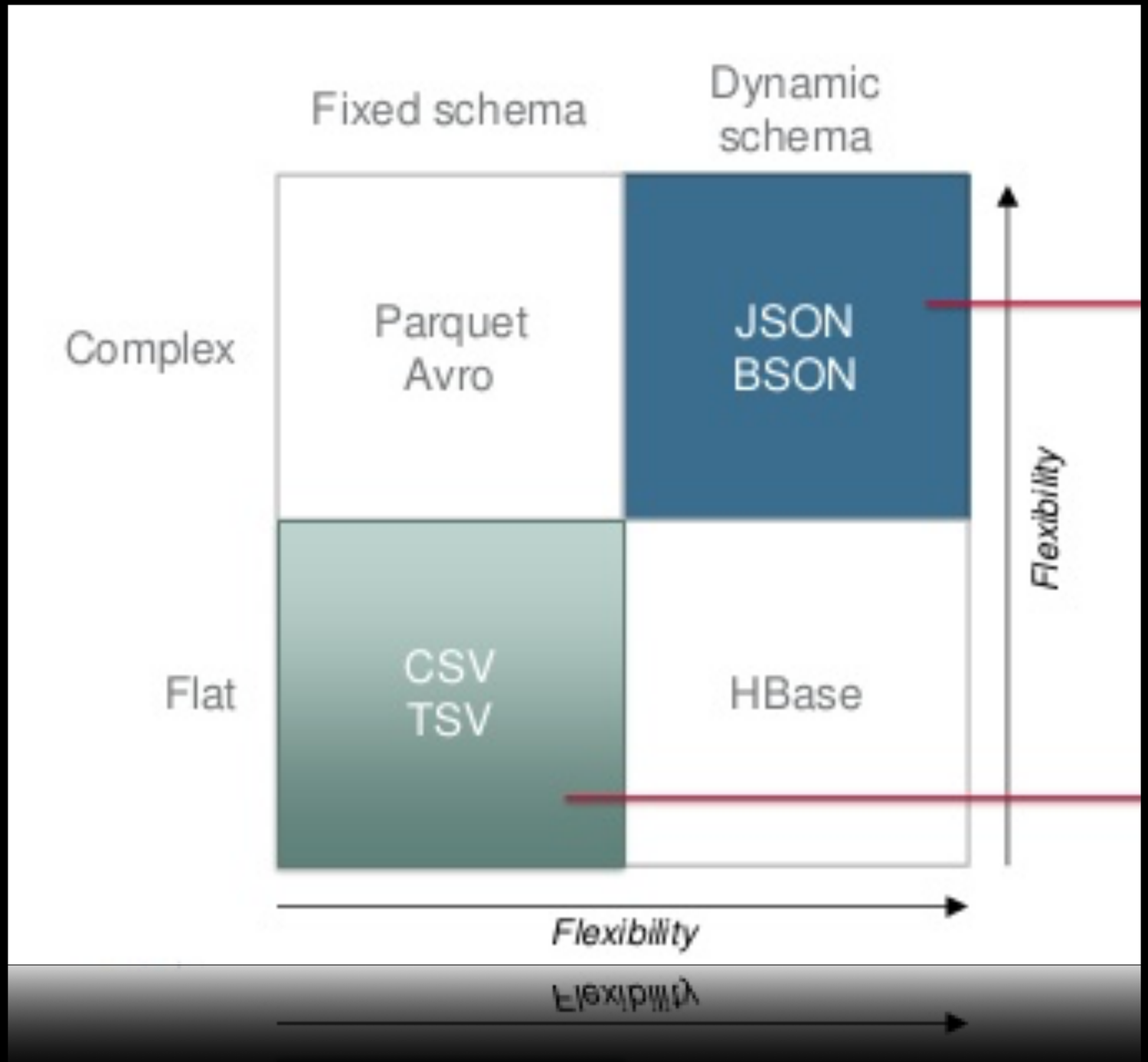
gtkcyber.com

Quick Demo

```
SELECT columns[40], cast(columns[19] as int) AS HR
FROM `baseball_csv/Teams.csv`
WHERE columns[0] = '1988'
ORDER BY HR desc;
```

Execution Time:
0232 seconds!!

Drill is Versatile



NoSQL, No Problem

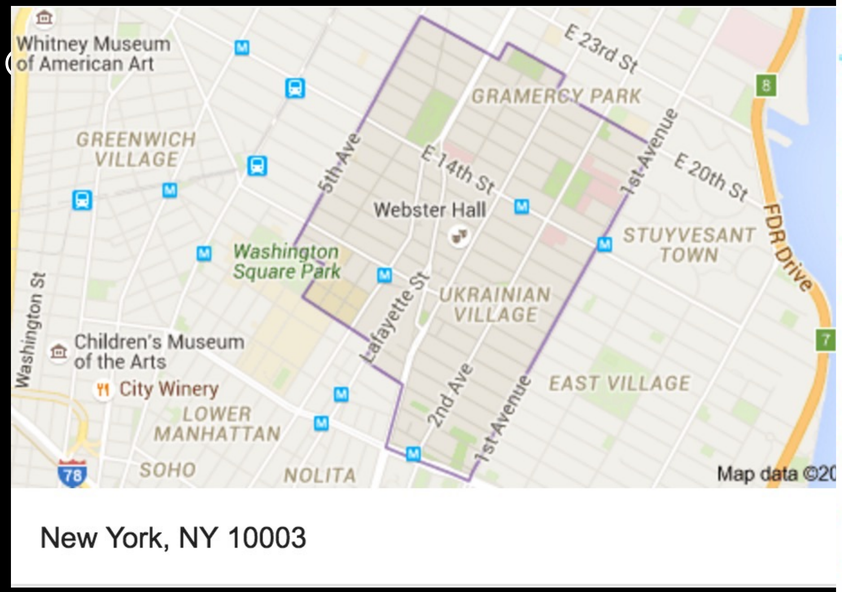
NoSQL, No Problem


```
{
  "address": {
     "building": "1007",
     "coord": [ -73.856077, 40.848447 ],
     "street": "Morris Park Ave",
     "zipcode": "10462"
  },
  "borough": "Bronx",
  "cuisine": "Bakery",
  "grades": [
     { "date": { "$date": 1393804800000 }, "grade": "A", "score": 2 },
     { "date": { "$date": 1378857600000 }, "grade": "A", "score": 6 },
     { "date": { "$date": 1358985600000 }, "grade": "A", "score": 10 },
     { "date": { "$date": 1322006400000 }, "grade": "A", "score": 9 },
     { "date": { "$date": 1299715200000 }, "grade": "B", "score": 14 }
  ],
  "name": "Morris Park Bake Shop",
  "restaurant_id": "30075445"
}
```


https://raw.githubusercontent.com/mongodb/docs-assets/primer-dataset/primer-dataset.json

NoSQL, No Problem

```
SELECT t.address.zipcode AS zip, count(name) AS rests
FROM `restaurants` t
GROUP BY t.address.zipc
ORDER BY rests DESC
LIMIT 10;
```

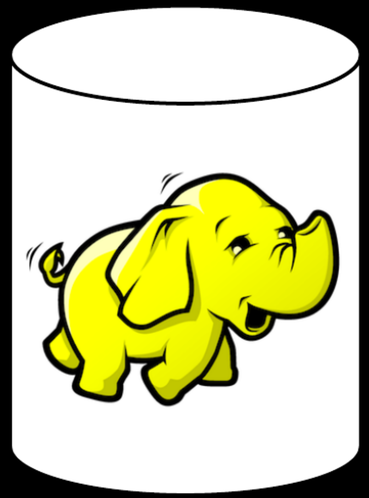


zip	rests
10003	686
10019	675
10036	611
10001	520
10022	485
10013	480
10002	471
10011	467
10016	433
10014	428

https://raw.githubusercontent.com/mongodb/docs-assets/primer-dataset/primer-dataset.json

Querying Across Silos

Querying Across Silos



Farmers Market Data

Restaurant Data

Querying Across Silos

```
SELECT t1.Borough, t1.markets, t2.rests, cast(t1.markets AS
FLOAT)/ cast(t2.rests AS FLOAT) AS ratio
FROM (
 SELECT Borough, count(`Farmers Markets Name`) AS markets
 FROM `farmers_markets.csv`
 GROUP BY Borough ) t1
 JOIN (
   SELECT borough, count(name) AS rests
   FROM mongo.test.`restaurants`
   GROUP BY borough
   ) t2
ON t1.Borough=t2.borough
ORDER BY ratio DESC;
```

Querying Across Silos

Borough	markets	rests	ratio
Bronx	18	2338	0.007698888
Brooklyn	34	6086	0.005586592
Manhattan	36	10259	0.003509114
Queens	12	5656	0.0021216408
Staten Island	1	969	0.0010319918

Execution Time: 0.502 Seconds

Querying Across Silos

data.world

GTK CYBER
GET ON THE FAST TRACK

Querying Across Silos

+

APACHE DRILL

data.world

Querying Across Silos

+

APACHE DRILL =

data.world

Querying Across Silos

bin — java -Xms4G -Xmx4G -XX:MaxDirectMemorySize=8G -XX:ReservedCodeCacheSize=1G -Ddrill.exec.ena...

...rive/github/drill-xml-plugin — -bash | java -Xms4G -Xmx...jdbc:drill:zk=local | ~/OneDrive/metabase — -bash

```
0: jdbc:drill:zk=local> SELECT companyName FROM dw.cgivre.`mac-address-manufacturers`.`20170426mac_address.csv/20170426mac_address` WHERE country='CN' LIMIT  20;
+-----------------------------------------------------------+
|                        companyName                        |
+-----------------------------------------------------------+
| Shenzhen ViewAt Technology Co.,Ltd.                       |
| ShenZhen ANYK Technology Co.,LTD                          |
| SOYEA Technology Co.,Ltd.                                 |
| Nanjing Shining Electric Automation Co., Ltd              |
| Anhui comhigher tech co.,ltd                              |
| Wei Fang Goertek Electronics Co.,Ltd                      |
| zte corporation                                           |
| SHENZHEN DAJIAHAO TECHNOLOGY CO.,LTD                      |
| SHENZHEN CLOU ELECTRONICS CO. LTD.                        |
| HONG KONG TECON TECHNOLOGY                                |
| Zhehua technology limited                                 |
| Chengdu Fuhuaxin Technology co.,Ltd                       |
| Shanghai LISTEN TECH.LTD                                  |
| Shenzhen ChuangDao & Perpetual Eternal Technology Co.,Ltd |
| Beijing Novel Super Digital TV Technology Co., Ltd        |
| URadio Systems Co., Ltd                                   |
| Jiangsu Qinheng Co., Ltd.                                 |
| TP-LINK TECHNOLOGIES CO.,LTD.                             |
| BEIJING GEHUA CATV NETWORK CO.,LTD                        |
| TP-LINK TECHNOLOGIES CO.,LTD.                             |
+-----------------------------------------------------------+
20 rows selected (1.087 seconds)
0: jdbc:drill:zk=local>
```

Querying Across Silos

```
SELECT SUBSTRING( REGEXP_REPLACE( MACAddressSource, ':', '' ),1,6 ) AS MacAddress,
MACAddressSource, dw.companyName, dw.country
FROM dfs.test.`pcapview` AS p
JOIN dw.cgivre.`mac-address-manufacturers`.`20170426mac_address.csv/20170426mac_address`
AS dw ON dw.prefix = SUBSTRING( REGEXP_REPLACE( MACAddressSource, ':', '' ),1,6 )
```

MacAddress	MACAddressSource	companyName	country
080027	08:00:27:38:DB:ED	PCS Systemtechnik GmbH	US
080027	08:00:27:97:3F:45	PCS Systemtechnik GmbH	US

2 rows selected (3.775 seconds)

Why **aren’t** you using Drill?

Installing & Configuring Drill

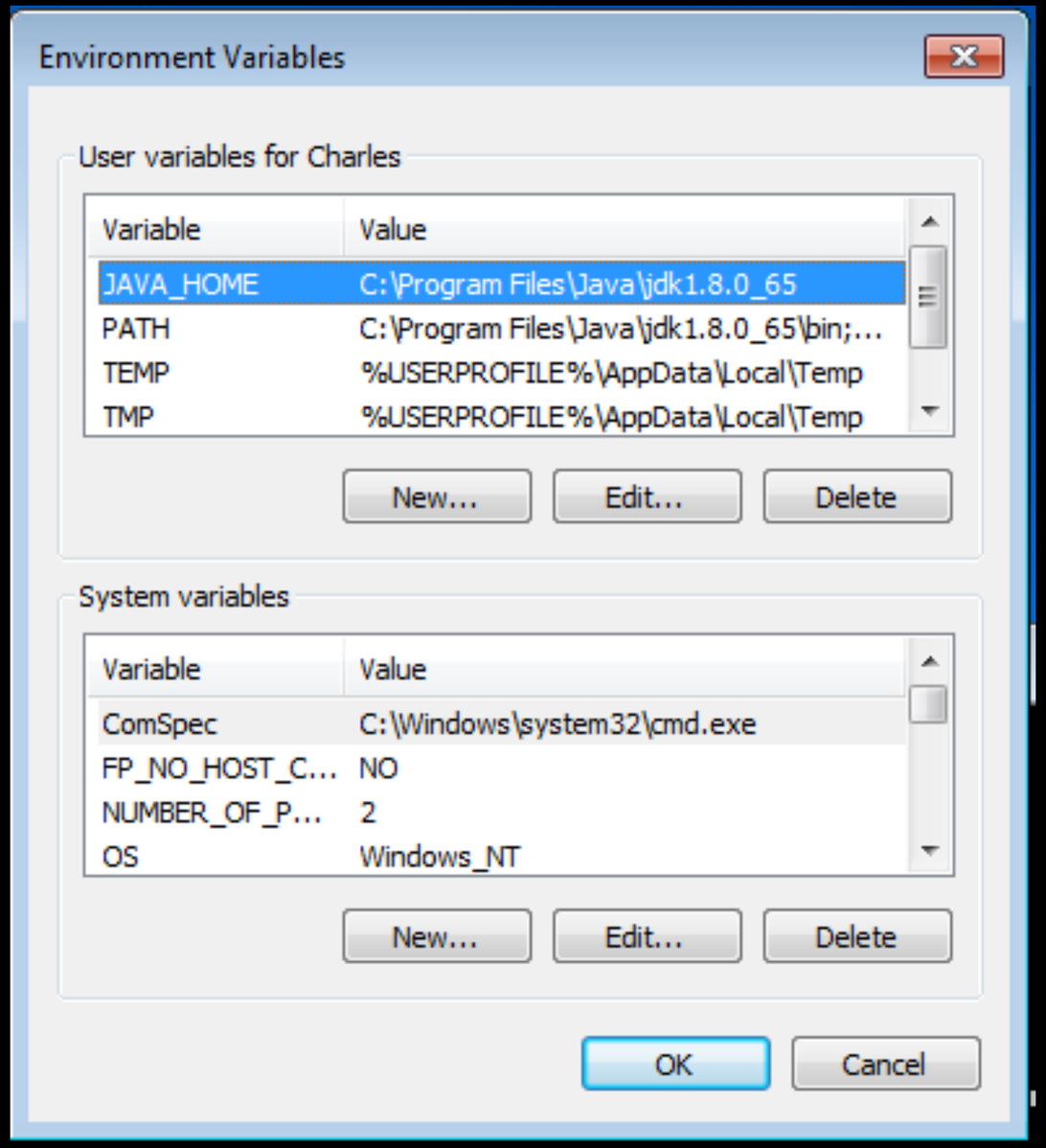
Embedded

Distributed

Step 1: Download Drill: drill.apache.org/download/

Drill Requirements

- Oracle Java SE Development Kit (JDK 7) or higher. (Verify this by opening a command prompt and typing `java -version`)
- On Windows machines, you will need to set the JAVA_HOME and PATH variables.


Environment Variables
User variables for Charles
Variable
Value
JAVA_HOME
C:\Program Files\Java\jdk1.8.0_65
PATH
C:\Program Files\Java\jdk1.8.0_65\bin;...
TEMP
%USERPROFILE%\AppData\Local\Temp
TMP
%USERPROFILE%\AppData\Local\Temp
New...
Edit...
Delete
System variables
Variable
Value
ComSpec
C:\Windows\system32\cmd.exe
FP_NO_HOST_C...
NO
NUMBER_OF_P...
2
OS
Windows_NT
New...
Edit...
Delete
OK
Cancel

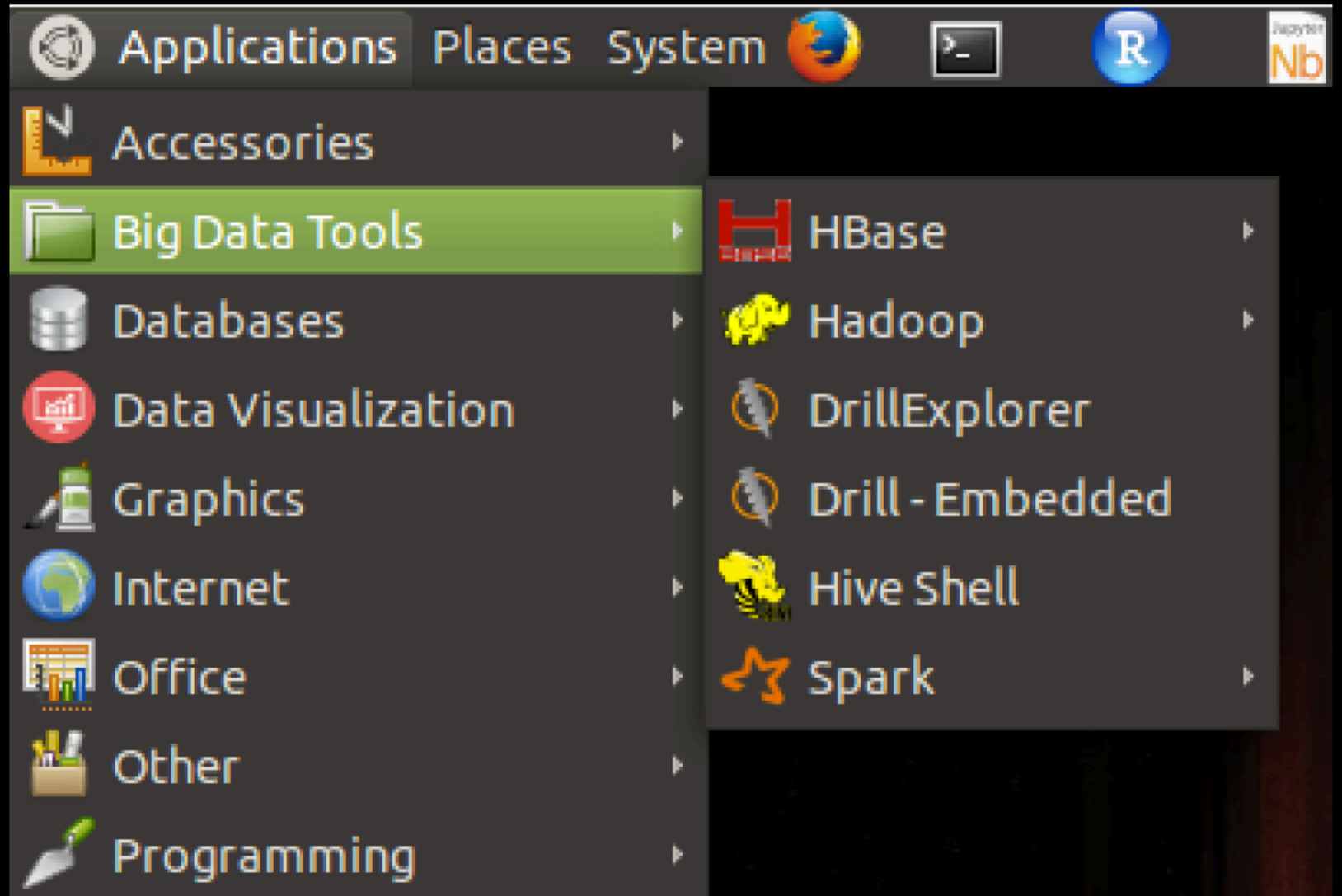
Starting Drill

Embedded Mode: For use on a standalone system

```
$./bin/drill-embedded
```

```
sqlline.bat -u "jdbc:drill:zk=local"
```

