Apache Drill Workshop

Charles S. Givre
givre_charles@bah.com
@cgivre
thedataist.com

DARK LABS

What is Drill?

DARK LABS

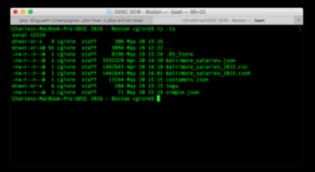
Data is not arranged in an optimal way for ad-hoc analysis

DARK LABS

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

ETL

DARK LABS



APACHE DRILL

hadoop

ETL

HIVE

CSV

JSON

darklabs.bah.com

DARK LABS

You just query the data...
no schema

Drill is NOT just SQL on Hadoop

DARK LABS

Drill scales

Drill is open source

Download Drill at: drill.apache.org

Quick Demo

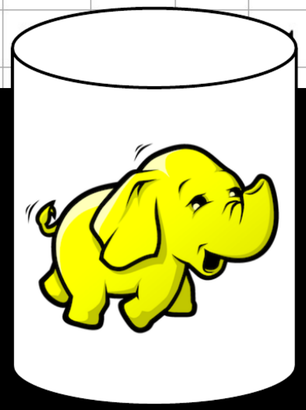
Thank you Jair Aguirre!!

DARK LABS

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

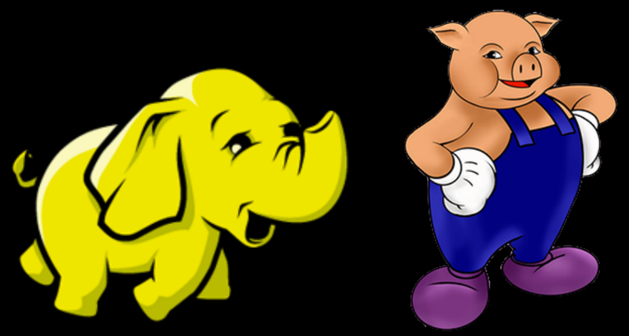


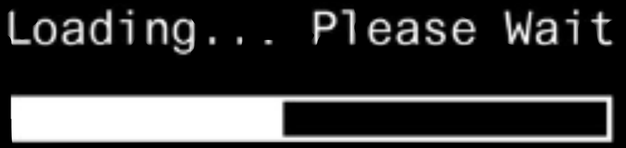
darklabs.bah.com

DARK LABS

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

DARK LABS

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;
```

APACHE DRILL

Execution Time:
0.89 seconds!!

NoSQL, No Problem

DARK LABS

NoSQL, No Problem

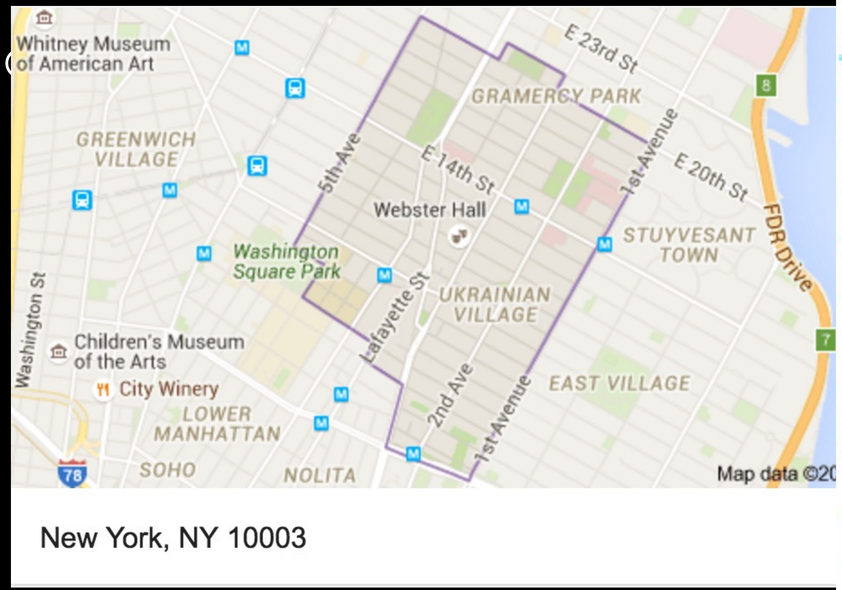