Starting Drill

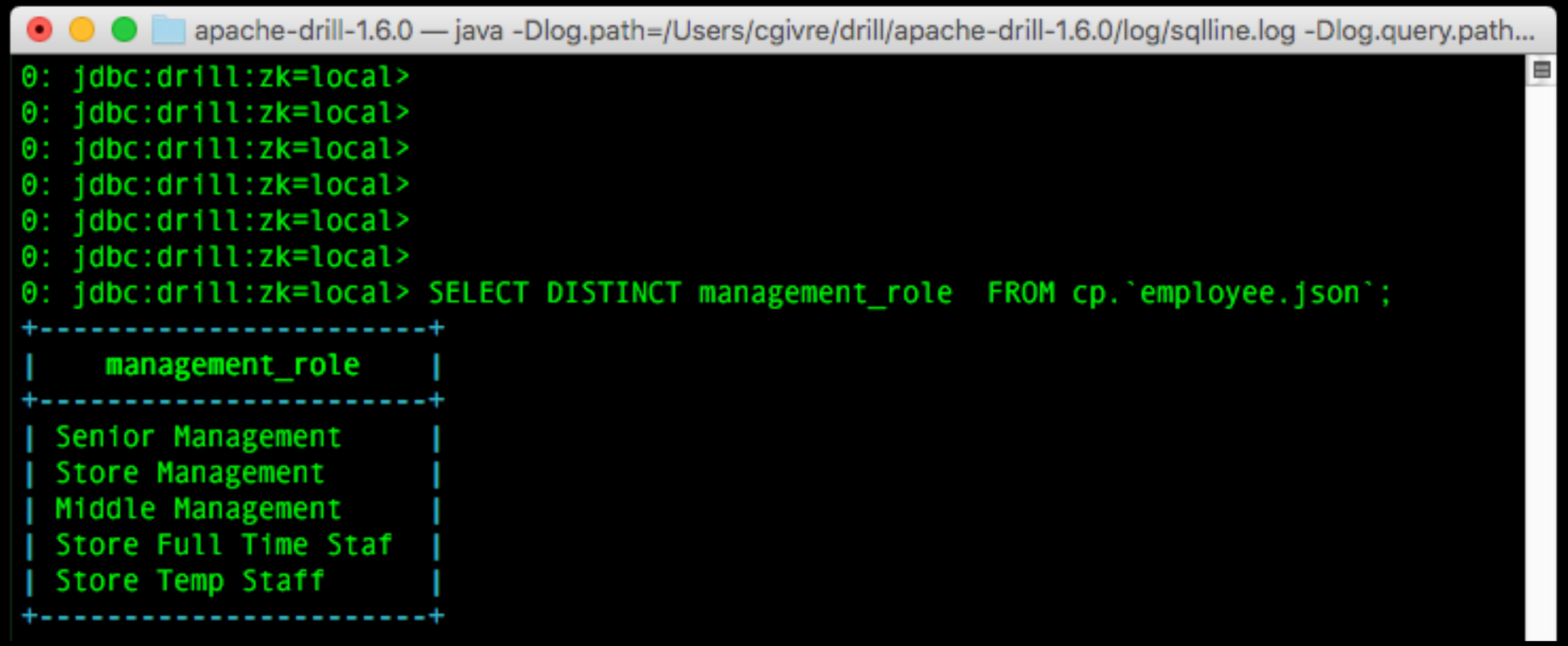


Drill's Command Line Interface

```
apache-drill-1.6.0 — java -Dlog.path=/Users/cgivre/drill/apache-drill-1.6.0/log/sqlline.log -Dl...
[Charless-MacBook-Pro:apache-drill-1.6.0 cgivre$ ./bin/drill-embedded
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M; s
upport was removed in 8.0
May 18, 2016 10:10:22 PM org.glassfish.jersey.server.ApplicationHandler initia
lize
INFO: Initiating Jersey application, version Jersey: 2.8 2014-04-29 01:25:26..
.
apache drill 1.6.0
"a little sql for your nosql"
0: jdbc:drill:zk=local>
```

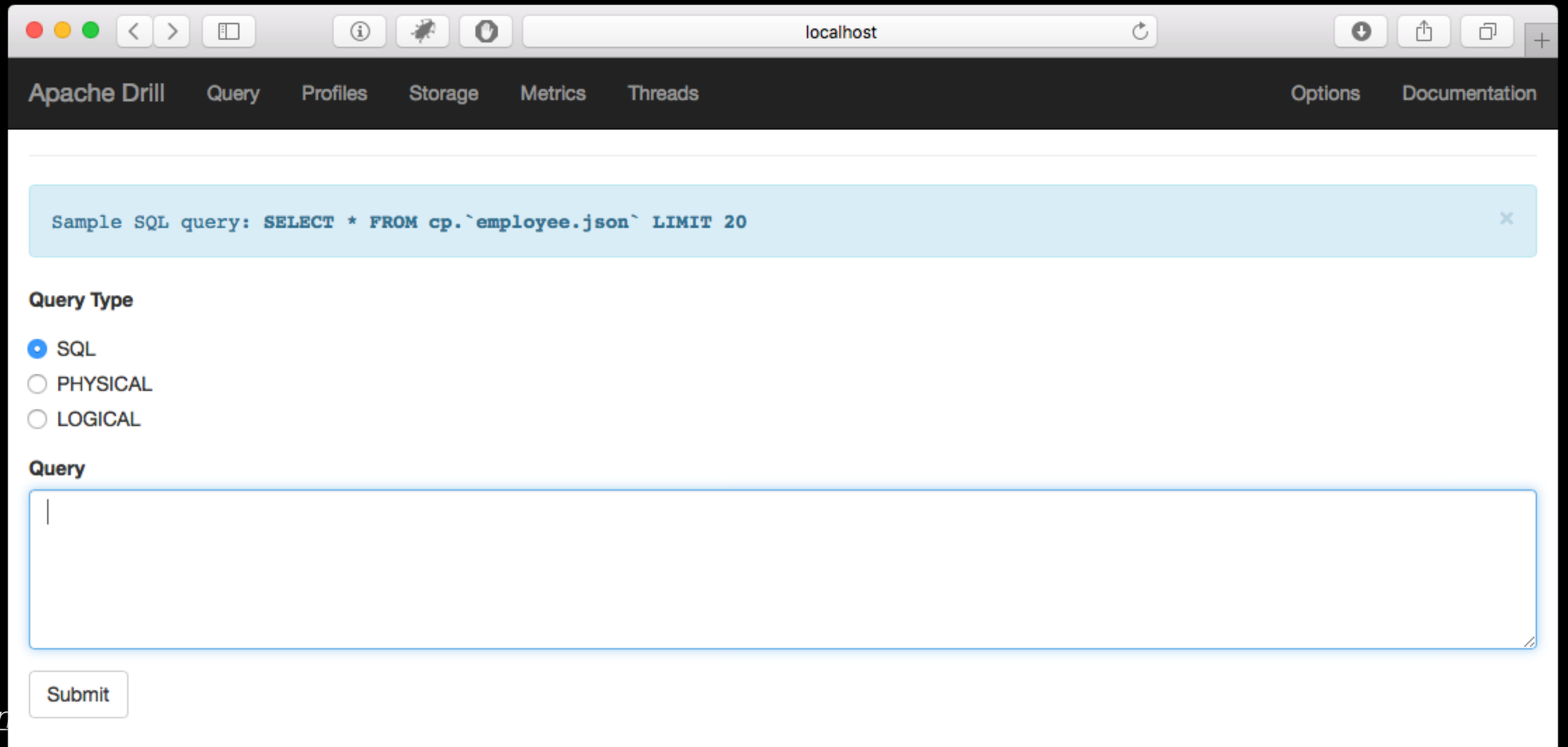
Drill's Command Line Interface

```
SELECT DISTINCT management_role FROM cp.`employee.json`;
```



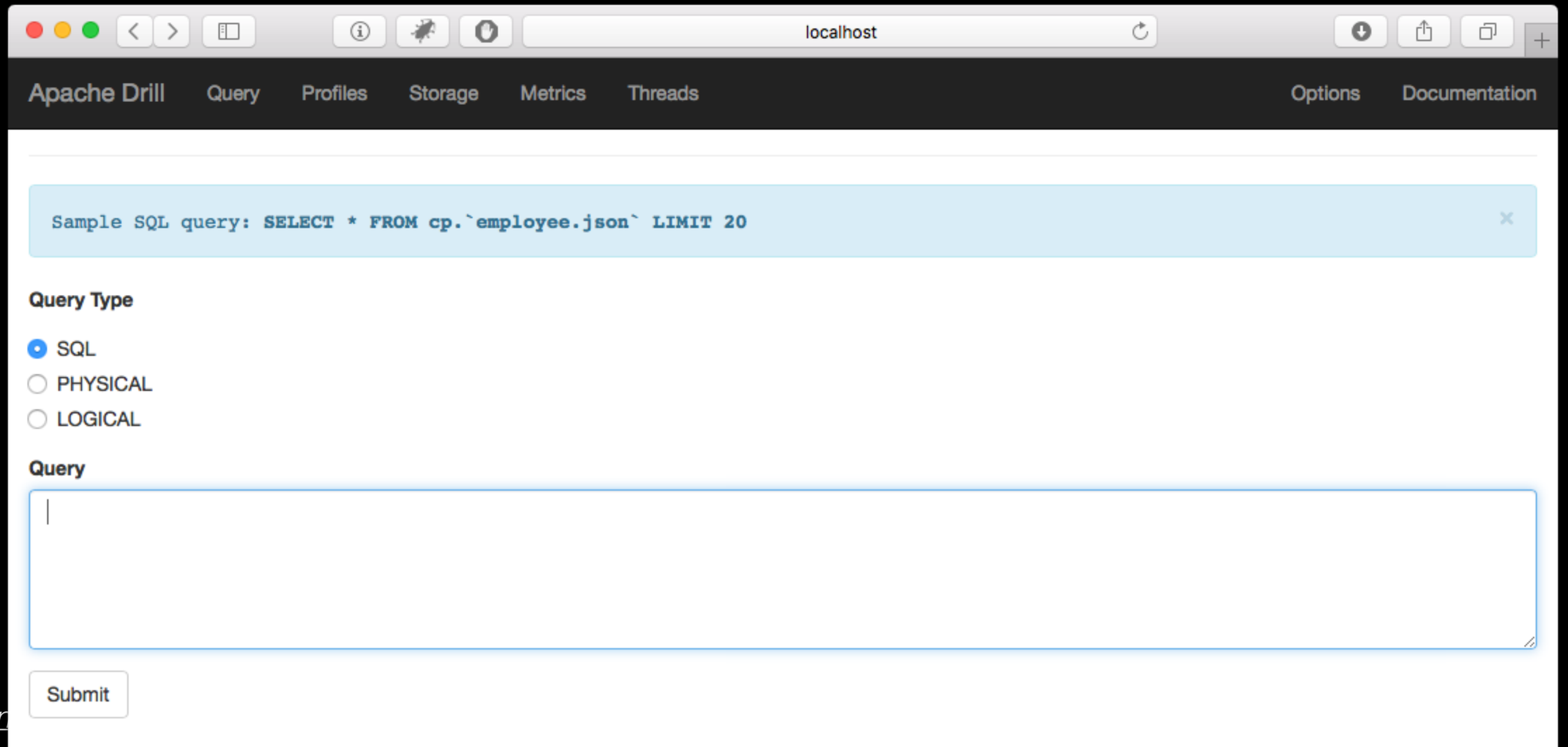
Drill Web UI

http://localhost:8047



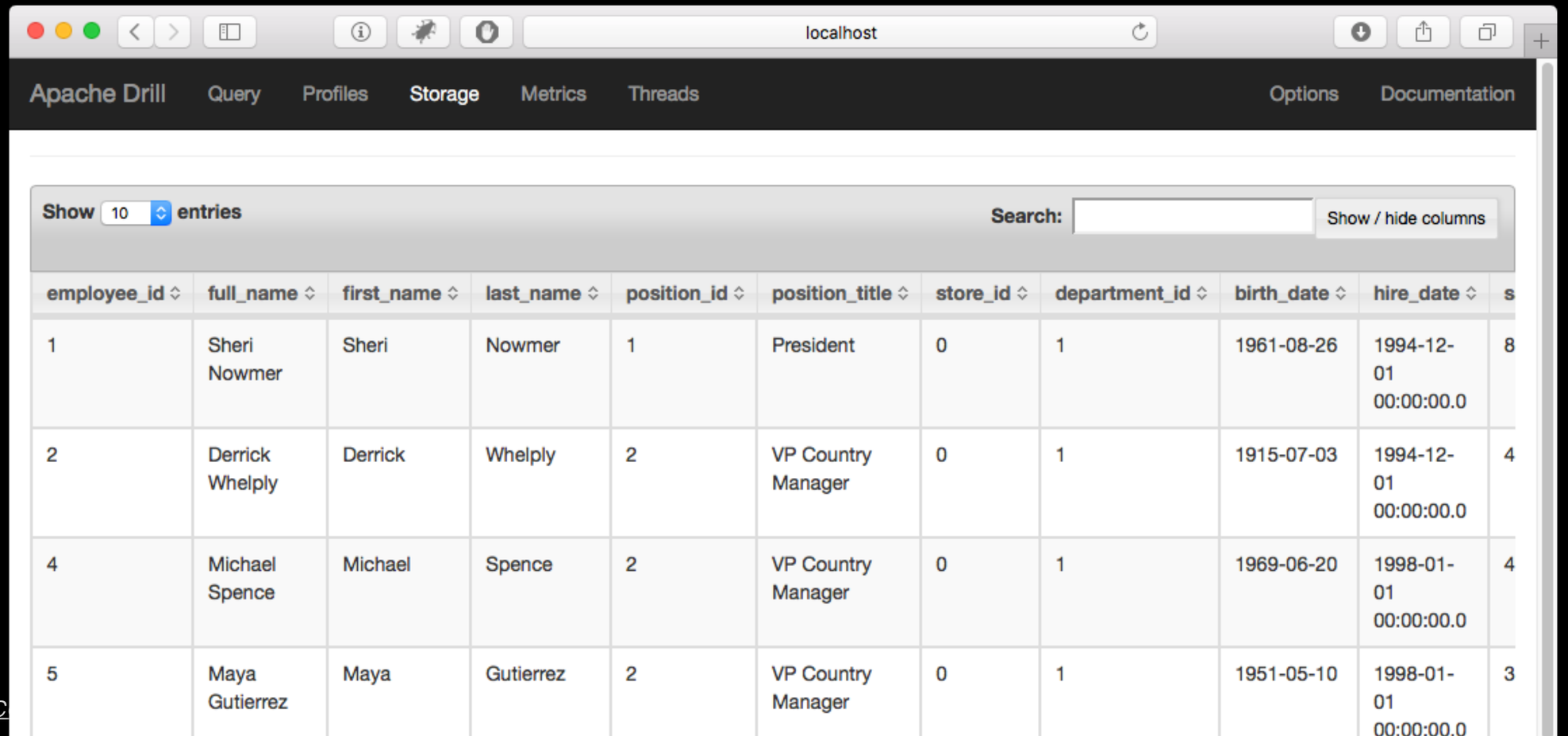
Drill Web UI

```
SELECT * FROM cp.`employee.json` LIMIT 20
```



Drill Web UI

```
SELECT * FROM cp.`employee.json` LIMIT 20
```

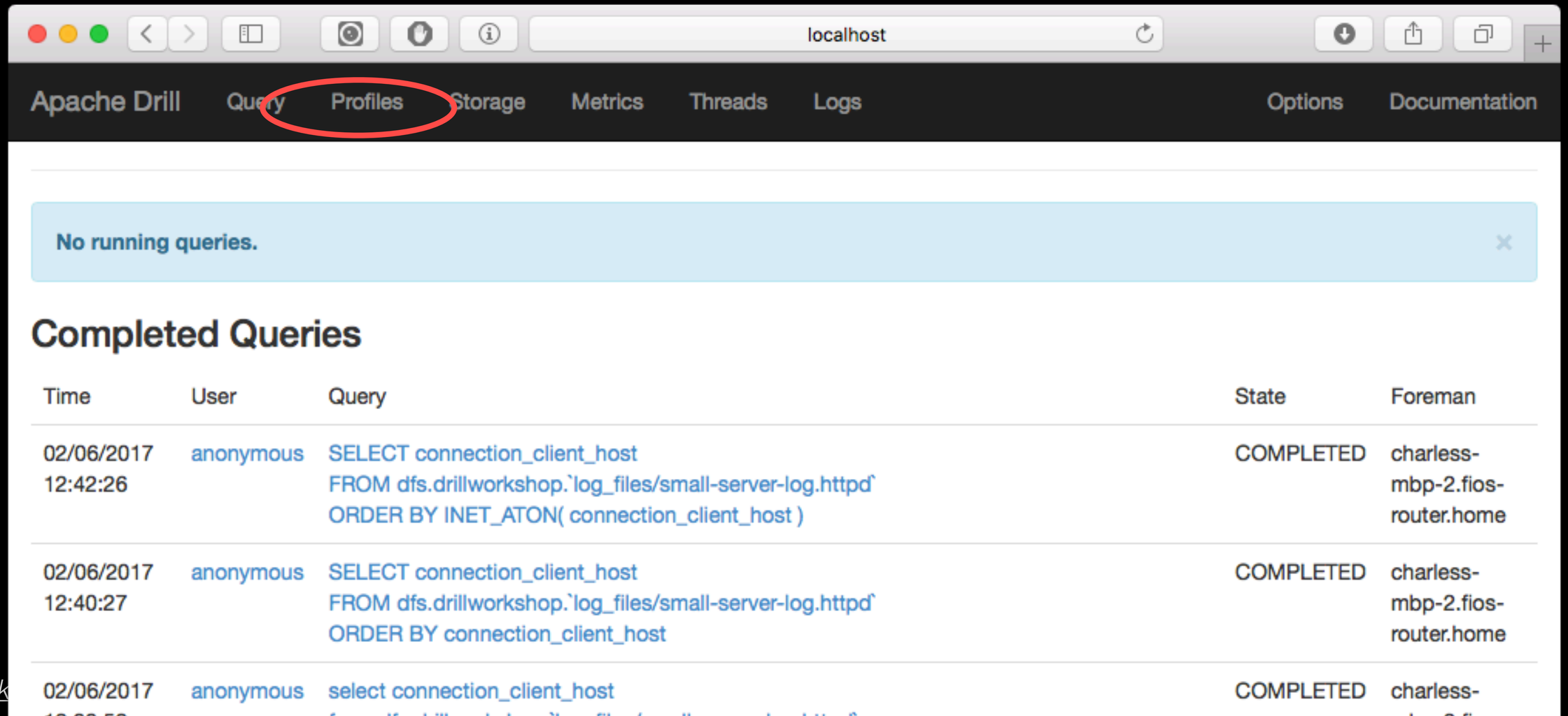


employee_id	full_name	first_name	last_name	position_id	position_title	store_id	department_id	birth_date	hire_date	s
1	Sheri Nowmer	Sheri	Nowmer	1	President	0	1	1961-08-26	1994-12-01 00:00:00.0	8
2	Derrick Whelply	Derrick	Whelply	2	VP Country Manager	0	1	1915-07-03	1994-12-01 00:00:00.0	4
4	Michael Spence	Michael	Spence	2	VP Country Manager	0	1	1969-06-20	1998-01-01 00:00:00.0	4
5	Maya Gutierrez	Maya	Gutierrez	2	VP Country Manager	0	1	1951-05-10	1998-01-01 00:00:00.0	3

Drill isn’t case sensitive*

* Except when Drill is case sensitive

Drill Web UI


localhost
Apache Drill
Query
Profiles
Storage
Metrics
Threads
Logs
Options
Documentation
No running queries.
Completed Queries
Time
User
Query
State
Foreman
02/06/2017 12:42:26
anonymous
SELECT connection_client_host FROM dfs.drillworkshop.`log_files/small-server-log.httpd` ORDER BY INET_ATON(connection_client_host)
COMPLETED
charless-mbp-2.fios-router.home
02/06/2017 12:40:27
anonymous
SELECT connection_client_host FROM dfs.drillworkshop.`log_files/small-server-log.httpd` ORDER BY connection_client_host
COMPLETED
charless-mbp-2.fios-router.home
02/06/2017
anonymous
select connection_client_host
COMPLETED
charless-

Drill Web UI

localhost

Apache Drill | Query | Profiles | Storage | Metrics | Threads | Logs | Options | Documentation

Enabled Storage Plugins

cp	Update	Disable
dfs	Update	Disable

Disabled Storage Plugins

hbase	Update	Enable
hive	Update	Enable
kudu	Update	Enable
mongo	Update	Enable
s3	Update	Enable

Drill Web UI

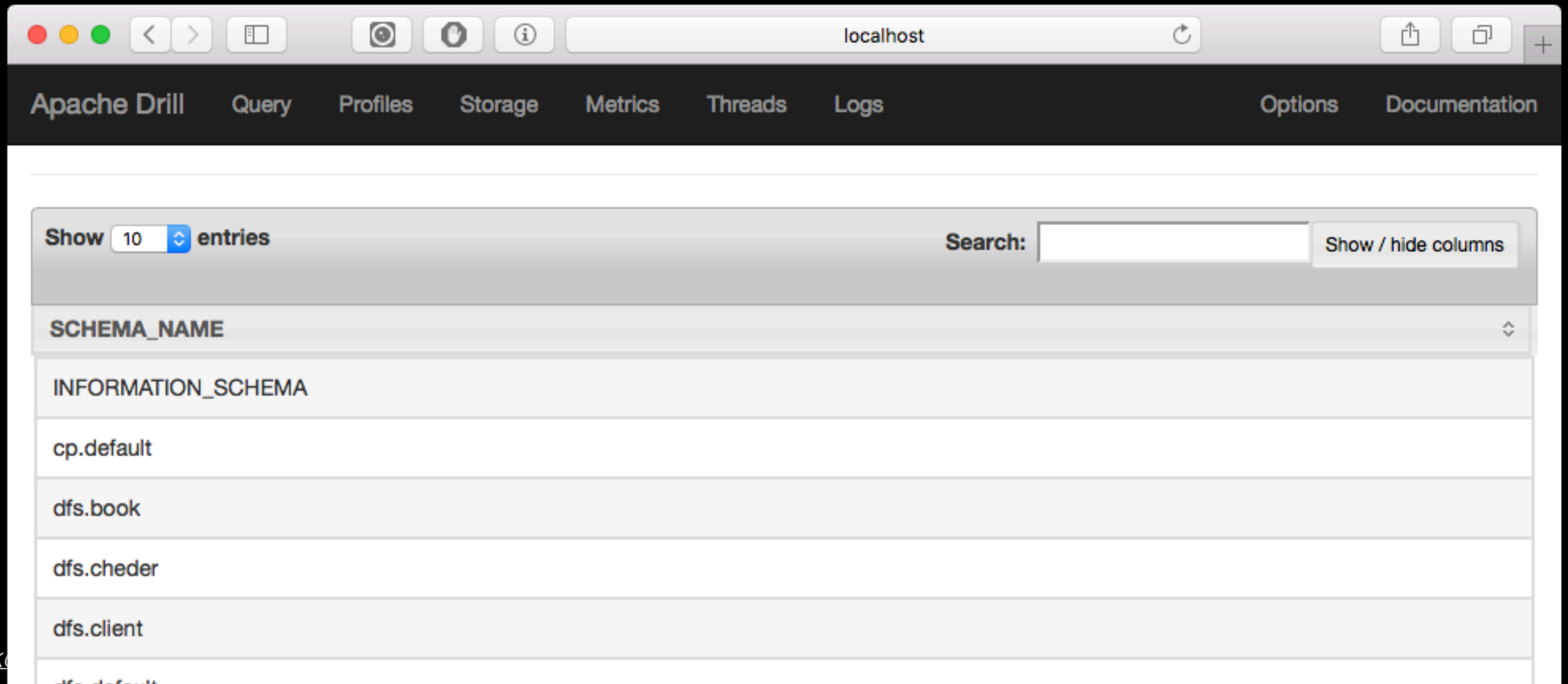
```
{
  "type": "file",
  "enabled": true,
  "connection": "file:///",
  "config": null,
  "workspaces": "workspaces": {
    "root": {
      "location": "/",
      "writable": false,
      "defaultInputFormat": null
    }…
  },
  "formats": {
    "csv": {
      "type": "text",
      "extensions": [
        "csv"
      ]      ]
    }
…
  }
```

Workspaces in Drill

- Workspaces are shortcuts to the file system. You'll want to use them when you have lengthy file paths.
- They work in any "file based" storage plugin (IE: S3, Hadoop, Local File System)

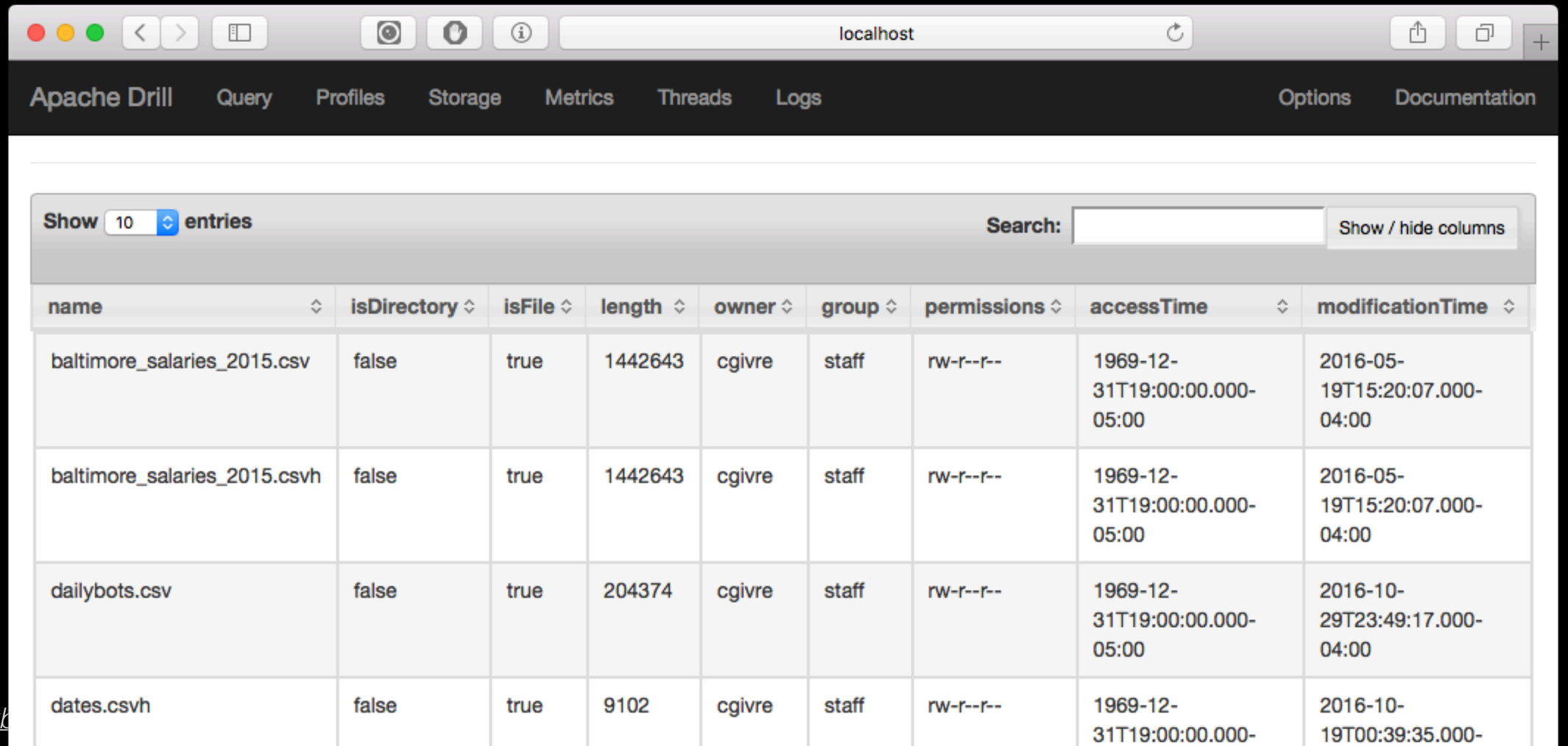
Drill Web UI

SHOW DATABASES



Drill Web UI

SHOW FILES IN <workspace>



name	isDirectory	isFile	length	owner	group	permissions	accessTime	modificationTime
baltimore_salaries_2015.csv	false	true	1442643	cgivre	staff	rw-r--r--	1969-12-31T19:00:00.000-05:00	2016-05-19T15:20:07.000-04:00
baltimore_salaries_2015.csvh	false	true	1442643	cgivre	staff	rw-r--r--	1969-12-31T19:00:00.000-05:00	2016-05-19T15:20:07.000-04:00
dailybots.csv	false	true	204374	cgivre	staff	rw-r--r--	1969-12-31T19:00:00.000-05:00	2016-10-29T23:49:17.000-04:00
dates.csvh	false	true	9102	cgivre	staff	rw-r--r--	1969-12-31T19:00:00.000-	2016-10-19T00:39:35.000-

Workspaces in Drill

```
SELECT field1, field2
FROM dfs.`/Users/cgivre/github/projects/drillclass/file1.csv`
```

or

```
SELECT field1, field2
FROM dfs.drilldata.`file1.csv`
```

In Class Exercise: Create a Workspace

In this exercise we are going to create a workspace called '**`drillclass`**', which we will use for future exercises.

1. First, download all the files from https://github.com/gtkcyber/drillworkshop and put them in a folder of your choice on your computer. **Remember the complete file path.**

2. Open the Drill Web UI and go to Storage->dfs->update

3. Paste the following into the 'workspaces' section and click update

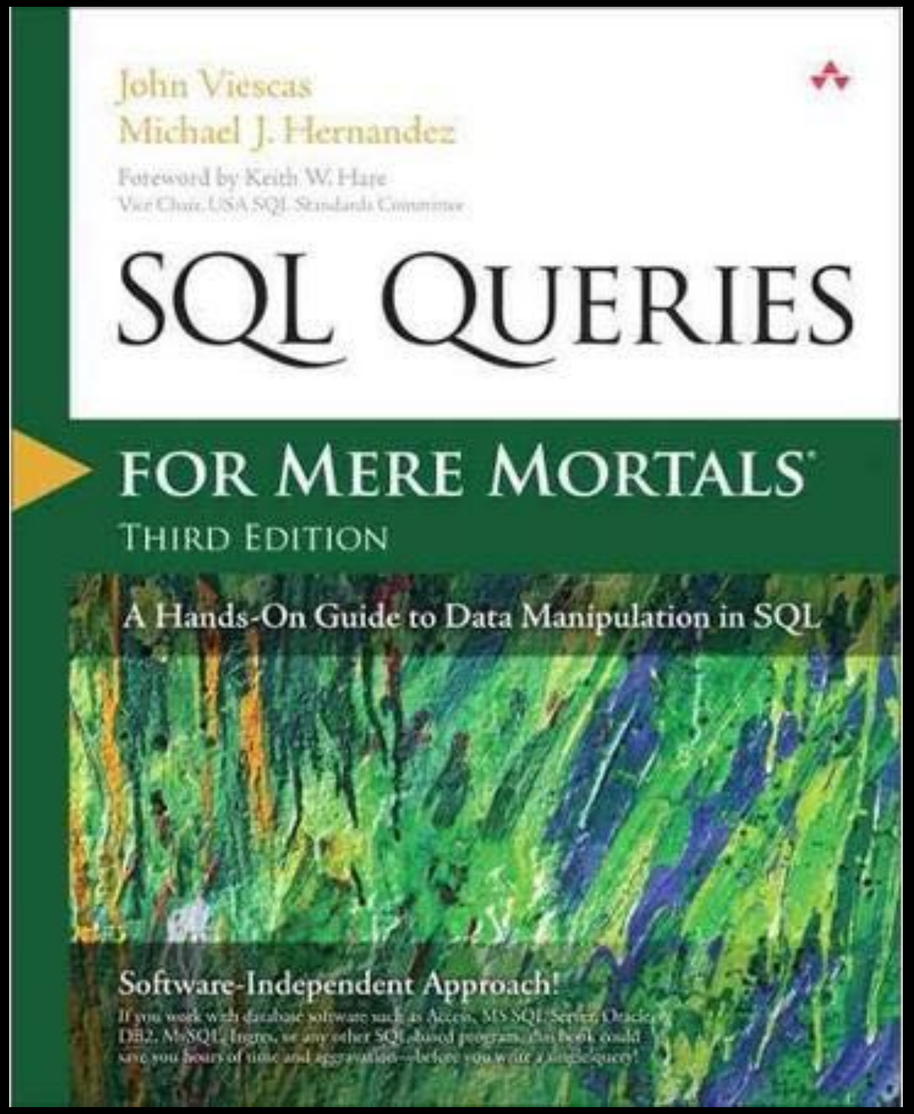
```
"drillclass": {
  "location": "<path to your files>",
  "writable": true,
  "defaultInputFormat": null
}
```

4. Execute a `show databases` query to verify that your workspace was added.

Querying Simple Delimited Data

Everything you need to know about SQL*… in 10 minutes

* well…not quite everything, but enough to get you through this session



http://amzn.to/2IlD8yi


```
SELECT <fields>
FROM <data source>
```

Please open people.csv

	A	B	C	D	E	F
1	id	first_name	last_name	email	gender	ip_address
2	1	Philip	Richardson	prichardson0@samsung.com	Male	76.20.6.56
3	2	Todd	James	tjames1@hostgator.com	Male	0.118.22.98
4	3	Jimmy	Mendoza	jmendoza2@reuters.com	Male	118.191.95.200
5	4	Jose	Morris	jmorris3@example.com	Male	106.180.20.171
6	5	Dorothy	Fernandez	dfernandez4@ask.com	Female	171.75.36.3
7	6	Patrick	Bradley	pbradley5@elpais.com	Male	20.244.138.119
8	7	Nicholas	Bishop	nbishop6@fotki.com	Male	80.54.11.247
9	8	Michael	Kelly	mkelly7@imageshack.us	Male	191.190.150.242
10	9	Russell	Coleman	rcoleman8@pbs.org	Male	50.176.191.152
11	10	Frances	Rodriguez	frodriguez9@github.io	Female	23.118.13.92
12	11	Nancy	Nelson	nnelsona@biglobe.ne.jp	Female	225.182.18.156
13	12	Theresa	Russell	trussellb@hexun.com	Female	16.14.226.63
14	13	Frances	Greene	fgreenec@sbwire.com	Female	192.88.64.253
15	14	Julia	Alvarez	jalvarezd@livejournal.com	Female	197.237.43.239

Giving a shout out to
https://www.mockaroo.com
which I used to generate most of my data for this class.


```
SELECT *
FROM <data source>
```


```
SELECT first_name,
last_name,
gender
FROM <data source>
```

Tip: Use BACK TICKS around field names in Drill

```
SELECT `first_name`,
`last_name`,
`gender`
FROM <data source>
```


```
SELECT `first_name`,
`last_name`,
`gender`
FROM <data source>
```

Querying Drill

```
FROM dfs.logs.`/data/customers.csv`
```

Storage Plugin

Workspace

Table

Querying Drill

```
FROM dfs.logs.`/data/customers.csv`
```

Storage Plugin — Workspace — Table

```
FROM dfs.`/var/www/mystore/sales/data/customers.csv`
```


```
SELECT `first_name`,
`last_name`,
`gender`
FROM dfs.drillclass.`people.csvh`
```

Try it yourself!!


```
SELECT `first_name`,
`last_name`,
`gender`
FROM dfs.drillclass.`people.csvh`
```

Apache Drill Query Profiles Storage Metrics Threads Logs Options Documentation

Show 10 entries Search: Show / hide columns

first_name	last_name	gender
Philip	Richardson	Male
Todd	James	Male
Jimmy	Mendoza	Male


```
SELECT <fields>
FROM <data source>
WHERE <logical condition>
```


```
SELECT `first_name`,
`last_name`,
`gender`
FROM dfs.drillclass.`people.csvh`
WHERE `gender` = 'Female'
```


```
SELECT `first_name`,
`last_name`,
`gender`
FROM dfs.drillclass.`people.csvh`
WHERE `gender` = 'Female'
```

Apache Drill | Query | Profiles | Storage | Metrics | Threads | Logs | Options | Documentation

Show 10 entries | Search: | Show / hide columns

first_name	last_name	gender
Dorothy	Fernandez	Female
Frances	Rodriguez	Female
Nancy	Nelson	Female


```
SELECT <fields>
FROM <data source>
WHERE <logical condition>
ORDER BY <field> (ASC|DESC)
```


```
SELECT `first_name`,
`last_name`,
`gender`
FROM dfs.drillclass.`people.csvh`
ORDER BY `last_name`, `first_name` ASC
```


```
SELECT `first_name`,
`last_name`,
`gender`
FROM dfs.drillclass.`people.csvh`
ORDER BY `last_name`, `first_name` ASC
```

Apache Drill | Query | Profiles | Storage | Metrics | Threads | Logs | Options | Documentation

Show 10 entries | Search: | Show / hide columns

first_name	last_name	gender
Julia	Alvarez	Female
Nicholas	Bishop	Male
Marie	Bradley	Female


```
SELECT
FUNCTION( <field> ) AS new_field
FROM <data source>
```


```
SELECT first_name,
LENGTH( `first_name` ) AS
fname_length
FROM dfs.drillclass.`people.csvh`
ORDER BY fname_length DESC
```


```
SELECT first_name,
LENGTH( `first_name` ) AS
fname_length
FROM dfs.drillclass.`people.csvh`
ORDER BY fname_length DESC
```

Apache Drill | Query | Profiles | Storage | Metrics | Threads | Logs | Options | Documentation

Show 10 entries | Search: | Show / hide columns

first_name	fname_length
Clarence	8
Nicholas	8
Theresa	7
Frances	7


```
SELECT <fields>
FROM <data source>
GROUP BY <field>
```


```
SELECT `gender`,
COUNT( * ) AS gender_count
FROM dfs.drillclass.`people.csvh`
GROUP BY `gender`
```


```
SELECT `gender`,
COUNT( * ) AS gender_count
FROM dfs.drillclass.`people.csvh`
GROUP BY `gender`
```

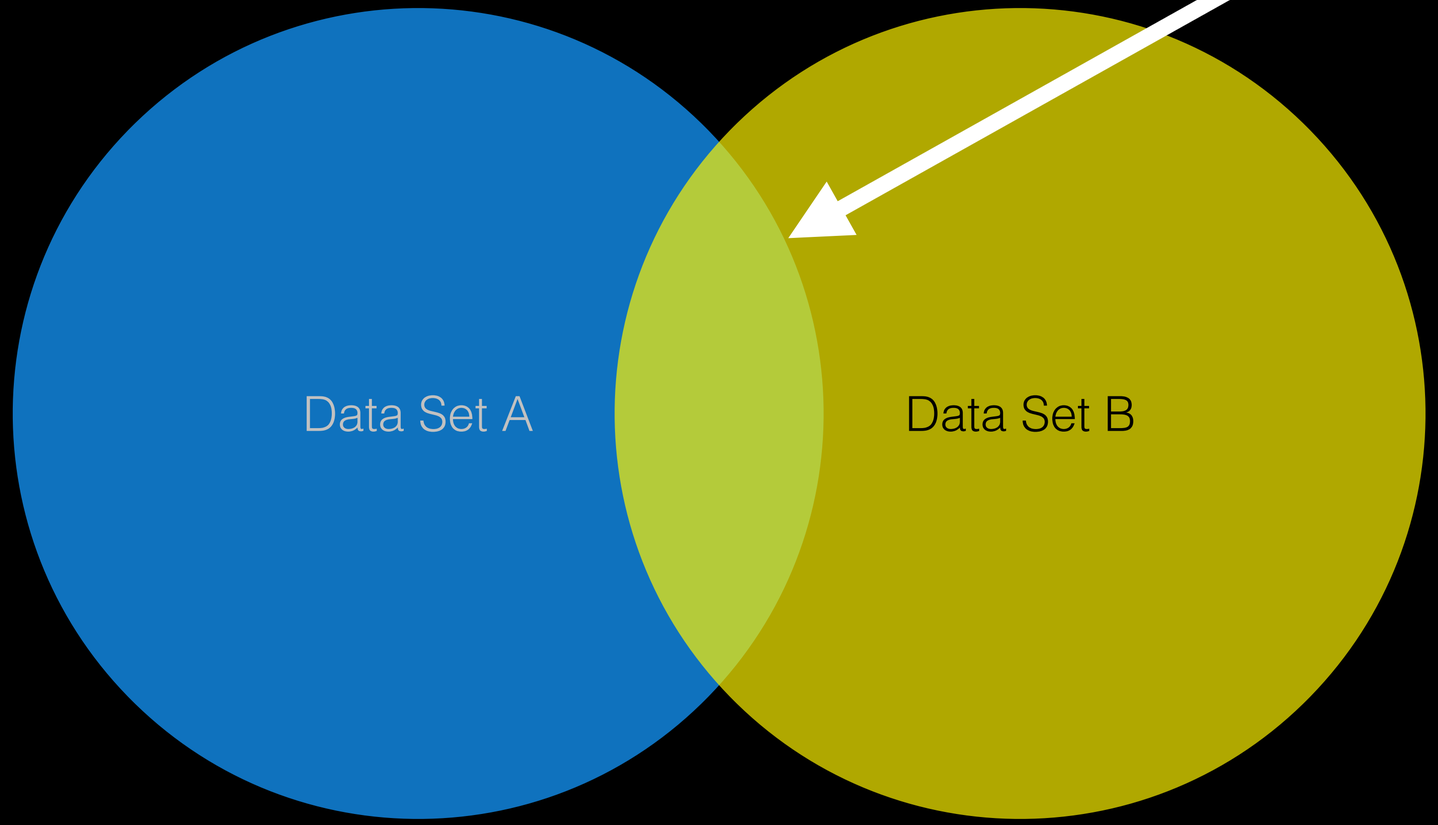
Apache Drill | Query | Profiles | Storage | Metrics | Threads | Logs | Options | Documentation

Show 10 entries | Search: | Show / hide columns

gender	gender_count
Male	11
Female	9

Joining Datasets

Referred to as an Inner Join

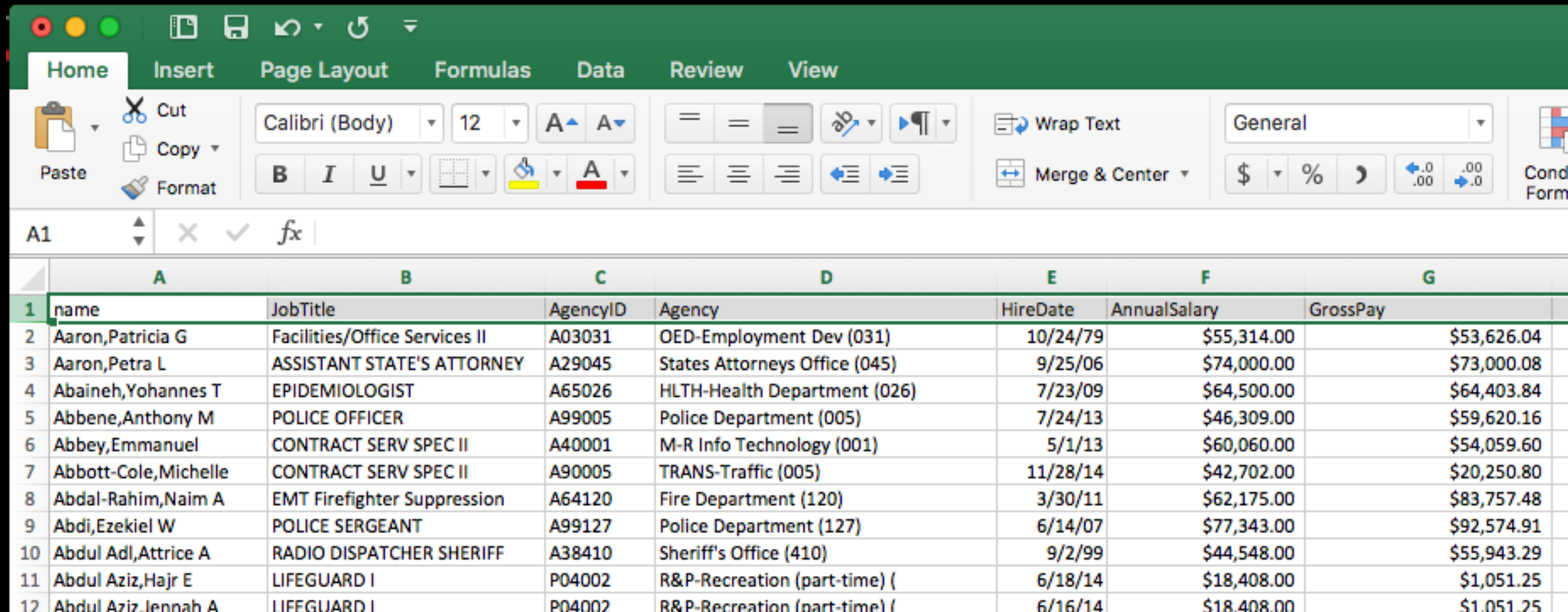



```
SELECT <fields>
FROM <data source 1> AS table1
INNER JOIN <data source 2> AS table2
ON table1.`id` = table2.`id`
```

Questions?