```
{
  "address": {
     "building": "1007",
     "coord": [ -73.856077, 40.848447 ],
     "street": "Morris Park Ave",
     "zipcode": "10462"
  },
  "borough": "Bronx",
  "cuisine": "Bakery",
  "grades": [
     { "date": { "$date": 1393804800000 }, "grade": "A", "score": 2 },
     { "date": { "$date": 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

DARK LABS

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

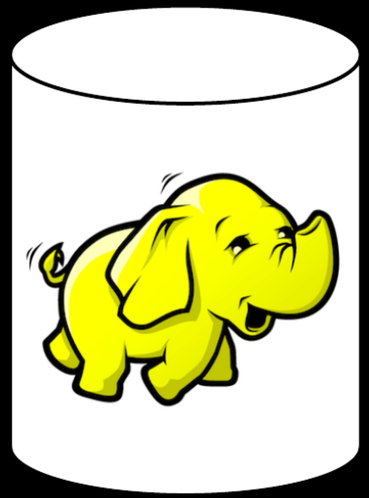
https://raw.githubusercontent.com/mongodb/docs-assets/primer-dataset/primer-dataset.json

DARK LABS

Querying Across Silos

DARK LABS

Querying Across Silos



Farmers Market Data

Restaurant Data

DARK LABS

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;
```

DARK LABS

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


WARNING
DANGER
DO NOT PULL
HANDLE
TOMCAT
RANGER
CV 1336
VCENTURION

DARK LABS

Installing Drill

DARK LABS

Installing Drill

1. Download Tarball from drill.apache.org
2. Unzip Tarball.

DARK LABS

Starting Drill