Querying Drill

Please take a look at baltimore_salaries_2016.csv.
This data is available at: http://bit.ly/balt-sal



	A	B	C	D	E	F	G
1	name	JobTitle	AgencyID	Agency	HireDate	AnnualSalary	GrossPay
2	Aaron,Patricia G	Facilities/Office Services II	A03031	OED-Employment Dev (031)	10/24/79	$55,314.00	$53,626.04
3	Aaron,Petra L	ASSISTANT STATE'S ATTORNEY	A29045	States Attorneys Office (045)	9/25/06	$74,000.00	$73,000.08
4	Abaineh,Yohannes T	EPIDEMIOLOGIST	A65026	HLTH-Health Department (026)	7/23/09	$64,500.00	$64,403.84
5	Abbene,Anthony M	POLICE OFFICER	A99005	Police Department (005)	7/24/13	$46,309.00	$59,620.16
6	Abbey,Emmanuel	CONTRACT SERV SPEC II	A40001	M-R Info Technology (001)	5/1/13	$60,060.00	$54,059.60
7	Abbott-Cole,Michelle	CONTRACT SERV SPEC II	A90005	TRANS-Traffic (005)	11/28/14	$42,702.00	$20,250.80
8	Abdal-Rahim,Naim A	EMT Firefighter Suppression	A64120	Fire Department (120)	3/30/11	$62,175.00	$83,757.48
9	Abdi,Ezekiel W	POLICE SERGEANT	A99127	Police Department (127)	6/14/07	$77,343.00	$92,574.91
10	Abdul Adl,Attrice A	RADIO DISPATCHER SHERIFF	A38410	Sheriff's Office (410)	9/2/99	$44,548.00	$55,943.29
11	Abdul Aziz,Hajr E	LIFEGUARD I	P04002	R&P-Recreation (part-time) (	6/18/14	$18,408.00	$1,051.25
12	Abdul Aziz,Jennah A	LIFEGUARD I	P04002	R&P-Recreation (part-time) (	6/16/14	$18,408.00	$1,051.25

In Class Exercise:
Create a Simple Report

For this exercise we will use the baltimore_salaries_2016.csvh file.

1. Create a query which returns each person's: name, jobtitle, and gross pay.

2. Create a report which contains each employee's name, job title, 2015 salary and 2016 salary. NOTE: This query requires the use of a JOIN.


```
SELECT EmpName, JobTitle, GrossPay
FROM dfs.drillclass.`baltimore_salaries_2016.csvh`
LIMIT 10
```

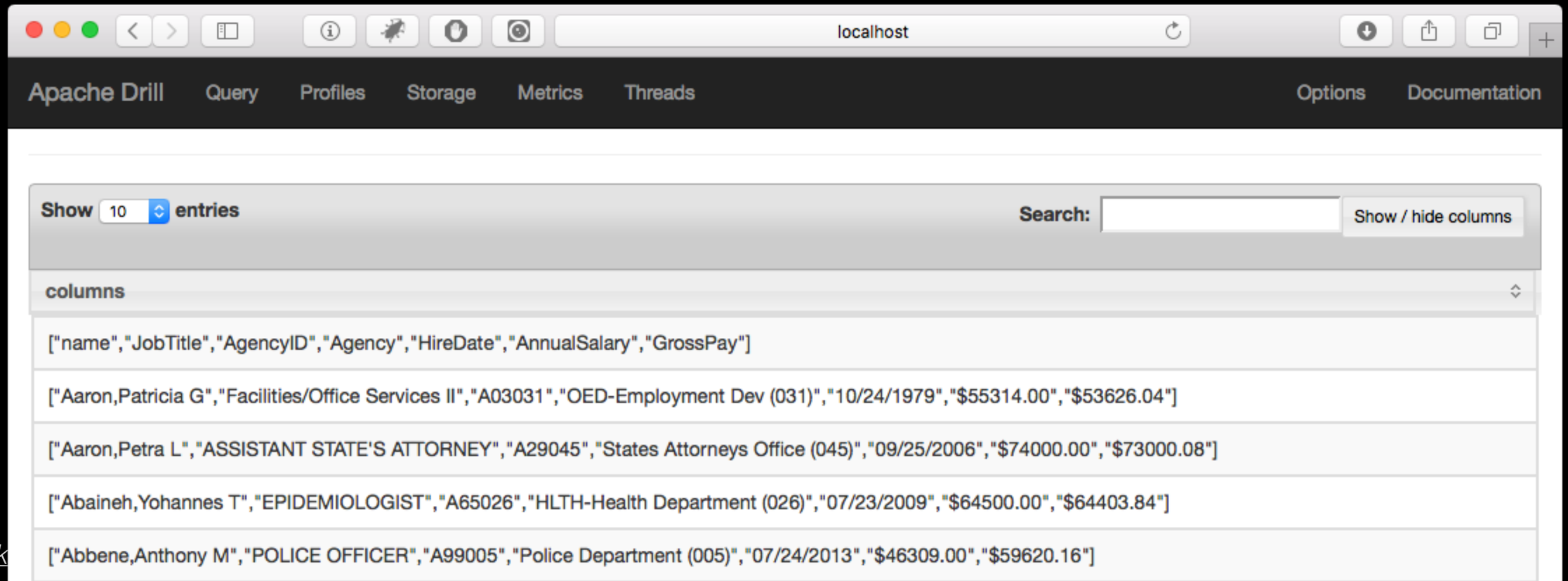

```
SELECT data2016.`EmpName`,
data2016.`JobTitle`,
data2016.`AnnualSalary` AS salary_2016,
data2015.`AnnualSalary` AS salary_2015
FROM dfs.drillclass.`baltimore_salaries_2016.csvh` AS data2016
INNER JOIN dfs.drillclass.`baltimore_salaries_2015.csvh` AS data2015
ON data2016.`EmpName` = data2015.`EmpName`
```

Querying Drill

```
SELECT *
FROM dfs.drillclass.`csv/baltimore_salaries_2016.csv`
LIMIT 10
```

Drill Data Types

```
SELECT *
FROM dfs.drillclass.`baltimore_salaries_2016.csv`
LIMIT 10
```



Drill Data Types

Simple Data Types

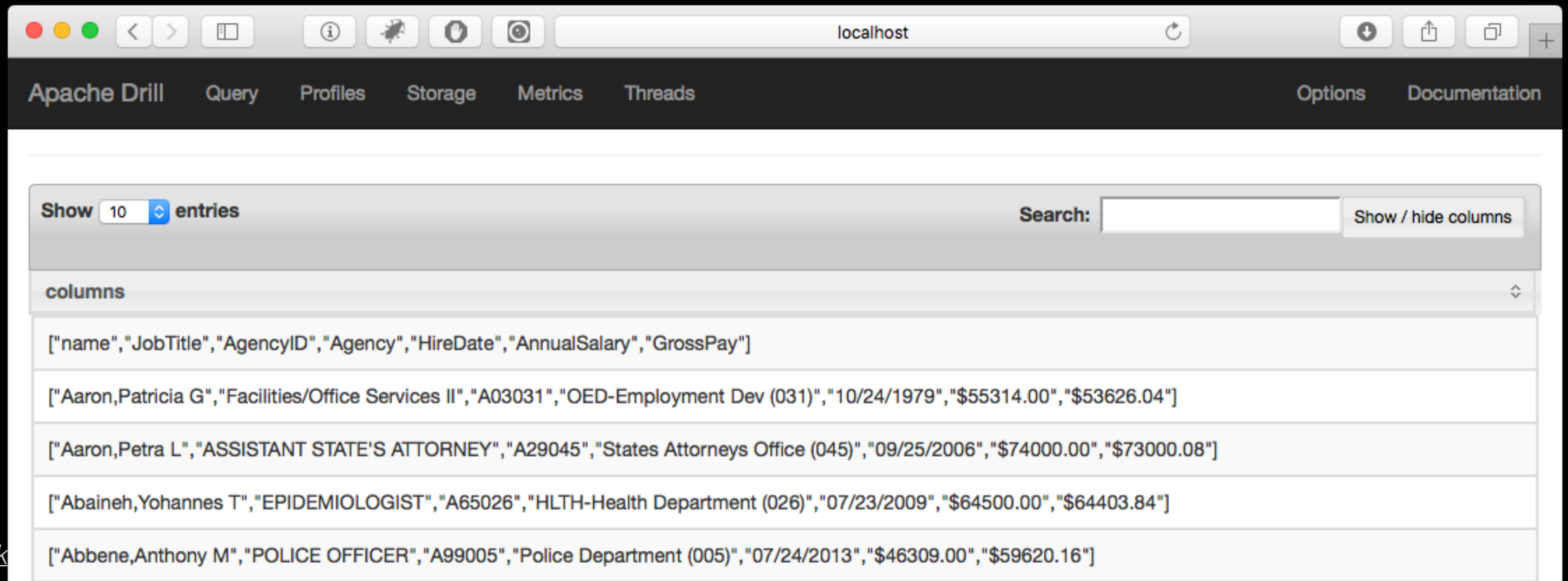
- Integer/BigInt/SmallInt
- Float/Decimal/Double
- Varchar/Binary
- Date/Time/Interval/Timestamp

Complex Data Types

- Arrays
- Maps

Querying Drill

["Aaron, Patricia G" "Facilities/Office Services"…]



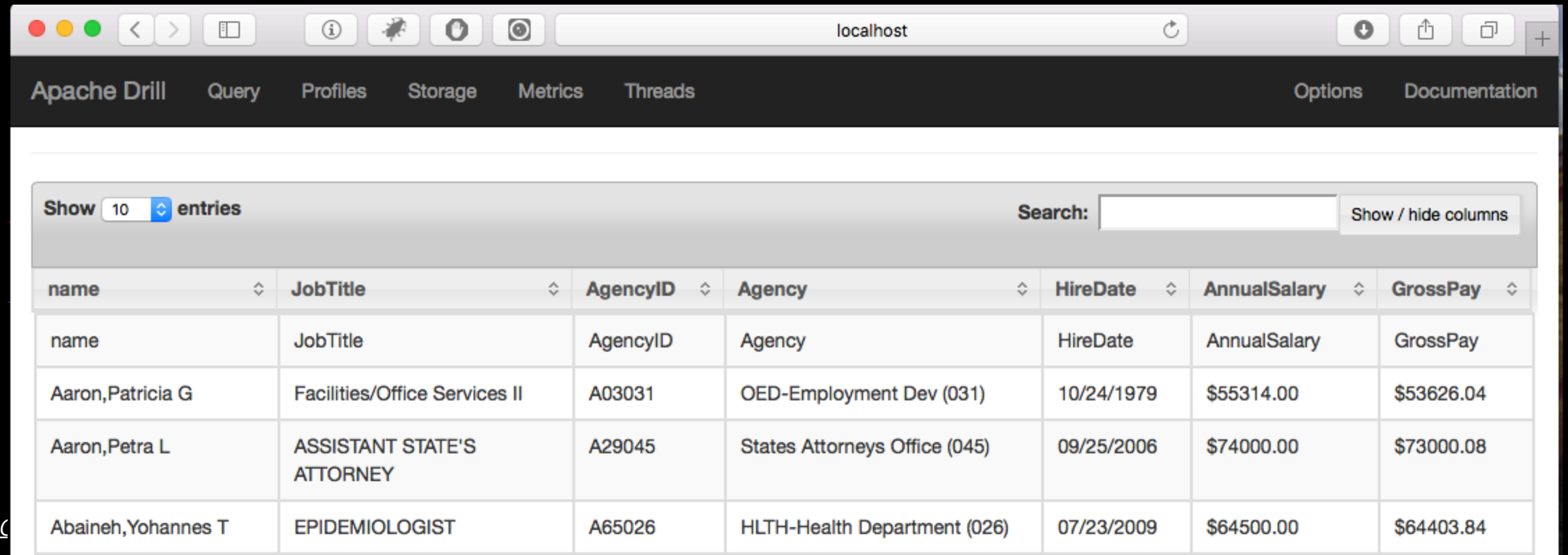
columns[*n*]

Querying Drill

```
SELECT columns[0] AS name,
columns[1] AS JobTitle,
columns[2] AS AgencyID,
columns[3] AS Agency,
columns[4] AS HireDate,
columns[5] AS AnnualSalary,
columns[6] AS GrossPay
FROM dfs.drillclass.`csv/baltimore_salaries_2016.csv`
LIMIT 10
```

Querying Drill

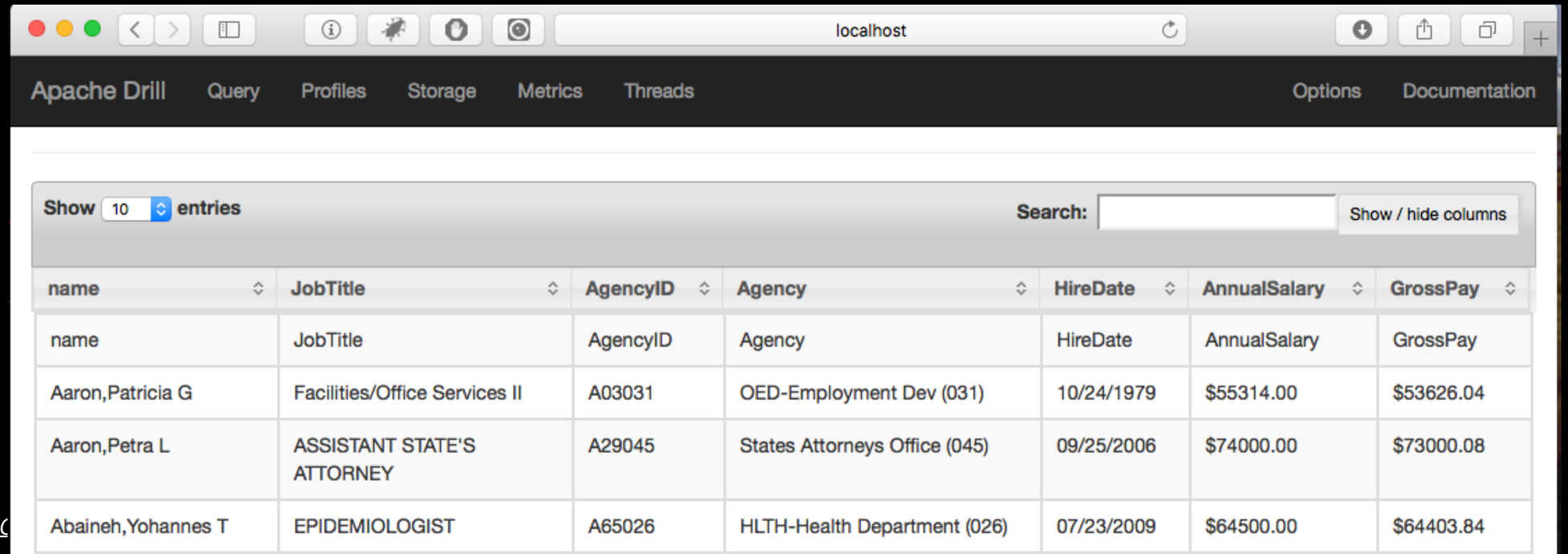
```
SELECT columns[0] AS name,
columns[1] AS JobTitle,
. . .
FROM dfs.drillclass.`csv/baltimore_salaries_2016.csv`
LIMIT 10
```



name	JobTitle	AgencyID	Agency	HireDate	AnnualSalary	GrossPay
name	JobTitle	AgencyID	Agency	HireDate	AnnualSalary	GrossPay
Aaron,Patricia G	Facilities/Office Services II	A03031	OED-Employment Dev (031)	10/24/1979	$55314.00	$53626.04
Aaron,Petra L	ASSISTANT STATE'S ATTORNEY	A29045	States Attorneys Office (045)	09/25/2006	$74000.00	$73000.08
Abaineh,Yohannes T	EPIDEMIOLOGIST	A65026	HLTH-Health Department (026)	07/23/2009	$64500.00	$64403.84

Querying Drill

```
SELECT columns[0] AS name,
columns[1] AS JobTitle,
. . .
FROM dfs.drillclass.`baltimore_salaries_2016.csv`
LIMIT 10
```



name	JobTitle	AgencyID	Agency	HireDate	AnnualSalary	GrossPay
name	JobTitle	AgencyID	Agency	HireDate	AnnualSalary	GrossPay
Aaron,Patricia G	Facilities/Office Services II	A03031	OED-Employment Dev (031)	10/24/1979	$55314.00	$53626.04
Aaron,Petra L	ASSISTANT STATE'S ATTORNEY	A29045	States Attorneys Office (045)	09/25/2006	$74000.00	$73000.08
Abaineh,Yohannes T	EPIDEMIOLOGIST	A65026	HLTH-Health Department (026)	07/23/2009	$64500.00	$64403.84

Querying Drill

```
"csvh": {
  "type": "text",
  "extensions": [
    "csvh"
  ],
  "extractHeader": true,
  "delimiter": ","
}
```

Querying Drill

File Extension	File Type
.psv	Pipe separated values
.csv	Comma separated value files
.csvh	Comma separated value with header
.tsv	Tab separated values
.json	JavaScript Object Notation files
.avro	Avro files (experimental)
.seq	Sequence Files

Querying Drill

Options	Description
comment	What character is a comment character
escape	Escape character
delimiter	The character used to delimit fields
quote	Character used to enclose fields
skipFirstLine	true/false
extractHeader	Reads the header from the CSV file

```
SELECT *
FROM table(
  dfs.drillclass.`baltimore_salaries_2016.csv`
  (
    type => 'text',
    extractHeader => true,
    fieldDelimiter => ‘,’
  )
)
```

Problem: Find the average salary of each Baltimore City job title

Aggregate Functions

Function	Argument Type	Return Type
AVG(expression)	Integer or Floating point	Floating point
COUNT(*)		BIGINT
COUNT([DISTINCT] <expression>)	any	BIGINT
MIN/MAX(<expression>)	Any numeric or date	same as argument
SUM(<expression>)	Any numeric or interval	same as argument

Querying Drill

```
SELECT JobTitle, AVG( AnnualSalary) AS avg_salary,
COUNT( DISTINCT name ) AS number
FROM drillclass.`baltimore_salaries_2016.csvh`
GROUP BY JobTitle
Order By avg_salary DESC
```

Querying Drill

Query Failed: An Error Occurred

org.apache.drill.common.exceptions.UserRemoteException: SYSTEM ERROR: SchemaChangeException: Failure while trying to materialize incoming schema. Errors: Error in expression at index -1. Error: Missing function implementation: [castINT(BIT-OPTIONAL)]. Full expression: --UNKNOWN EXPRESSION--.. Fragment 0:0 [Error Id: af88883b-f10a-4ea5-821d-5ff065628375 on 10.251.255.146:31010]

Querying Drill

```
SELECT JobTitle, AVG( AnnualSalary) AS avg_salary,
COUNT( DISTINCT name ) AS number
FROM dfs.drillclass.`baltimore_salaries_2016.csvh`
GROUP BY JobTitle
Order By avg_salary DESC
```

Querying Drill

```
SELECT JobTitle,
AVG( AnnualSalary ) AS avg_salary,
COUNT( DISTINCT name ) AS number
FROM dfs.drillclass.`baltimore_salaries_2016.csvh`
GROUP BY JobTitle
Order By avg_salary DESC
```

AnnualPay has extra characters

AnnualPay is a string

Querying Drill

Function	Return Type
BYTE_SUBSTR	BINARY or VARCHAR
CHAR_LENGTH	INTEGER
CONCAT	VARCHAR
ILIKE	BOOLEAN
INITCAP	VARCHAR
LENGTH	INTEGER
LOWER	VARCHAR
LPAD	VARCHAR
LTRIM	VARCHAR
POSITION	INTEGER
REGEXP_REPLACE	VARCHAR
RPAD	VARCHAR
RTRIM	VARCHAR
SPLIT	ARRAY
STRPOS	INTEGER
SUBSTR	VARCHAR
TRIM	VARCHAR
UPPER	VARCHAR

In Class Exercise:
Clean the field.

In this exercise you will use one of the string functions to remove the dollar sign from the ‘AnnualPay’ column.

Complete documentation can be found here:

https://drill.apache.org/docs/string-manipulation/

In Class Exercise:
Clean the field.

In this exercise you will use one of the string functions to remove the dollar sign from the ‘AnnualPay’ column.

Complete documentation can be found here:

https://drill.apache.org/docs/string-manipulation/

```
SELECT LTRIM( AnnualPay, '$' ) AS annualPay
FROM dfs.drillclass.`baltimore_salaries_2016.csvh`
```

Drill Data Types

Data type	Description
Bigint	8 byte signed integer
Binary	Variable length byte string
Boolean	True/false
Date	yyyy-mm-dd
Double / Float	8 or 4 byte floating point number
Integer	4 byte signed integer
Interval	A day-time or year-month interval
Time	HH:mm:ss
Timestamp	JDBC Timestamp
Varchar	UTF-8 encoded variable length string

cast(<expression> AS <data type>)

In Class Exercise: Convert to a number

In this exercise you will use the cast() function to convert AnnualPay into a number.

Complete documentation can be found here:

https://drill.apache.org/docs/data-type-conversion/#cast

In Class Exercise: Convert to a number

In this exercise you will use the cast() function to convert AnnualPay into a number.

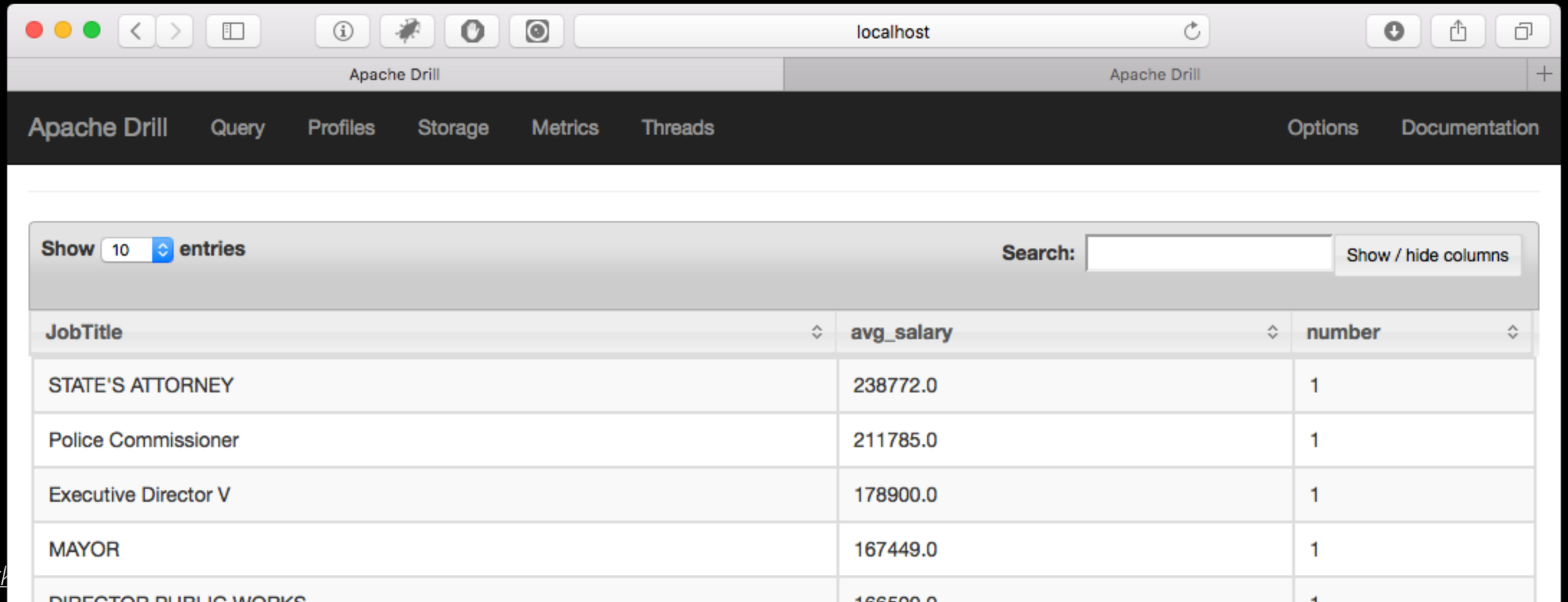
Complete documentation can be found here:

https://drill.apache.org/docs/data-type-conversion/#cast

```
SELECT CAST( LTRIM( AnnualPay, '$' ) AS FLOAT ) AS
annualPay
FROM dfs.drillclass.`csv/baltimore_salaries_2016.csvh`
```


```
SELECT JobTitle,
AVG(
   CAST(
      LTRIM( AnnualSalary, '$' ) AS FLOAT ) ) AS avg_salary,
COUNT( DISTINCT name ) AS number
FROM dfs.drillclass.`baltimore_salaries_2016.csvh`
GROUP BY JobTitle
Order By avg_salary DESC
```


```
SELECT JobTitle,
AVG( CAST( LTRIM( AnnualSalary, '$' ) AS FLOAT) ) AS avg_salary,
COUNT( DISTINCT name ) AS number
FROM dfs.drillclass.`baltimore_salaries_2016.csvh`
GROUP BY JobTitle
Order By avg_salary DESC
```



JobTitle	avg_salary	number
STATE'S ATTORNEY	238772.0	1
Police Commissioner	211785.0	1
Executive Director V	178900.0	1
MAYOR	167449.0	1
DIRECTOR PUBLIC WORKS	166500.0	1

TO_NUMBER(<field>, <format>)

TO_NUMBER(<field>, <format>)

Symbol	Meaning
0	Digit
#	Digit, zero shows as absent
.	Decimal separator or monetary separator
-	Minus Sign
,	Grouping Separator
%	Multiply by 100 and show as percentage
‰ \u2030	Multiply by 1000 and show as per mille value
¤ \u00A4	Currency symbol

In Class Exercise:
Convert to a number using **TO_NUMBER()**

In this exercise you will use the **TO_NUMBER()** function to convert AnnualPay into a numeric field.

Complete documentation can be found here:

https://drill.apache.org/docs/data-type-conversion/#to_number

In Class Exercise: Convert to a number using **TO_NUMBER()**

In this exercise you will use the **TO_NUMBER()** function to convert AnnualPay into a numeric field.

Complete documentation can be found here:

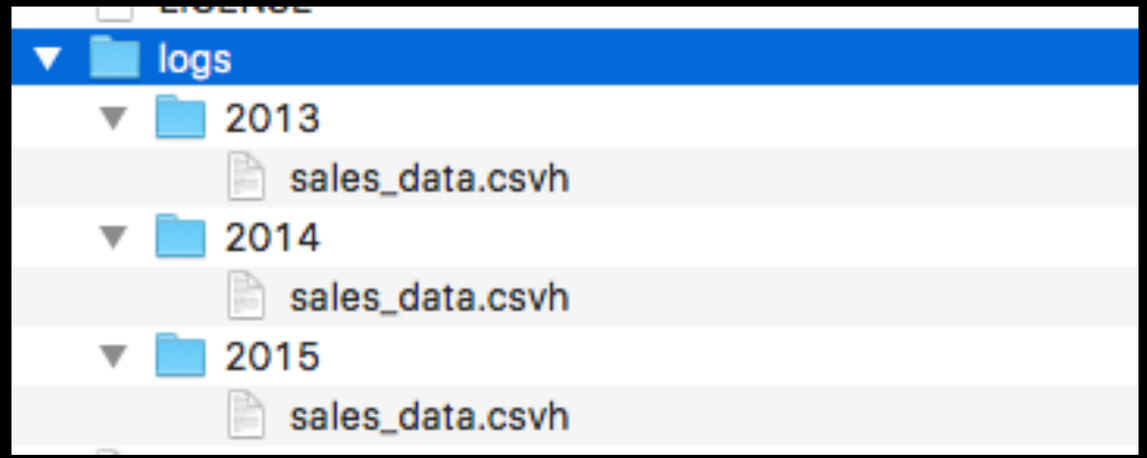
https://drill.apache.org/docs/data-type-conversion/#to_number

```
SELECT JobTitle,
AVG( TO_NUMBER( AnnualSalary, '¤' )) AS avg_salary,
COUNT( DISTINCT `EmpName` ) AS number
FROM dfs.drillclass.`baltimore_salaries_2016.csvh`
GROUP BY JobTitle
Order BY avg_salary DESC
```

Problem: You have files spread across many directories which you would like to analyze

Problem: You have multiple log files which you would like to analyze

- In the sample data files, there is a folder called ‘logs’ which contains the following structure:




```
SELECT *
FROM dfs.drillclass.`logs/`
LIMIT 10
```


```
SELECT *
FROM dfs.drillclass.`logs/`
LIMIT 10
```

localhost

Apache Drill | Apache Drill | Apache Drill

Apache Drill Query Profiles Storage Metrics Threads Options Documentation

Show 10 entries Search: Show / hide columns

customer_id	item_count	amount_spent	dir0
1169	2	1.05	2013
813	4	9.76	2013
373	1	6.69	2013
877	3	6.28	2013
959	4	1.74	2013

dir***n*** accesses the subdirectories

dir*n* accesses the subdirectories

```
SELECT *
FROM dfs.drilldata.`logs/`
WHERE dir0 = '2013'
```

Directory Functions

Function	Description
MAXDIR(), MINDIR()	Limit query to the first or last directory
IMAXDIR(), IMINDIR()	Limit query to the first or last directory in case insensitive order.

```
WHERE dir<n> = MAXDIR('<plugin>.<workspace>', '<filename>')
```

In Class Exercise:

Find the total number of items sold by year and the total dollar sales in each year.

HINT: Don’t forget to CAST() the fields to appropriate data types

In Class Exercise:

Find the total number of items sold by year and the total dollar sales in each year.

HINT: Don't forget to CAST() the fields to appropriate data types

```
SELECT dir0 AS data_year,
SUM( CAST( item_count AS INTEGER ) ) as total_items,
SUM( CAST( amount_spent AS FLOAT ) ) as total_sales
FROM dfs.drillclass.`logs/`
GROUP BY dir0
```

Questions?

Exercises

Using the Baltimore Salaries dataset write queries that answer the following questions:

1. In 2016, calculate the average difference in GrossPay and Annual Salary by Agency. HINT: Include `WHERE NOT( GrossPay ='' )` in your query. For extra credit, calculate the number of people in each Agency, and the min/max for the salary delta as well.

2. Find the top 10 individuals whose salaries changed the most between 2015 and 2016, both gain and loss.

3. (Optional Extra Credit) Using the various string manipulation functions, **split** the name function into two columns for the last name and first name. HINT: Don't overthink this, and review the sides about the columns array if you get stuck.

Homework

Using the Baltimore Salaries dataset write queries that answer the following questions:

1. In 2016, calculate the average difference in GrossPay and Annual Salary by Agency. HINT: Include `WHERE NOT( GrossPay ='' )` in your query. For extra credit, calculate the number of people in each Agency, and the min/max for the salary delta as well.

2. Find the top 10 individuals whose salaries changed the most between 2015 and 2016, both gain and loss.

3. (Optional Extra Credit) Using the various string manipulation functions, **split** the name function into two columns for the last name and first name. HINT: Don't overthink this, and review the sides about the columns array if you get stuck.

Exercises

Using the Baltimore Salaries dataset write queries that answer the following questions:

1. In 2016, calculate the average difference in GrossPay and Annual Salary by Agency. HINT: Include **`WHERE NOT( GrossPay ='' )`** in your query. For extra credit, calculate the number of people in each Agency, and the min/max for the salary delta as well.

```
SELECT Agency,
AVG( TO_NUMBER( `AnnualSalary`, '¤' ) - TO_NUMBER( `GrossPay`, '¤' )) AS avg_SalaryDelta,
COUNT( DISTINCT `EmpName` ) as emp_count,
MIN( TO_NUMBER( `AnnualSalary`, '¤' ) - TO_NUMBER( `GrossPay`, '¤' ) ) min_salary_delta,
MAX( TO_NUMBER( `AnnualSalary`, '¤' ) - TO_NUMBER( `GrossPay`, '¤' ) ) max_salary_delta
FROM dfs.drillclass.`baltimore_salaries_2016.csvh`
WHERE NOT( GrossPay ='' )
GROUP BY Agency
ORDER BY avg_SalaryDelta DESC
```

Exercises

Find the top 10 individuals whose salaries changed the most between 2015 and 2016, both gain and loss.

```
SELECT data2016.`EmpName`,
data2016.`JobTitle` AS JobTitle_2016,
data2015.`JobTitle` AS JobTitle_2015,
data2016.`AnnualSalary` AS salary_2016,
data2015.`AnnualSalary` AS salary_2015,
(TO_NUMBER( data2016.`AnnualSalary`, '¤' ) -
TO_NUMBER( data2015.`AnnualSalary`, '¤' )) AS salary_delta
FROM dfs.drillclass.`baltimore_salaries_2016.csvh` AS data2016
INNER JOIN dfs.drillclass.`baltimore_salaries_2015.csvh` AS
data2015
ON data2016.`EmpName` = data2015.`EmpName`
ORDER BY salary_delta DESC
LIMIT 10
```

Exercises

(Optional Extra Credit) Using the various string manipulation functions, **split** the name function into two columns for the last name and first name. HINT: Don't overthink this, and review the sides about the columns array if you get stuck.

```
SELECT `EmpName`,
SPLIT( `EmpName`,',' )[0] AS last_name,
SPLIT( `EmpName`,',' )[1] AS first_name
FROM dfs.drillclass.`baltimore_salaries_2016.csvh`
```

Working with Dates & Times

Working with Dates & Times

CAST(<field> **AS DATE**)

CAST(<field> **AS TIME**)

Working with Dates & Times

TO_DATE(<field>, '<format>')

TO_TIMESTAMP(<field>, '<format>')

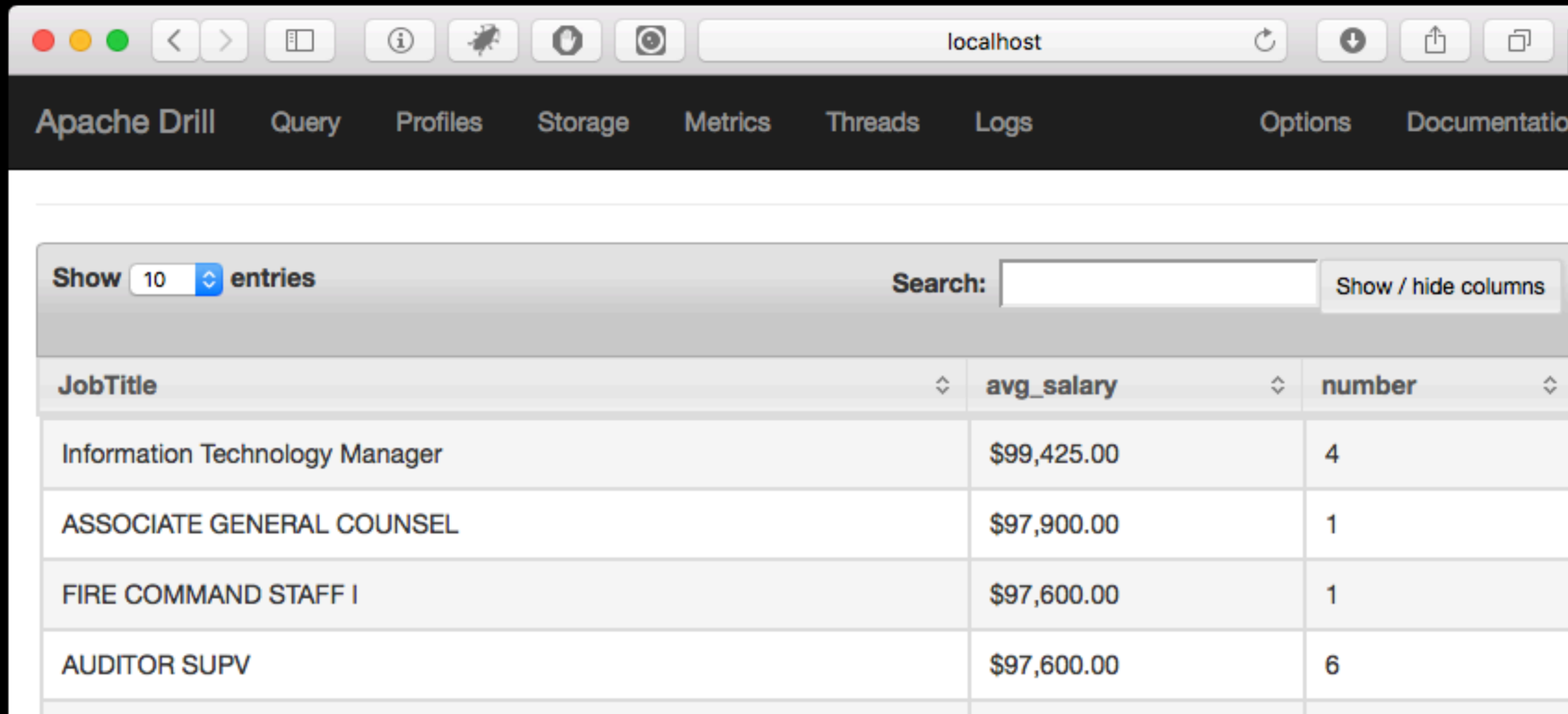
Working with Dates & Times

Symbol	Meaning	Presentation	Examples
G	era	text	AD
C	century of era (>=0)	number	20
Y	year of era (>=0)	year	1996
x	weekyear	year	1996
w	week of weekyear	number	27
e	day of week	number	2
E	day of week	text	Tuesday; Tue
y	year	year	1996
D	day of year	number	189
M	month of year	month	July; Jul; 07
d	day of month	number	10
a	halfday of day	text	PM
K	hour of halfday (0~11)	number	0
h	clockhour of halfday (1~12)	number	12
H	hour of day (0~23)	number	0
k	clockhour of day (1~24)	number	24
m	minute of hour	number	30
s	second of minute	number	55
S	fraction of second	number	978
z	time zone	text	Pacific Standard Time; PST
Z	time zone offset/id	zone	-0800; -08:00; America/Los_Angeles
	escape for text	delimiter	
'	single quote	literal	

TO_CHAR(<field>, <format>)

TO_CHAR(<field>, <format>)

```
SELECT JobTitle,
TO_CHAR( AVG( TO_NUMBER( AnnualSalary, '¤' )), '¤#,###.00' ) AS avg_salary,
COUNT( DISTINCT name ) AS number
FROM dfs.drillclass.`baltimore_salaries_2016.csvh`
GROUP BY JobTitle
Order By avg_salary DESC
```



Intervals

```
SELECT date2,
date5,
(TO_DATE( date2, 'MM/dd/yyyy' ) - TO_DATE( date5, 'yyyy-MM-dd' )) as date_diff
FROM dfs.drillclass.`dates.csvh`
```

date_diff
P249D
P-5D
P-312D
P-315D
P-171D

Intervals

P249D

- P (Period) marks the beginning of a period of time.
- Y follows a number of years.
- M follows a number of months.
- D follows a number of days.
- H follows a number of hours 0-24.
- M follows a number of minutes.
- S follows a number of seconds and optional milliseconds

Intervals

```
SELECT date2,
date5,
 (TO_DATE( date2, 'MM/dd/yyyy' ) - TO_DATE( date5, 'yyyy-MM-dd
)) as date_diff,
EXTRACT( day FROM (TO_DATE( date2, 'MM/dd/yyyy' ) -
TO_DATE( date5, 'yyyy-MM-dd' )))
FROM dfs.drillclass.`dates.csvh`
```

date_diff	EXPR$3
P249D	249
P-5D	-5
P-312D	-312
P-315D	-315

Other Date/Time Functions

- AGE(timestamp):
- EXTRACT(field FROM time_exp): Extract a part of a date, time or interval
- CURRENT_DATE()/CURRENT_TIME()/NOW()
- DATE_ADD()/DATE_SUB(): Adds or subtracts two dates

For complete documentation: http://drill.apache.org/docs/date-time-functions-and-arithmetic/

In Class Exercise:
Parsing Dates and Times

In this exercise you will find a data file called dates.csvh which contains 5 columns of random dates in various formats:

- **date1** is in ISO 8601 format
- **date2** is MM/DD/YYYY ie: 03/12/2016
- **date3** is: Sep 19, 2016
- **date4** is formatted: Sun, 19 Mar 2017 00:15:28 -0700
- **date5** is formatted like database dates: YYYY-mm-dd: 2016-10-03

For this exercise, complete the following steps:

1. Using the various methods, (**CAST(), TO_DATE()**) we have discussed, convert each column into a date (or time) as appropriate.
2. Reformat **date5** so that it is in the same format as **date3**.
3. Find all the dates rows where **date3** occurs after **date5**.
4. Create a histogram table of **date2** by weekday: IE: Sunday 5, Monday 4, etc
5. Find all the entries in **date5** that are **more than 1 year old**


```
SELECT CAST( date1 AS DATE )
FROM dfs.drillclass.`dates.csvh`

SELECT date2, TO_DATE( date2, 'MM/dd/yyyy' )
FROM dfs.drillclass.`dates.csvh`

SELECT date3, TO_DATE( date3, 'MMM dd, yyyy' )
FROM dfs.drillclass.`dates.csvh`

SELECT date4, TO_TIMESTAMP( date4, 'EEE, dd MMM yyyy HH:mm:ss
Z' )
FROM dfs.drillclass.`dates.csvh`

SELECT date5, TO_DATE( date5, 'yyyy-MM-dd' )
FROM dfs.drillclass.`dates.csvh`
```

Nested Data

Complex data types: A data type which holds more than one value

Complex data types: A data type which holds more than one value

- **Array**: A complex data type **indexed by number**
- **Map**: A complex data type **indexed by a key**

Arrays in Drill

columns
["Robert","Hernandez","5/3/67"]
["Steve","Smith","8/4/84"]
["Anne","Raps","9/13/91"]
["Alice","Muller","4/15/75"]

Arrays in Drill

columns
["Robert","Hernandez","5/3/67"]
["Steve","Smith","8/4/84"]
["Anne","Raps","9/13/91"]
["Alice","Muller","4/15/75"]

```
SELECT columns[0] AS first_name,
columns[1] AS last_name,
columns[2] AS birthday
FROM dfs.drillclass.`customer_data.csv`
```

Arrays in Drill

```
SELECT columns[0] AS first_name,
columns[1] AS last_name,
columns[2] AS birthday
FROM dfs.drillclass.`customer_data.csv`
```

first_name	last_name	birthday
Robert	Hernandez	5/3/67
Steve	Smith	8/4/84
Anne	Raps	9/13/91
Alice	Muller	4/15/75

Showing 1 to 4 of 4 entries

Maps (Key/Value Pairs) in Drill

```
SELECT parse_user_agent( columns[0] ) AS ua
FROM dfs.drillclass.`user-agents.csv`
```

Documentation for this function is available at: https://github.com/cgivre/drill-useragent-function

Maps (Key/Value Pairs) in Drill

```
SELECT parse_user_agent( columns[0] ) AS ua
FROM dfs.drillclass.`user-agents.csv`
```

ua
{"DeviceClass":"Desktop","DeviceName":"Desktop","DeviceBrand":"Unknown","OperatingSystemClass":"Desktop","OperatingSystemName":"Windows NT","OperatingSystemVersion":"Windows XP","OperatingSystemNameVersion":"Windows XP","LayoutEngineClass":"Browser","LayoutEngineName":"Gecko","LayoutEngineVersion":"35.0","LayoutEngineVersionMajor":"35","LayoutEngineNameVersion":"Gecko 35.0","LayoutEngineNameVersionMajor":"Gecko 35","LayoutEngineBuild":"20100101","AgentClass":"Browser","AgentName":"Firefox","AgentVersion":"35.0","AgentVersionMajor":"35","AgentNameVersion":"Firefox 35.0","AgentNameVersionMajor":"Firefox 35"}
{"DeviceClass":"Desktop","DeviceName":"Desktop","DeviceBrand":"Unknown","OperatingSystemClass":"Desktop","OperatingSystemName":"Windows NT","OperatingSystemVersion":"Windows XP","OperatingSystemNameVersion":"Windows XP","LayoutEngineClass":"Browser","LayoutEngineName":"Gecko","LayoutEngineVersion":"35.0","LayoutEngineVersionMajor":"35","LayoutEngineNameVersion":"Gecko 35.0","LayoutEngineNameVersionMajor":"Gecko 35","LayoutEngineBuild":"20100101","AgentClass":"Browser","AgentName":"Firefox","AgentVersion":"35.0","AgentVersionMajor":"35","AgentNameVersion":"Firefox 35.0","AgentNameVersionMajor":"Firefox 35"}