DARK LABS

Starting Drill

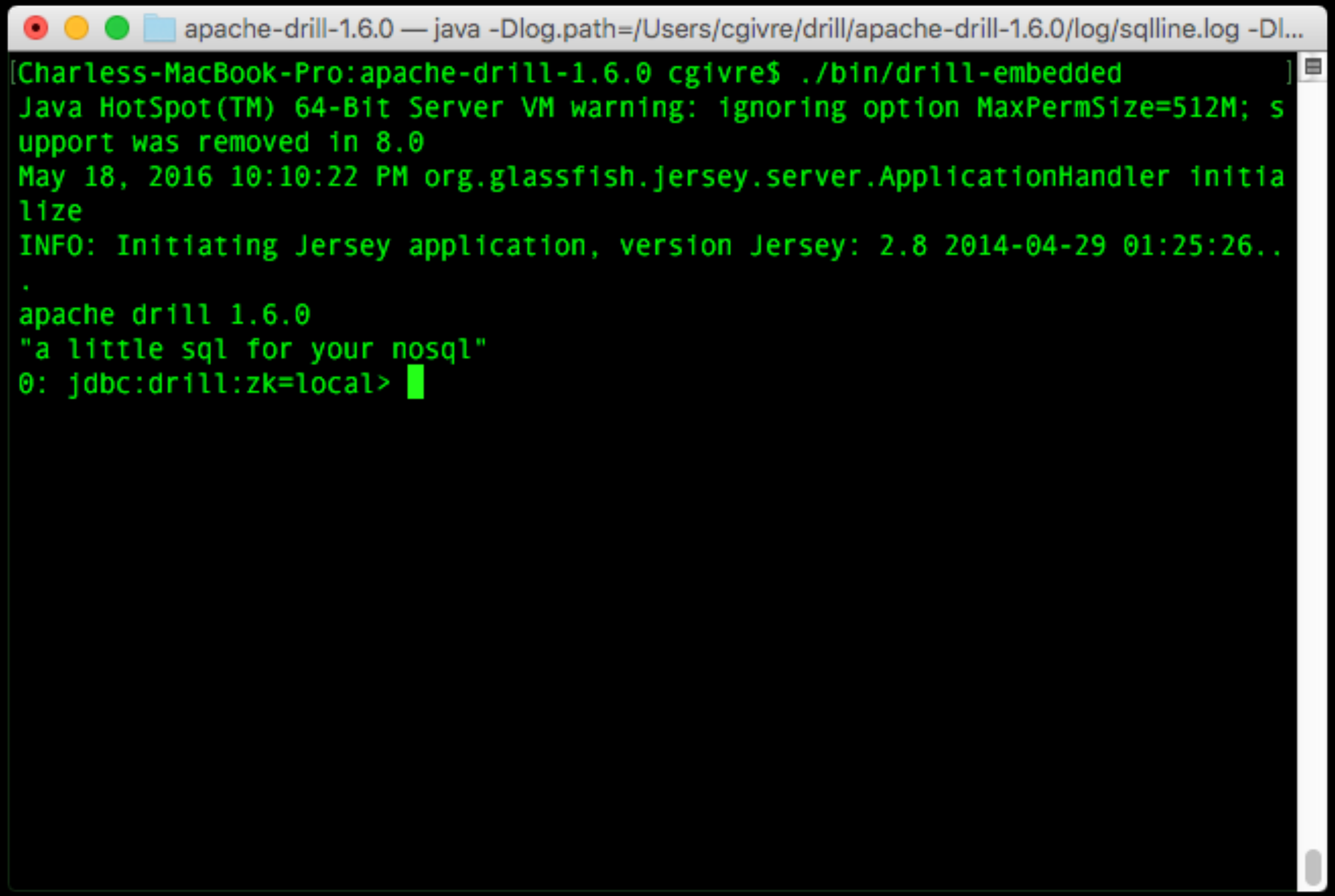
Embedded Mode: For use on a standalone system

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

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

DARK LABS

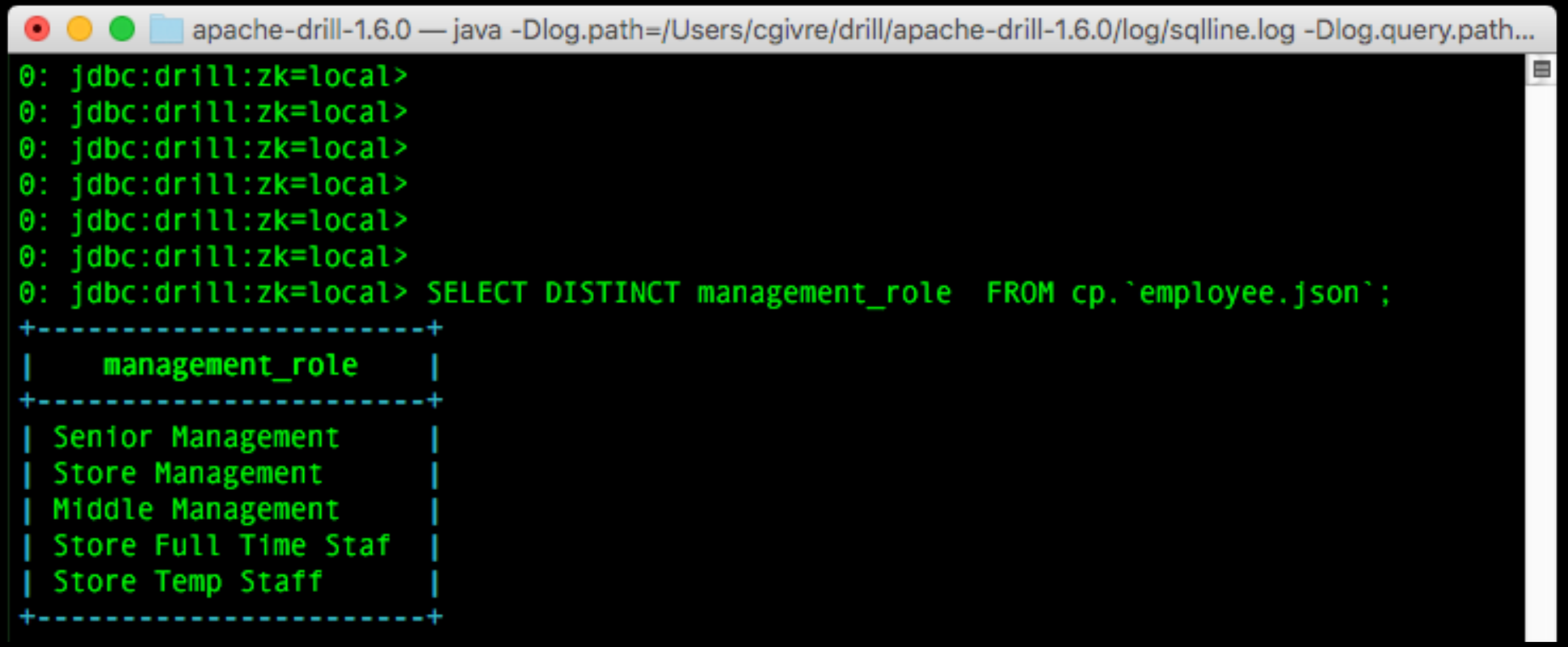
Querying Drill



DARK LABS

Querying Drill