Maps (Key/Value Pairs) in Drill

```
SELECT parse_user_agent( columns[0] ) AS ua
FROM dfs.drillclass.`user-agents.csv`
```

```
{
  "DeviceClass":"Desktop",
  "DeviceName":"Macintosh",
  "DeviceBrand":"Apple",
  "OperatingSystemClass":"Desktop",
  "OperatingSystemName":"Mac OS X",
  …
   "AgentName":"Chrome",
  "AgentVersion":"39.0.2171.99",
  "AgentVersionMajor":"39",
  "AgentNameVersion":"Chrome 39.0.2171.99",
  "AgentNameVersionMajor":"Chrome 39",
  "DeviceCpu":"Intel"
}
```

table.map.key

Maps (Key/Value Pairs) in Drill

```
SELECT uadata.ua.OperatingSystemName AS OS_Name
FROM (
    SELECT parse_user_agent( columns[0] ) AS ua
    FROM dfs.drillclass.`user-agents.csv`
) AS uadata
```

In Class Exercise:

The file user-agents.csv is a small sample of a list of user agents gathered from a server log during an attempted attack. Using this data, answer the following questions:

1. What was the most common OS?
2. What was the most common browser?

In Class Exercise:

The file user-agents.csv is a small sample of a list of user agents gathered from a server log during an attempted attack. Using this data, answer the following questions:

1. What was the most common OS?
2. What was the most common browser?

```
SELECT uadata.ua.AgentNameVersion AS Browser,
COUNT( * ) AS BrowserCount
FROM (
   SELECT parse_user_agent( columns[0] ) AS ua
   FROM dfs.drillclass.`user-agents.csv`
) AS uadata
GROUP BY uadata.ua.AgentNameVersion
ORDER BY BrowserCount DESC
```

Querying JSON Data

records.json

```
[
    {
        "first_name":"Robert",
        "last_name":"Hernandez",
        "birthday":"5\\/3\\/67"
    },{
        "first_name":"Steve",
        "last_name":"Smith",
        "birthday":"8\\/4\\/84"},
    }
]
```


```
SELECT *
FROM dfs.drillclass.`records.json`
```

first_name	last_name	birthday
Robert	Hernandez	5V3V67
Steve	Smith	8V4V84
Anne	Raps	9V13V91
Alice	Muller	4V15V75

Showing 1 to 4 of 4 entries

Previous 1 Next

GTK CYBER

GET ON THE FAST TRACK

gtkcyber.com

Please open **split.json** in a text editor


```
{
"columns":
    [
     "first_name",
      "last_name",
      "birthday"
    ],
"data":
[
 [
  "Robert",
  "Hernandez",
  "5\\/3\\/67"
 ],
 [
  "Steve",
  "Smith",
  "8\\/4\\/84"
  ],
 ]
}
```

split.json

`FLATTEN( <json array> )`
separates elements in a repeated field into individual records.


```
SELECT data
FROM dfs.drillclass.`split.json`
```

data
[["Robert","Hernandez","5\\V3\\V67"],["Steve","Smith","8\\V4\\V84"],["Anne","Raps","9\\V13\\V91"],["Alice","Muller","4\\V15\\V75"]]

Showing 1 to 1 of 1 entries

Previous 1 Next


```
SELECT FLATTEN(data) AS row_data
FROM dfs.drillclass.`split.json`
```

row_data
["Robert","Hernandez","5\\V3\\V67"]
["Steve","Smith","8\\V4\\V84"]
["Anne","Raps","9\\V13\\V91"]
["Alice","Muller","4\\V15\\V75"]

Showing 1 to 4 of 4 entries

Previous 1 Next


```
SELECT row_data[0] AS first_name,
row_data[1] AS last_name,
row_data[2] AS birthday
FROM
(
  SELECT FLATTEN( data ) AS row_data
  FROM dfs.drillclass.`split.json`
) AS split_data
```

first_name	last_name	birthday
Robert	Hernandez	5\/3\/67
Steve	Smith	8\/4\/84
Anne	Raps	9\/13\/91
Alice	Muller	4\/15\/75

Showing 1 to 4 of 4 entries

Previous 1 Next

Please open **columns.json** in a text editor


```
{
 "first_name":
  {
   "0":"Robert",
   "1":"Steve",
   "2":"Anne",
   "3":"Alice"
  },
 “last_name": {
   "0":"Hernandez",
   "1":"Smith",
   "2":"Raps",
   “3”:"Muller"
   },
"birthday":{
   "0":"5\\/3\\/67",
   "1":"8\\/4\\/84",
   "2":"9\\/13\\/91",
   "3":"4\\/15\\/75"
   }
}
```

`KVGEN(<map>)`

generates key/value pairs from a column with repeated data. Often used in combination with **`FLATTEN()`**.


```
SELECT KVGEN( first_name ) AS kvgen_firstname
FROM dfs.drillclass.`columns.json`
```

kvgen_firstname
[{"key":"0","value":"Robert"},{"key":"1","value":"Steve"},{"key":"2","value":"Anne"},{"key":"3","value":"Alice"}]

Showing 1 to 1 of 1 entries

Previous 1 Next


```
SELECT FLATTEN(
  KVGEN( first_name )
) AS kvgen_firstname
FROM dfs.drillclass.`columns.json`
```


```
SELECT FLATTEN(
  KVGEN( first_name )
) AS kvgen_firstname
FROM dfs.drillclass.`columns.json`
```

kvgen_firstname
{"key":"0","value":"Robert"}
{"key":"1","value":"Steve"}
{"key":"2","value":"Anne"}
{"key":"3","value":"Alice"}

Showing 1 to 4 of 4 entries Previous 1 Next


```
SELECT FLATTEN( KVGEN( first_name ) )['value'] AS firstname
FROM dfs.drillclass.`columns.json`
```

firstname
Robert
Steve
Anne
Alice

Showing 1 to 4 of 4 entries

Previous 1 Next


```
SELECT first_name, last_name, birthday
FROM
(
  SELECT row_number() OVER (ORDER BY `1') AS rownum,
  FLATTEN( KVGEN(first_name))['value'] AS first_name
  FROM dfs.drillclass.`columns.json`
) AS tbl1
JOIN
(
  SELECT row_number() OVER (ORDER BY `1') AS rownum,
  FLATTEN( KVGEN(last_name))['value'] AS last_name
  FROM dfs.drillclass.`columns.json`
) AS tbl2
ON tbl1.rownum=tbl2.rownum
JOIN
(
  SELECT row_number() OVER (ORDER BY `1') AS rownum,
  FLATTEN( KVGEN(birthday))['value'] AS birthday
  FROM dfs.drillclass.`columns.json`
) AS tbl3 ON tbl1.rownum=tbl3.rownum
```

Putting it all together…

Please run

```
ALTER SYSTEM SET `store.json.all_text_mode` = true;
```

in the Drill command line

Please open
`baltimore_salaries_2016.json`
in a text editor


```
{
  "meta" : {
    "view" : {
      "id" : "nsfe-bg53",
      "name" : "Baltimore City Employee Salaries FY2015",
      "attribution" : "Mayor's Office",
      "averageRating" : 0,
      "category" : "City Government",
      …
      "       "format" : { }
      },
},
  "data" : [ [ 1, "66020CF9-8449-4464-AE61-B2292C7A0F2D", 1, 1438255843, "393202",
1438255843, "393202", null, "Aaron,Patricia G", "Facilities/Office Services II",
"A03031", "OED-Employment Dev (031)", "1979-10-24T00:00:00", "55314.00", "53626.04" ]
, [ 2, "31C7A2FE-60E6-4219-890B-AFF01C09EC65", 2, 1438255843, "393202", 1438255843,
"393202", null, "Aaron,Petra L", "ASSISTANT STATE'S ATTORNEY", "A29045", "States
Attorneys Office (045)", "2006-09-25T00:00:00", "74000.00", "73000.08" ]
```


```
"meta" : {
      "view" : {
        "id" : "nsfe-bg53",
        "name" : "Baltimore City Employee Salaries FY2015",
        "attribution" : "Mayor's Office",
        "averageRating" : 0,
        "category" : "City Government",
        …
        "       "format" : { }
        },
},
  "data" : [ [ 1, "66020CF9-8449-4464-AE61-B2292C7A0F2D", 1, 1438255843, "393202",
1438255843, "393202", null, "Aaron,Patricia G", "Facilities/Office Services II",
"A03031", "OED-Employment Dev (031)", "1979-10-24T00:00:00", "55314.00", "53626.04" ]
, [ 2, "31C7A2FE-60E6-4219-890B-AFF01C09EC65", 2, 1438255843, "393202", 1438255843,
"393202", null, "Aaron,Petra L", "ASSISTANT STATE'S ATTORNEY", "A29045", "States
Attorneys Office (045)", "2006-09-25T00:00:00", "74000.00", "73000.08" ]
```


```
    "meta" : {
      "view" : {
        "id" : "nsfe-bg53",
        "name" : "Baltimore City Employee Salaries FY2015",
        "attribution" : "Mayor's Office",
        "averageRating" : 0,
        "category" : "City Government",
        …
        "       "format" : { }
        },
},
  "data" : [ [ 1, "66020CF9-8449-4464-AE61-B2292C7A0F2D", 1, 1438255843, "393202",
1438255843, "393202", null, "Aaron,Patricia G", "Facilities/Office Services II",
"A03031", "OED-Employment Dev (031)", "1979-10-24T00:00:00", "55314.00", "53626.04" ]
, [ 2, "31C7A2FE-60E6-4219-890B-AFF01C09EC65", 2, 1438255843,
"393202", 1438255843, "393202", null, "Aaron,Petra L",
"ASSISTANT STATE'S ATTORNEY", "A29045", "States Attorneys
Office (045)", "2006-09-25T00:00:00", "74000.00", "73000.08" ]
```


```
"data" : [
    [ 1,
    "66020CF9-8449-4464-AE61-B2292C7A0F2D",
    1,
    1438255843,
    "393202",
    1438255843,
    “393202",
     null,
    "Aaron,Patricia G",
    "Facilities/Office Services II",
    "A03031",
    "OED-Employment Dev (031)",
    "1979-10-24T00:00:00",
    “55314.00",
     “53626.04"
     ]
```

In Class Exercise

Using the Baltimore Salaries JSON file, recreate the earlier query to find the average salary by job title and how many people have each job title.

HINT: Don't forget to CAST() the columns...

HINT 2: GROUP BY does NOT support aliases.

In Class Exercise

Using the JSON file, recreate the earlier query to find the average salary by job title and how many people have each job title.

```
SELECT raw_data[9] AS job_title,
AVG( CAST( raw_data[13] AS DOUBLE ) ) AS avg_salary,
COUNT( DISTINCT raw_data[8] ) AS person_count
FROM
(
   SELECT FLATTEN( data ) AS raw_data
   FROM dfs.drillclass.`baltimore_salaries_2016.json`
)
GROUP BY raw_data[9]
ORDER BY avg_salary DESC
```

HTTPD Log Files

HTTPD Log Files

195.154.46.135 - - [25/Oct/2015:04:11:25 +0100] "GET /linux/doing-pxe-without-dhcp-control HTTP/1.1" 200 24323 "http://howto.basjes.nl/" "Mozilla/5.0 (Windows NT 5.1; rv:35.0) Gecko/20100101 Firefox/35.0"
23.95.237.180 - - [25/Oct/2015:04:11:26 +0100] "GET /join_form HTTP/1.0" 200 11114 "http://howto.basjes.nl/" "Mozilla/5.0 (Windows NT 5.1; rv:35.0) Gecko/20100101 Firefox/35.0"
23.95.237.180 - - [25/Oct/2015:04:11:27 +0100] "POST /join_form HTTP/1.1" 302 9093 "http://howto.basjes.nl/join_form" "Mozilla/5.0 (Windows NT 5.1; rv:35.0) Gecko/20100101 Firefox/35.0"
158.222.5.157 - - [25/Oct/2015:04:24:31 +0100] "GET /join_form HTTP/1.0" 200 11114 "http://howto.basjes.nl/" "Mozilla/5.0 (Windows NT 6.3; WOW64; rv:34.0) Gecko/20100101 Firefox/34.0 AlexaToolbar/alxf-2.21"
158.222.5.157 - - [25/Oct/2015:04:24:32 +0100] "POST /join_form HTTP/1.1" 302 9093 "http://howto.basjes.nl/join_form" "Mozilla/5.0 (Windows NT 6.3; WOW64; rv:34.0) Gecko/20100101 Firefox/34.0 AlexaToolbar/alxf-2.21"

For complete documentation: https://gist.github.com/cgivre/47f07a06d44df2af625fc6848407ae7c

HTTPD Log Files

195.154.46.135 - - [25/Oct/2015:04:11:25 +0100] "GET /linux/doing-pxe-without-dhcp-control HTTP/1.1" 200 24323 "http://howto.basjes.nl/" "Mozilla/5.0 (Windows NT 5.1; rv:35.0) Gecko/20100101 Firefox/35.0"
23.95.237.180 - - [25/Oct/2015:04:11:26 +0100] "GET /join_form HTTP/1.0" 200 11114 "http://howto.basjes.nl/" "Mozilla/5.0 (Windows NT 5.1; rv:35.0) Gecko/20100101 Firefox/35.0"
23.95.237.180 - - [25/Oct/2015:04:11:27 +0100] "POST /join_form HTTP/1.1" 302 9093 "http://howto.basjes.nl/join_form" "Mozilla/5.0 (Windows NT 5.1; rv:35.0) Gecko/20100101 Firefox/35.0"
158.222.5.157 - - [25/Oct/2015:04:24:31 +0100] "GET /join_form HTTP/1.0" 200 11114 "http://howto.basjes.nl/" "Mozilla/5.0 (Windows NT 6.3; WOW64; rv:34.0) Gecko/20100101 Firefox/34.0 AlexaToolbar/alxf-2.21"
158.222.5.157 - - [25/Oct/2015:04:24:32 +0100] "POST /join_form HTTP/1.1" 302 9093 "http://howto.basjes.nl/join_form" "Mozilla/5.0 (Windows NT 6.3; WOW64; rv:34.0) Gecko/20100101 Firefox/34.0 AlexaToolbar/alxf-2.21"

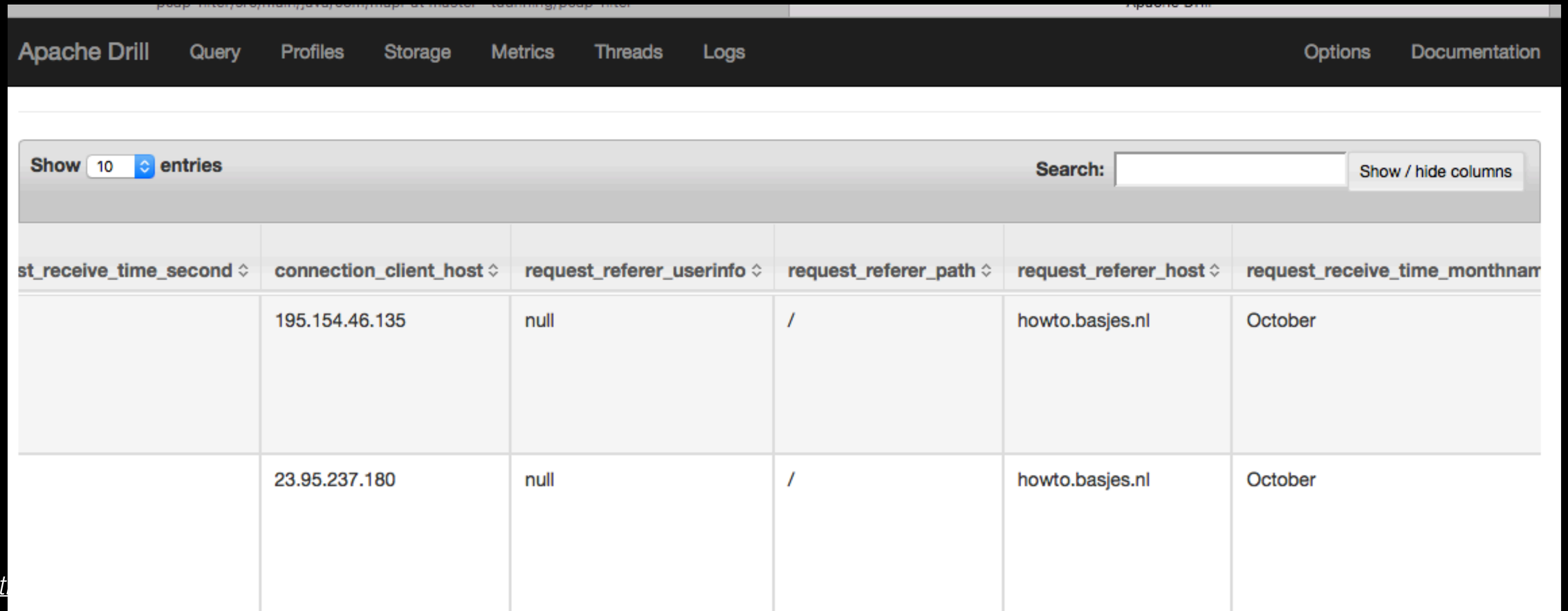
```
"httpd": {
      "type": "httpd",
      "logFormat": "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-agent}i\"",
      "timestampFormat": null
    },
```

HTTPD Log Files

```
SELECT *
FROM dfs.drillclass.`small-server-log.httpd`
```

HTTPD Log Files

```
SELECT *
FROM dfs.drillclass.`small-server-log.httpd`
```



HTTPD Log Files

```
SELECT request_referer, parse_url( request_referer ) AS url_data
FROM dfs.drillclass.`small-server-log.httpd`
```

HTTPD Log Files

```
SELECT request_referer, parse_url( request_referer ) AS url_data
FROM dfs.drillclass.`small-server-log.httpd`
```

Show 10 entries Search:

request_referer	url_data
http://howto.basjes.nl/	{"protocol":"http","authority":"howto.basjes.nl","host":"howto.basjes.nl","path":"/"}
http://howto.basjes.nl/	{"protocol":"http","authority":"howto.basjes.nl","host":"howto.basjes.nl","path":"/"}
http://howto.basjes.nl/join_form	{"protocol":"http","authority":"howto.basjes.nl","host":"howto.basjes.nl","path":"/join_form"}
http://howto.basjes.nl/	{"protocol":"http","authority":"howto.basjes.nl","host":"howto.basjes.nl","path":"/"}
http://howto.basjes.nl/join_form	{"protocol":"http","authority":"howto.basjes.nl","host":"howto.basjes.nl","path":"/join_form"}

http%3A%2F%2Fmysite.com%3Fuser%3Dcgivre%26password%3D1234%26firstname%3DCharles+S


```
SELECT urldecode( <field> )
FROM…
```

http://mysite.com?
user=cgivre&password=1234&first name=Charles S

SELECT **parse_query(urldecode(<field>))**
FROM…

```
{
  “username”:”Charles”,
  “password”:”1234”,
   “name”:”Charles S”
}
```

Networking Functions

Networking Functions

- inet_aton(<ip>): Converts an IPv4 Address to an integer
- inet_ntoa(<int>): Converts an integer to an IPv4 address
- is_private(<ip>): Returns true if the IP is private
- in_network(<ip>,<cidr>): Returns true if the IP is in the CIDR block
- getAddressCount(<cidr>): Returns the number of IPs in a CIDR block
- getBroadcastAddress(<cidr>): Returns the broadcast address of a CIDR block
- getNetmast(<cidr>): Returns the net mask of a CIDR block
- getLowAddress(<cidr>): Returns the low IP of a CIDR block
- getHighAddress(<cidr>): Returns the high IP of a CIDR block
- parse_user_agent(<ua_string>): Returns a map of user agent information
- urlencode(<url>): Returns a URL encoded string
- urldecode(<url>): Decodes a URL encoded string

In Class Exercise

There is a file in the repo called ‘hackers-access.httpd’ is a HTTPD server log. Write queries to determine:

1. What is the most common browser?

2. What is the most common operating system?

In Class Exercise

```
SELECT ua.uadata.OperatingSystemNameVersion AS operating_system,
COUNT( * ) AS os_count
FROM
(
  SELECT parse_user_agent(`request_user-agent`) AS uadata
  FROM dfs.drillclass.`hackers-access.httpd`
) AS ua
GROUP BY ua.uadata.OperatingSystemNameVersion
ORDER BY os_count DESC
```

A Quick Demo…

What if you wanted all the **unique** IP addresses in your server log?

A Quick Demo…

```
SELECT DISTINCT connection_client_host
FROM dfs.drillclass.`hackers-access.httpd`
```

A Quick Demo…

```
SELECT DISTINCT connection_client_host
FROM dfs.drillclass.`hackers-access.httpd`
```

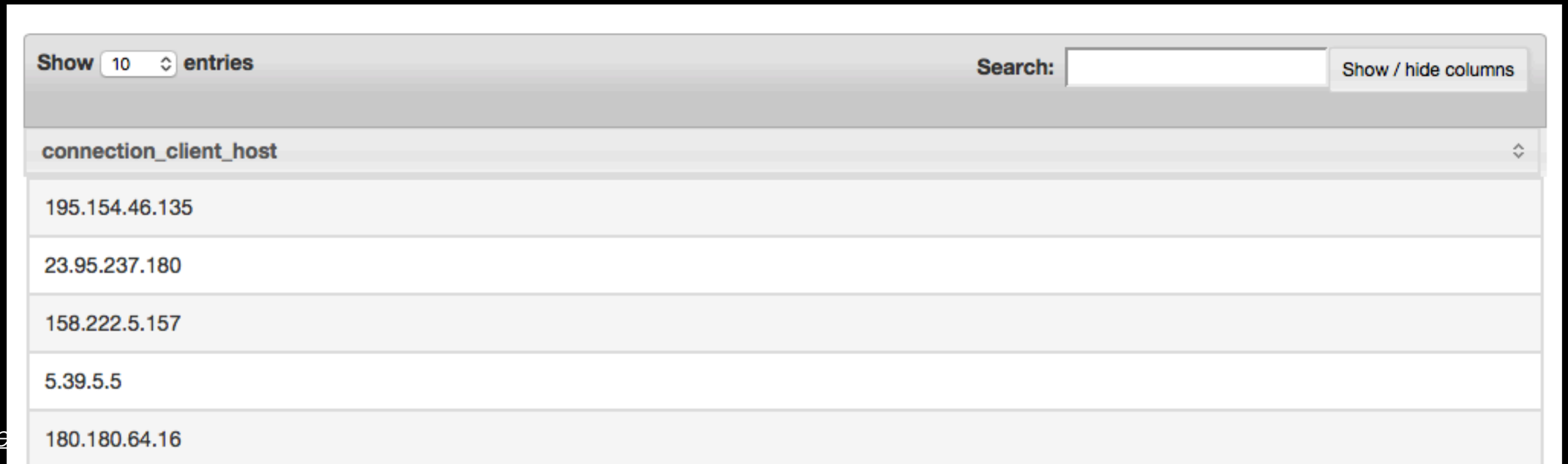
A Quick Demo…

```
SELECT DISTINCT connection_client_host
FROM dfs.drillclass.`hackers-access.httpd`
```

Now what if you wanted these IPs in order?

A Quick Demo…

```
SELECT DISTINCT connection_client_host
FROM dfs.drillclass.`hackers-access.httpd`
WHERE regexp_matches(`connection_client_host`, '(\d{1,3}\.){3}\d{1,3}')
```



A Quick Demo…

```
SELECT DISTINCT connection_client_host
FROM dfs.drillclass.`hackers-access.httpd`
WHERE regexp_matches(`connection_client_host`, '(\d{1,3}\.){3}\d{1,3}')
```

connection_client_host
1.0.189.90
1.0.190.144
1.0.190.64
1.0.191.52
101.231.46.34
101.71.27.120
103.27.239.39

What if we only wanted IPs within a certain range?

A Quick Demo…

```
SELECT DISTINCT connection_client_host
FROM dfs.drillclass.`hackers-access.httpd`
WHERE regexp_matches(`connection_client_host`, '(\d{1,3}\.)
{3}\d{1,3}') AND
inet_aton( `connection_client_host` ) >= inet_aton( '23.94.10.8' )
AND
inet_aton( `connection_client_host` ) < inet_aton( '31.187.79.31' )
ORDER BY inet_aton( `connection_client_host` ) ASC
```

What if we wanted to know what were the locations of IPs who requested certain pages?

A Quick Demo…

```
SELECT getCountryName( connection_client_host ) AS ip_country,
COUNT( DISTINCT connection_client_host ) AS unique_ips
FROM dfs.drillclass.`hackers-access.httpd`
WHERE regexp_matches(`connection_client_host`, '(\d{1,3}\.){3}\d{1,3}')
GROUP BY getCountryName( connection_client_host )
ORDER BY unique_ips DESC
```

Show 10 entries Search: Show / hide columns

ip_country	unique_ips
United States	299
Thailand	46
China	36
Romania	21
Germany	18

Log Files

Log Files

- Drill does not natively support reading log files… yet
- If you are NOT using Merlin, included in the GitHub repo are several .jar files. Please take a second and copy them to <drill directory>/jars/3rdparty

Log Files

```
070823 21:00:32       1 Connect     root@localhost on test1
070823 21:00:48       1 Query       show tables
070823 21:00:56       1 Query       select * from category
070917 16:29:01      21 Query       select * from location
070917 16:29:12      21 Query       select * from location where id = 1 LIMIT 1
```


```
"log": {
      "type": “log",
      "errorOnMismatch": false,
      "extensions": [
        "log"
      ],
      "fieldNames": [
        "date",
        "time",
        "pid",
        "action",
        "query"
      ],
      "pattern": "(\\d{6})\\s(\\d{2}:\\d{2}:\\d{2})\\s+(\\d+)\\s(\\w+)\\s+(.+)"
    }
  }
```


```
SELECT *
FROM dfs.drillclass.`mysql.log`
```


```
SELECT *
FROM dfs.drillclass.`mysql.log`
```

Show 10 entries Search: Show / hide columns

date	time	pid	action	query
070823	21:00:32	1	Connect	root@localhost on test1
070823	21:00:48	1	Query	show tables
070823	21:00:56	1	Query	select * from category
070917	16:29:01	21	Query	select * from location
070917	16:29:12	21	Query	select * from location where id = 1 LIMIT 1

Showing 1 to 5 of 5 entries Previous 1 Next

In Class Exercise

There is a file in the repo called ‘firewall.log’ which contains entries in the following format:

```
Dec 12 03:36:23    sshd[41875]: Failed password for root from 222.189.239.10 port 1350 ssh2
Dec 12 03:36:22    sshd[41875]: Failed password for root from 222.189.239.10 port 1350 ssh2
Dec 12 03:36:22    sshlockout[15383]: Locking out 222.189.239.10 after 15 invalid attempts
Dec 12 03:36:22    sshd[41875]: Failed password for root from 222.189.239.10 port 1350 ssh2
Dec 12 03:36:22    sshlockout[15383]: Locking out 222.189.239.10 after 15 invalid attempts
Dec 12 03:36:22    sshd[42419]: Failed password for root from 222.189.239.10 port 2646 ssh2
```

In this exercise:

1. Write a regex to extract the date, process type, PID, from IP and any other information you believe may be useful from this log

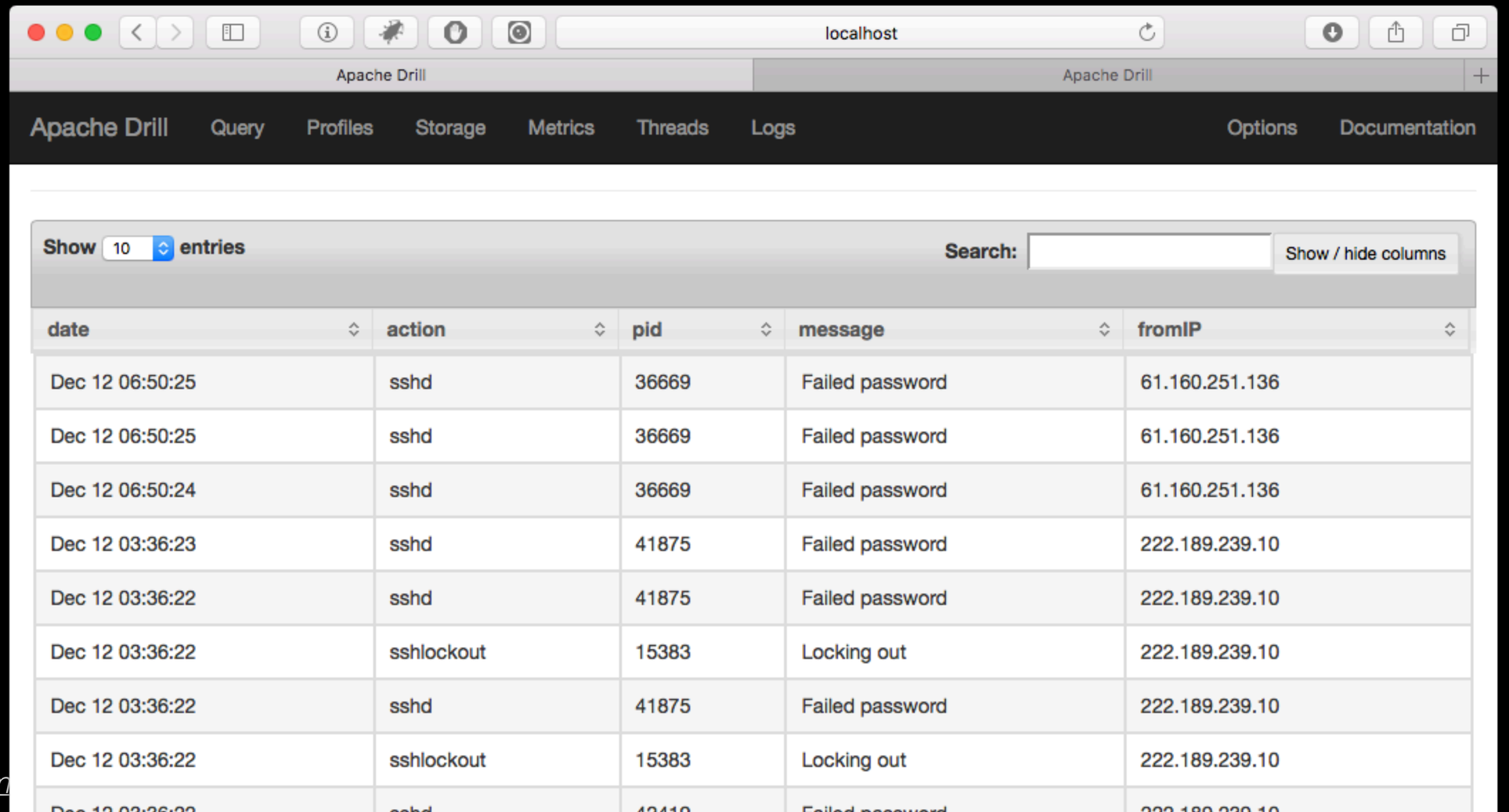
2. Use that regex to configure Drill to query this data.

3. Find all the records where the IP is in the CIDR block: 61.160.251.128/28

In Class Exercise

```
"ssdlog": {
      "type": "log",
      "extensions": [
        "ssdlog"
      ],
      "fieldNames": [
        "date",
        "action",
        "pid",
        "message",
        "fromIP"
      ],
      "pattern": "(\\w{3}\\s\\d{2}\\s\\d{2}:\\d{2}:\\d{2})\\s+(\\w+)\\[(\\d+)\\]:\\s(\\w+\\s\\w+).+?(\\d{1,3}\\.\\d{1,3}\\.\\d{1,3}\\.\\d{1,3})"
    }
```