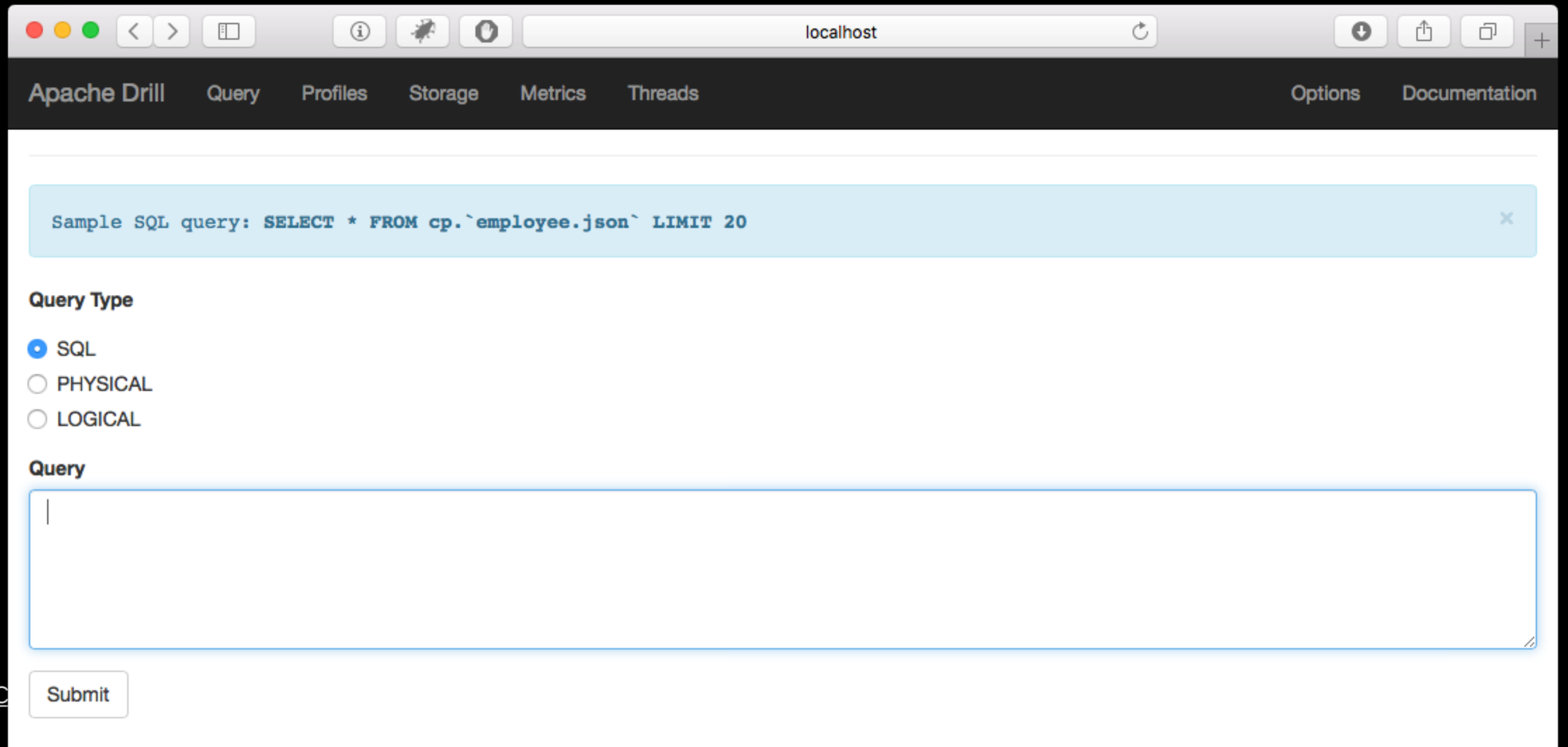
```
SELECT DISTINCT management_role FROM cp.`employee.json`;
```



DARK LABS

Querying Drill

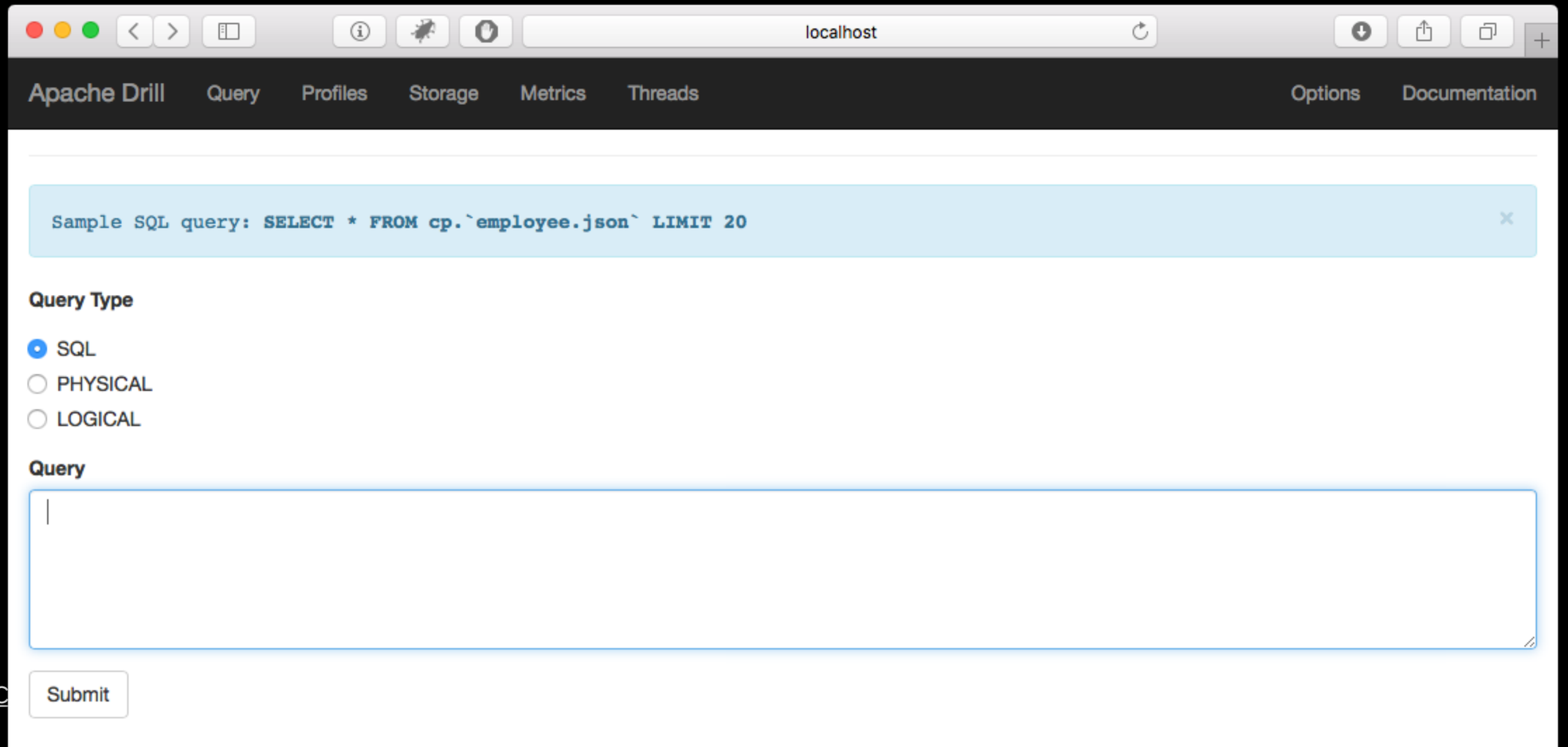
http://localhost:8047



DARK LABS

Querying Drill

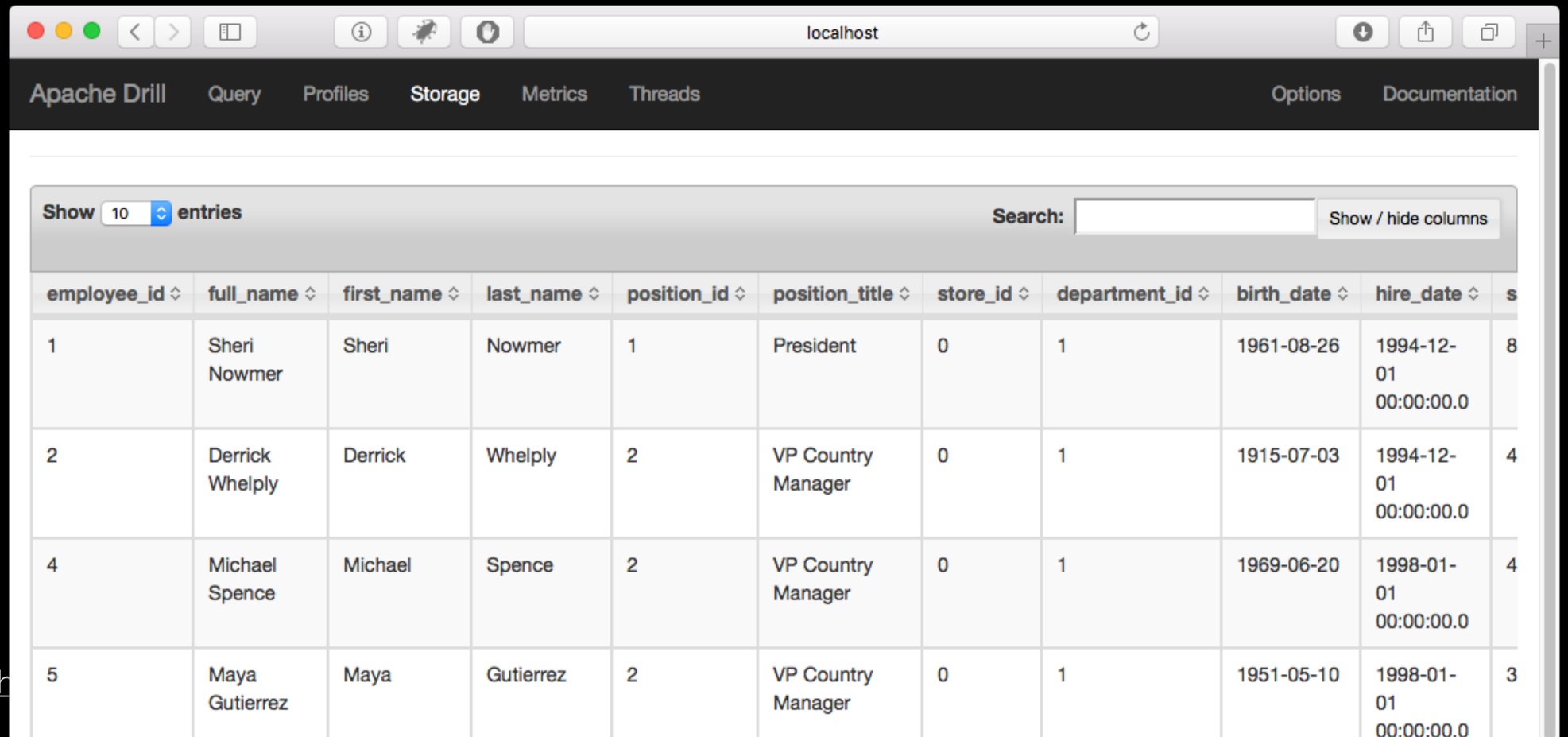
```
SELECT * FROM cp.`employee.json` LIMIT 20
```



DARK LABS

Querying Drill