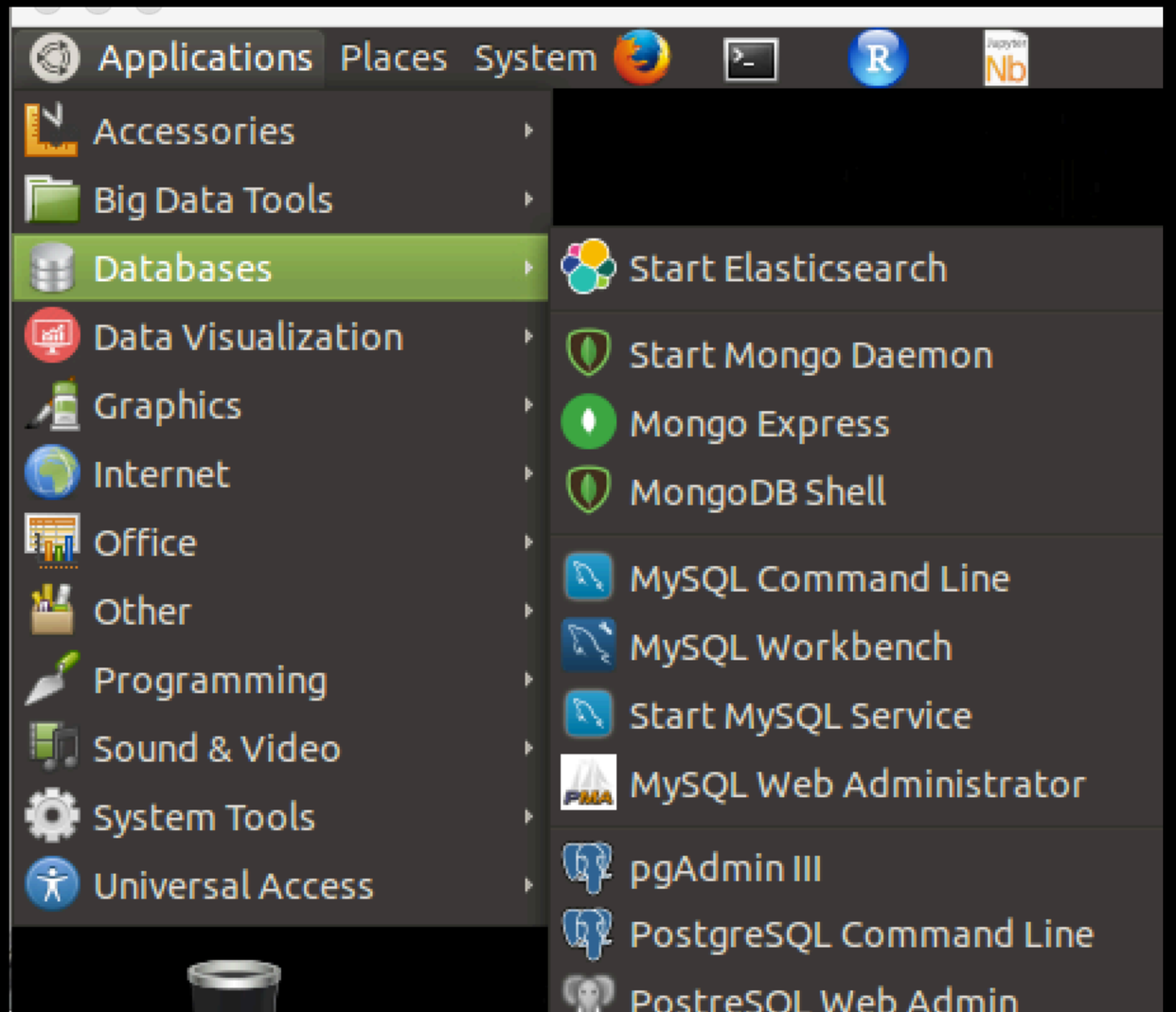
In Class Exercise



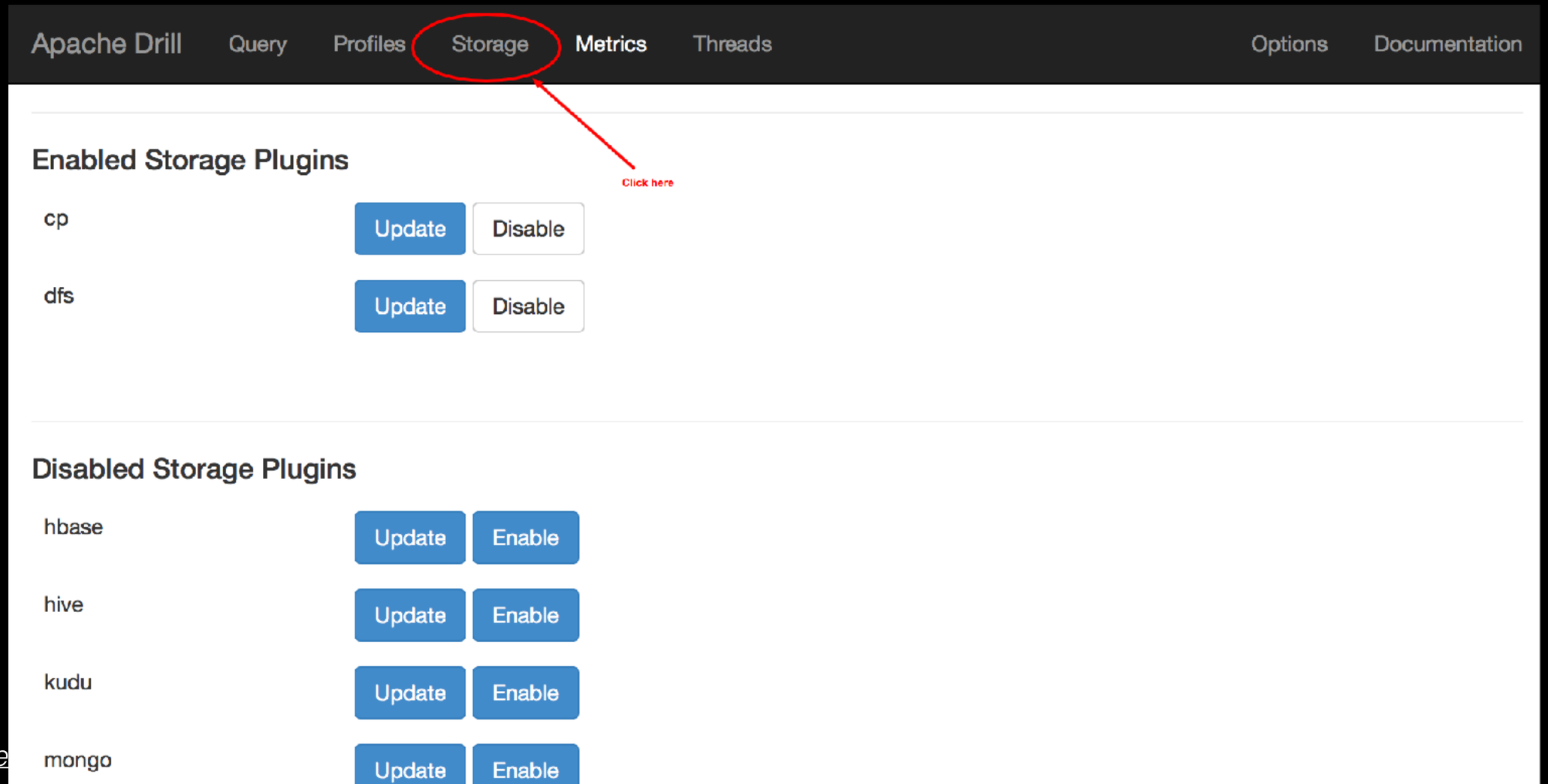
Connecting other Data Sources

Connecting other Data Sources

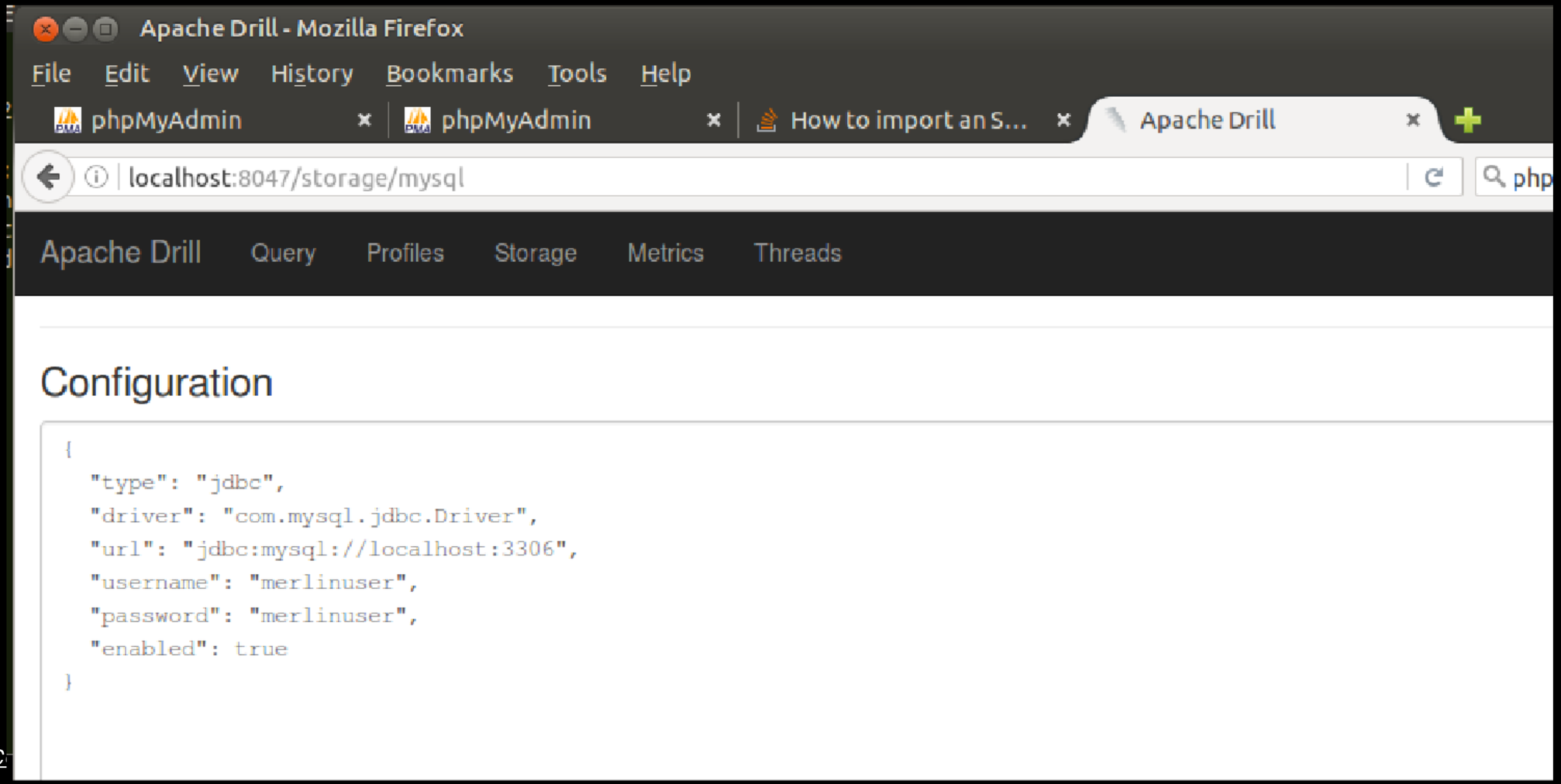
Connecting other Data Sources



Connecting other Data Sources



Connecting other Data Sources

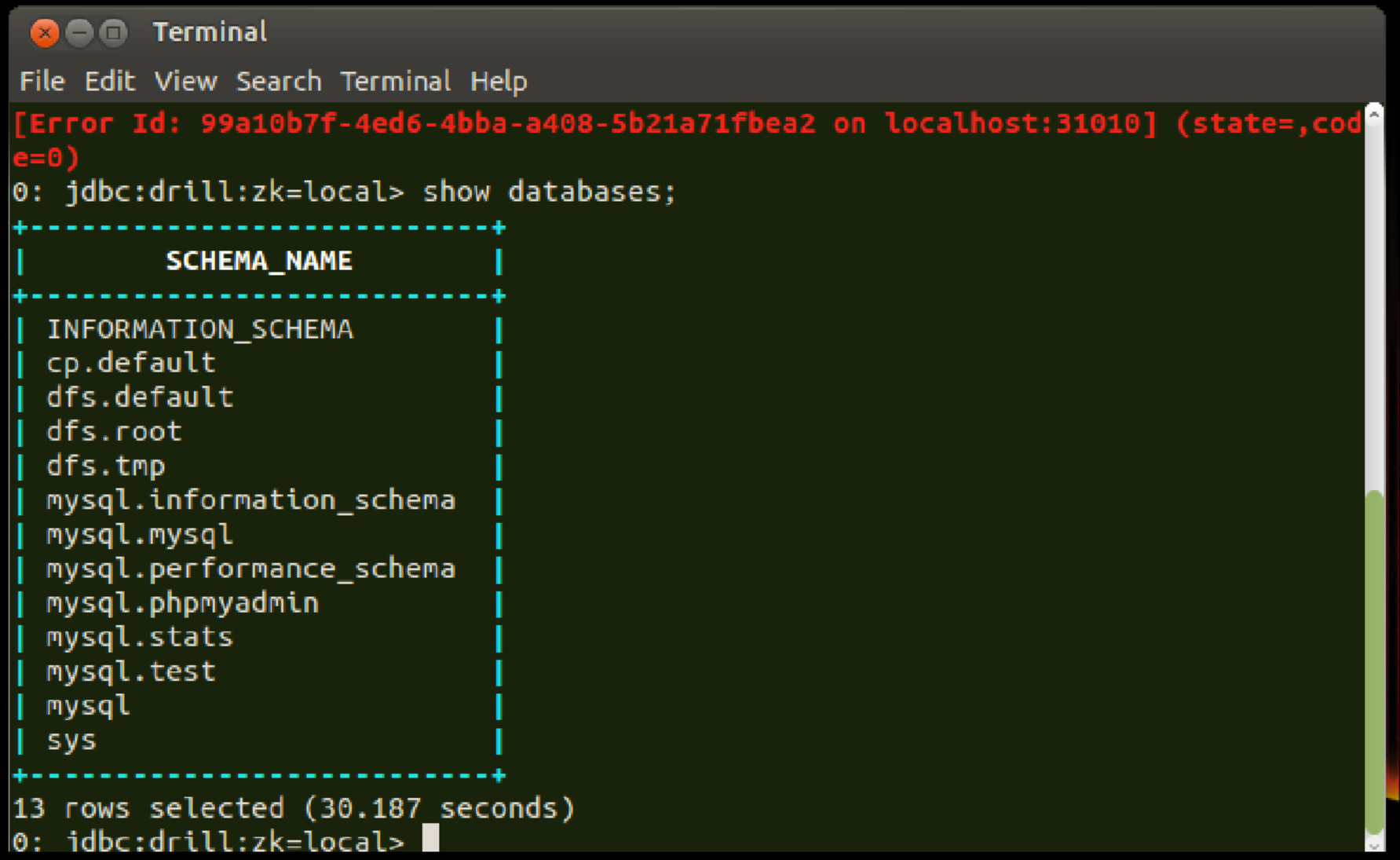


MySQL™

Connecting other Data Sources

```
{
 “type”:”jdbc”,
 “driver”:”com.mysql.jdbc.Driver”,
  “url”:”jdbc:mysql://localhost:3306",
  “username”:”merlinuser”,
  “password”:”merlinuser”,
  “enabled”:true
 }
```

Connecting other Data Sources



Connecting other Data Sources

```
SELECT teams.name, SUM( batting.HR ) as hr_total
FROM batting
INNER JOIN teams ON batting.teamID=teams.teamID
WHERE batting.yearID = 1988 AND teams.yearID = 1988
GROUP BY batting.teamID
ORDER BY hr_total DESC
```

Connecting other Data Sources

```
SELECT teams.name, SUM( batting.HR ) as hr_total
FROM batting
INNER JOIN teams ON batting.teamID=teams.teamID
WHERE batting.yearID = 1988 AND teams.yearID = 1988
GROUP BY batting.teamID
ORDER BY hr_total DESC
```

MySQL: 0.047 seconds

Connecting other Data Sources

```
SELECT teams.name, SUM( batting.HR ) as hr_total
FROM mysql.stats.batting
INNER JOIN mysql.stats.teams ON batting.teamID=teams.teamID
WHERE batting.yearID = 1988 AND teams.yearID = 1988
GROUP BY teams.name
ORDER BY hr_total DESC
```

MySQL: 0.047 seconds

Drill: 0.366 seconds


hadoop

Just like DFS, except you specify a link to the Hadoop namenode.

```
{
  "type": "file",
  "enabled": true,
  "connection": "hdfs://localhost:54310",
  "config": null,
  "workspaces": {
      "demodata": {
      "location": "/user/merlinuser/demo",
      "writable": true,
      "defaultInputFormat": null
    }
  },
```


```
SELECT name, SUM( CAST( HR AS INT)) AS HR_Total
FROM hdfs.demodata.`Teams.csvh`
WHERE yearID=1988
GROUP BY name
ORDER BY HR_Total DESC
```

Writing A Drill Function

Writing A Drill Function

```
function doSomething( arg1, arg2 )
{
    //Something happens…
    return result value
}
```

Writing A Drill Function

Step 1. In the workshop/udfs folder there is a .zip file called drill-workshop-function.zip, unzip that file.

Step 2. Open this folder using IntelliJ IDEA which can be found in the programming menu

Writing A Drill Function

blankFunction.java - drill-workshop-function - [~/OneDrive/github/drill-workshop-function]

drill-workshop-function › src › main › java › org › apache › drill › contrib › function › blankFunction

Project

drill-workshop-function ~/OneDrive/github/d
- .idea
- src
  - main
    - java
      - org.apache.drill.contrib.function
        - blankFunction
    - resources
  - test
- drill-workshop-function.iml
- pom.xml
- External Libraries

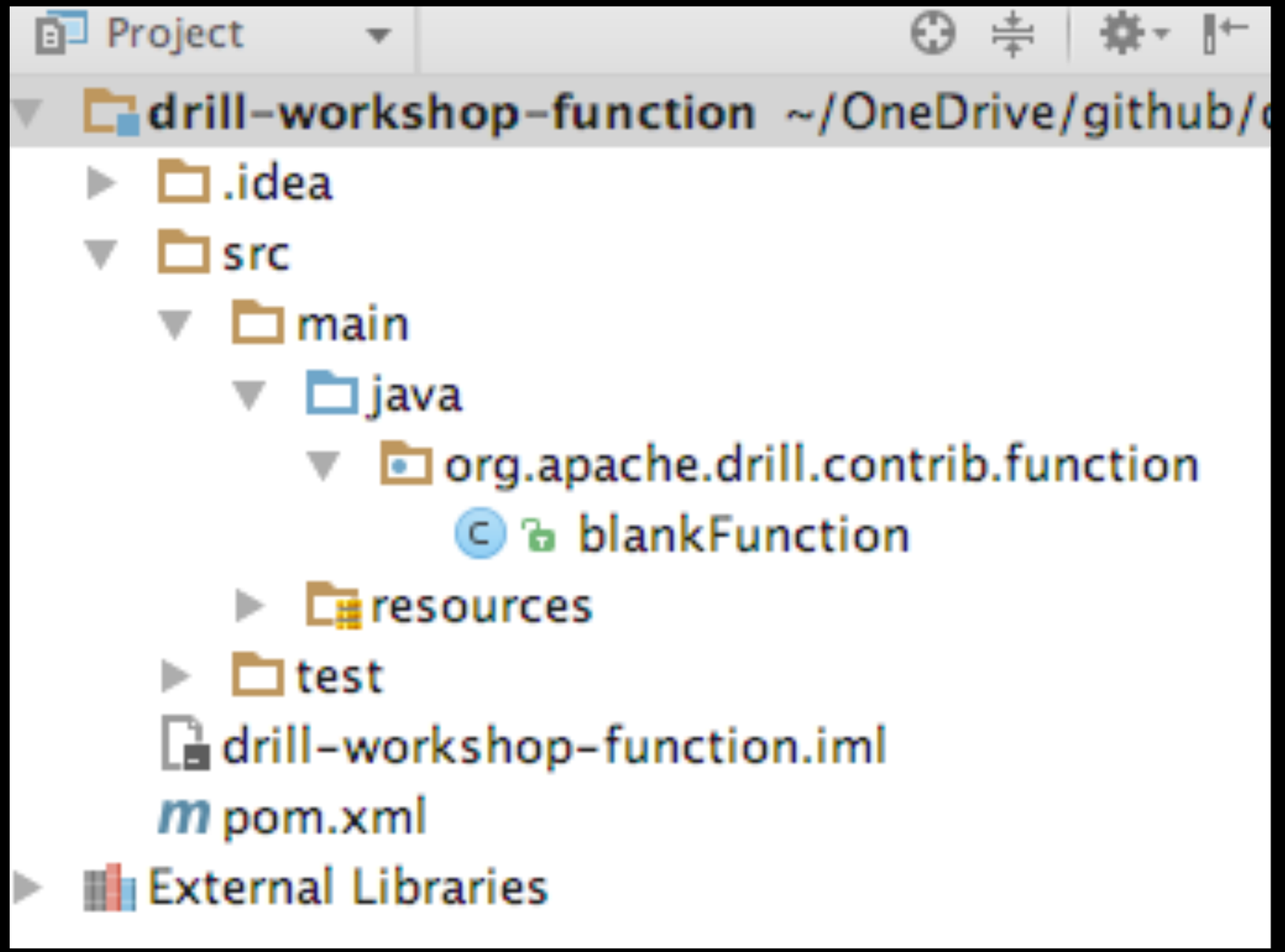
blankFunction.java ×

blankFunction | eval()

```
1  package org.apache.drill.contrib.function;
2
3  import io.netty.buffer.DrillBuf;
4  import org.apache.drill.exec.expr.DrillSimpleFunc;
5  import org.apache.drill.exec.expr.annotations.FunctionTemplate;
6  import org.apache.drill.exec.expr.annotations.Output;
7  import org.apache.drill.exec.expr.annotations.Param;
8  import org.apache.drill.exec.expr.annotations.Workspace;
9  import org.apache.drill.exec.expr.holders.BigIntHolder;
10 import org.apache.drill.exec.expr.holders.NullableVarCharHolder;
11
12 import javax.inject.Inject;
13
14
15 @FunctionTemplate(
16     name = "function_name",  //Put your function name here...
17     scope = FunctionTemplate.FunctionScope.SIMPLE,
18     nulls = FunctionTemplate.NullHandling.NULL_IF_NULL
19 )
20 public class blankFunction  implements DrillSimpleFunc {
21
22     @Param
23     NullableVarCharHolder inputTextA;  //put input parameters in this manner
24
25     @Output BigIntHolder out;  //output parameters like this
26
27     @Inject
28     DrillBuf buffer;
29
30     @Workspace
31     int x;  //Use the workspace to declare variables which are common for each function itera
```

1: Project | Z: Structure | Ant Build | Maven Projects

Writing A Drill Function



Charles' Rule #1 of Drill Functions:

Use an IDE!

Writing A Drill Function

```
@FunctionTemplate(
        name = "<function name>",
        scope = FunctionTemplate.FunctionScope.SIMPLE,
        nulls = FunctionTemplate.NullHandling.NULL_IF_NULL
)
public class <function file name> implements DrillSimpleFunc {


}
```

The function name is the actual function which will be called in a query. The file name **must** match the name in the class declaration.

Writing A Drill Function: Defining Input Parameters

```
@Param
NullableVarCharHolder inputTextA;

@Param
IntHolder someInt;

@Param
Float8Holder someDecimal;

@Param
ComplexHolder someMapOrArray;

import org.apache.drill.exec.expr.holders.XXX;
```

Writing A Drill Function

```
@FunctionTemplate(
        name = "addTwo",
        scope = FunctionTemplate.FunctionScope.SIMPLE,
        nulls = FunctionTemplate.NullHandling.NULL_IF_NULL
)
public class addTwoFunction implements DrillSimpleFunc {

    @Param
    IntHolder someInt;

}
```

Writing A Drill Function: Defining Output Parameters

```
@Output
BigIntHolder out;

@Inject
DrillBuf buffer;
```

Writing A Drill Function

```
@FunctionTemplate(
        name = "addTwo",
        scope = FunctionTemplate.FunctionScope.SIMPLE,
        nulls = FunctionTemplate.NullHandling.NULL_IF_NULL
)
public class addTwoFunction implements DrillSimpleFunc {

   @Param
  IntHolder someInt;

   @Output
   BigIntHolder out;

    @Inject
    DrillBuf buffer;

}
```

Writing A Drill Function: Workspace Parameters (Optional)

```
@Workspace
BigIntHolder temp1;
```

Writing A Drill Function: The function body

```
public void setup() {}

public void eval() {}
```

Writing A Drill Function: Accessing variables

For strings:

```
String s =
org.apache.drill.exec.expr.fn.impl.StringFunctionHelpers.toStringFromUTF8(inputTextA.
start, inputTextA.end, inputTextA.buffer);
```

For Numeric Fields:

```
long x = inputParam1.value;
Double y = inputParam2.value;
```

Writing A Drill Function: The function body

```
public void setup() {}


public void eval() {
    long x = inputParam1.value;
    long result = x + 2;
}
```

Would this work?

```
return result;
```

Writing A Drill Function: The function body

```
public void setup() {}


public void eval() {
    long x = inputParam1.value;
    long result = x + 2;
    out.value = result;
}
```

Writing A Drill Function: Returning values

Any numeric:

```
out.value = <number>
```

Any String

```
String outputValue = <some string>
out.buffer = buffer;
out.start = 0;
out.end = outputValue.getBytes().length;
buffer.setBytes(0, outputValue.getBytes());
```

Writing A Drill Function: Returning values

```
org.apache.drill.exec.vector.complex.writer.BaseWriter.MapWriter queryMapWriter =
outWriter.rootAsMap();

String[] arguments = queryString.split("&");

for (int i = 0; i < arguments.length; i++) {
   String[] queryParts = <some map or array>

   org.apache.drill.exec.expr.holders.VarCharHolder rowHolder = new
     org.apache.drill.exec.expr.holders.VarCharHolder();

  byte[] rowStringBytes = <value>.getBytes();
  outBuffer.reallocIfNeeded(rowStringBytes.length);
            outBuffer.setBytes(0, rowStringBytes);

   rowHolder.start = 0;
   rowHolder.end = rowStringBytes.length;
   rowHolder.buffer = outBuffer;
   queryMapWriter.varChar(<key>).write(rowHolder);
```

Writing A Drill Function: Compiling & Installing

```
mvn clean package -DskipTests
```

Next, if all went well, copy the .jar files from the <project>/target folder to: <drill path>/jars/3rdparty and restart Drill.

Now you are ready to try your function in a query!

In Class Exercise

Write a Drill UDF called `digit_to_string()` which takes a single digit as an argument and returns the text version of that integer.

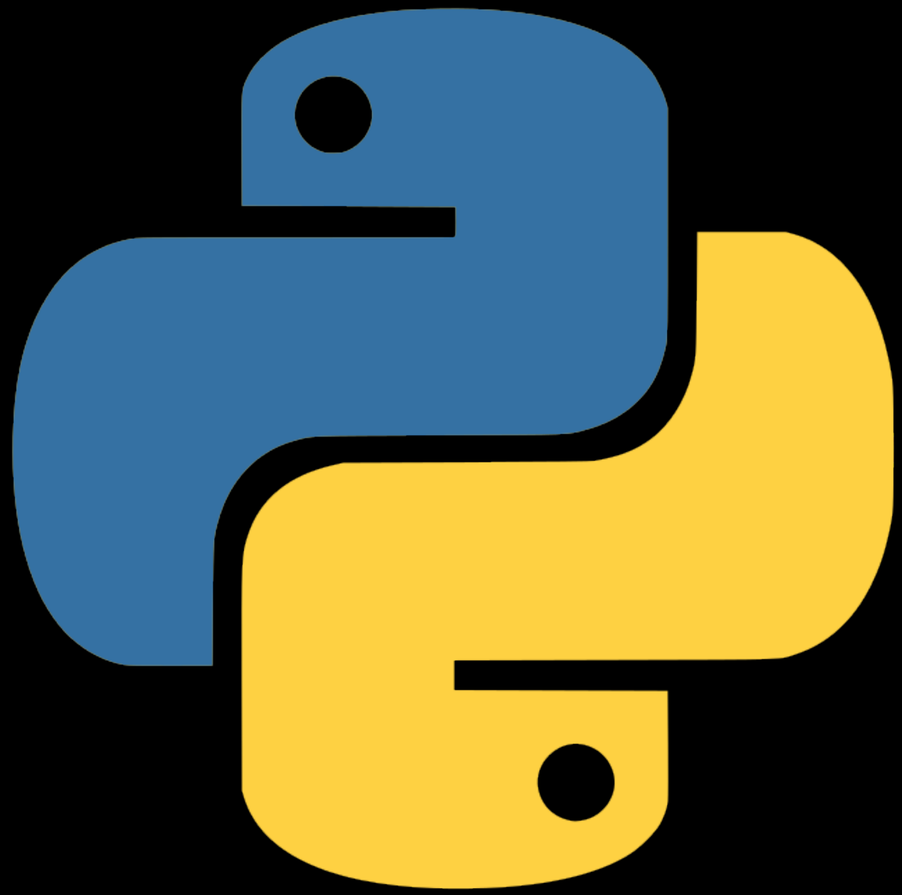
IE:

```
digit_to_string( 3 ) = three
```

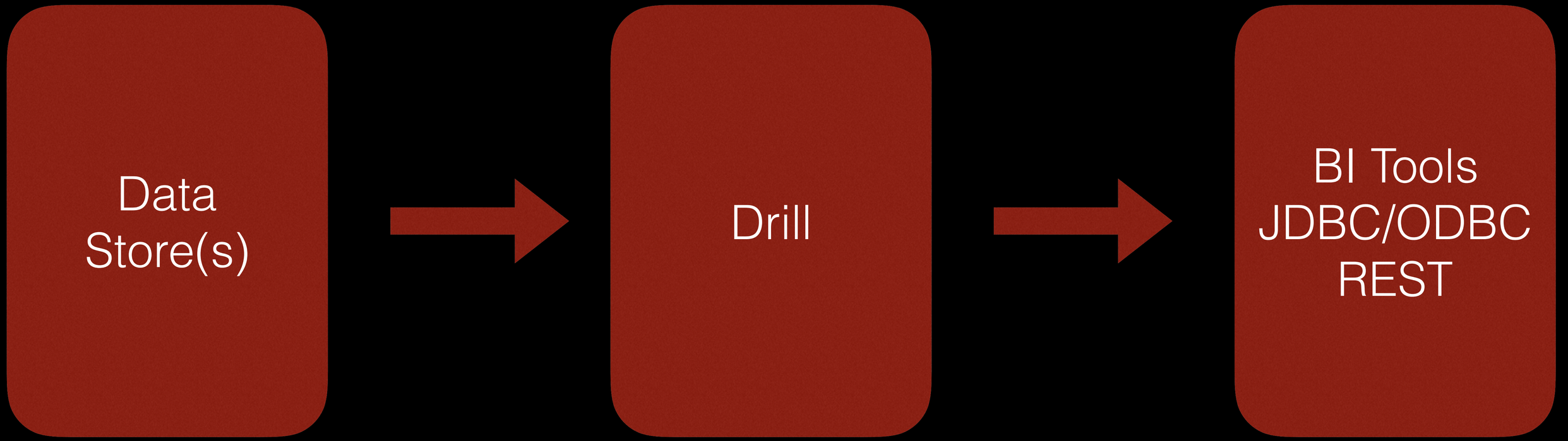
What is most important is understanding how to get variables in and out of the `eval()` function.

Connecting to Drill

Python



Connecting to Drill



Connecting to Drill

```
pip install pydrill
```

Connecting to Drill

```
from pydrill.client import PyDrill
```

Connecting to Drill

```
drill = PyDrill(host='localhost', port=8047)

if not drill.is_active():
    raise ImproperlyConfigured('Please run Drill first')
```

Connecting to Drill

```
query_result = drill.query('''
 SELECT JobTitle,
  AVG( CAST( LTRIM( AnnualSalary, '$' ) AS FLOAT) ) AS
avg_salary,
COUNT( DISTINCT name ) AS number
FROM dfs.drillclass.`*.csvh`
GROUP BY JobTitle
Order By avg_salary DESC
LIMIT 10
''')
```

Connecting to Drill

```
df = query_result.to_dataframe()
```

gtkcyber.com

Sergeant

- DBI
- RJDBC
- dplyr

See complete documentation: https://github.com/hrbrmstr/sergeant

```
devtools::install_github("hrbrmstr/sergeant")
```

See complete documentation: https://github.com/hrbrmstr/sergeant

```
library(sergeant)
connection <- drill_connection("localhost")
drill_active(connection)
query_result <- drill_query(connection,
"SELECT * FROM cp.`employee.json` limit 100”
)
```

See complete documentation: https://github.com/hrbrmstr/sergeant

In Class Exercise

Complete the Scripting Demonstration Worksheet.

Questions?

Thank you!!

Charles S. Givre
@cgivre
thedataist.com
linkedin.com/in/cgivre