```
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

DARK LABS

Querying Drill

```
SELECT <fields>
FROM <table>
WHERE <optional logical condition>
```

DARK LABS

Querying Drill

```
SELECT name, address, email
FROM customerData
WHERE age > 20
```

DARK LABS

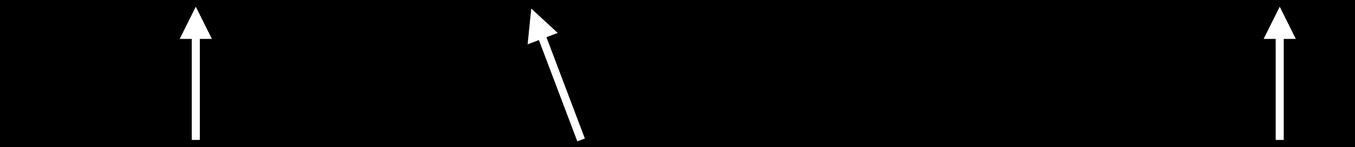
Querying Drill

```
SELECT name, address, email
FROM dfs.logs.`/data/customers.csv`
WHERE age > 20
```

DARK LABS

Querying Drill

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



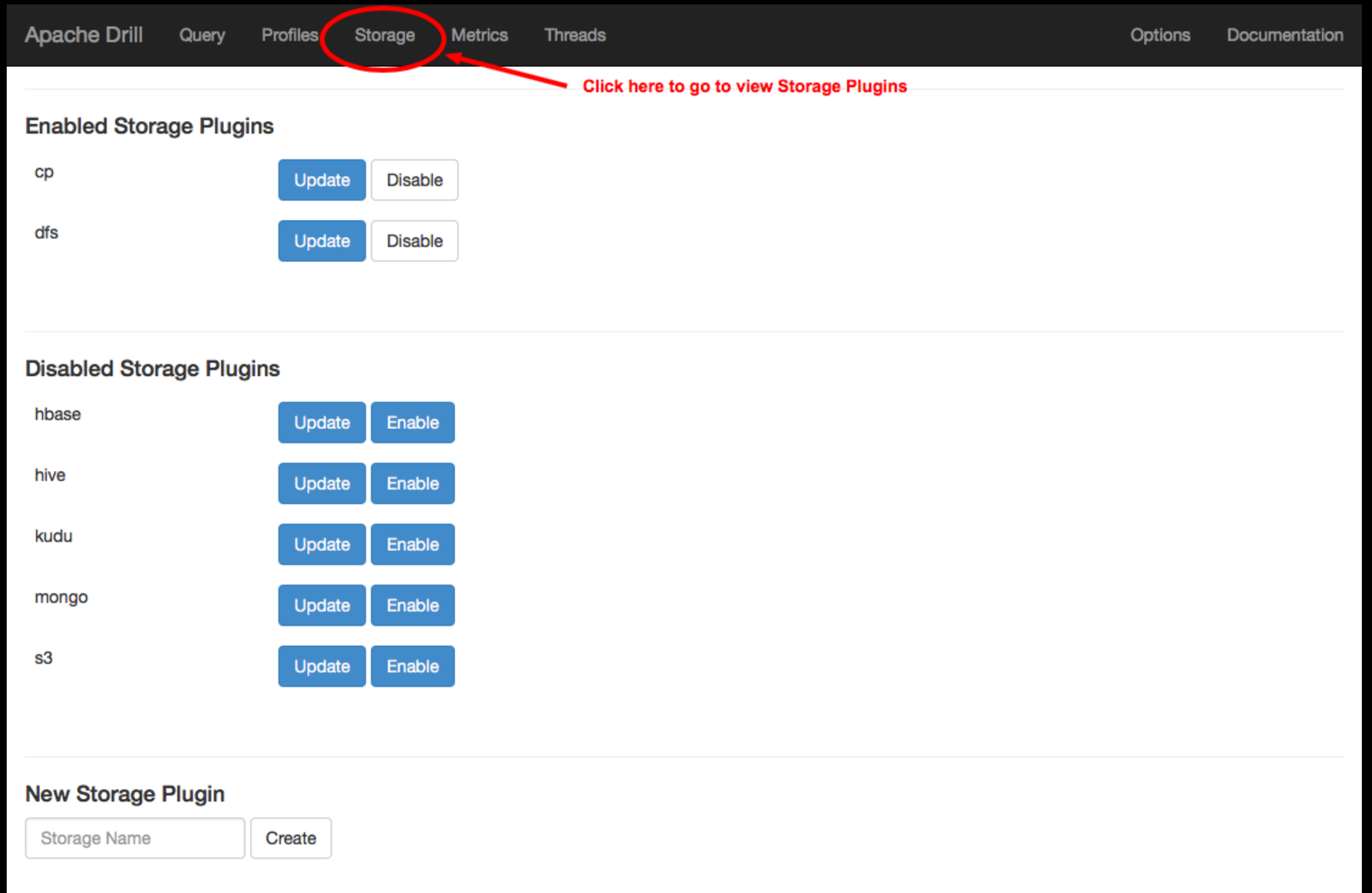
DARK LABS

Querying Drill

Plugins Supported	Description
cp	Queries files in the Java ClassPath
dfs	File System. Can connect to remote filesystems such as Hadoop
hbase	Connects to HBase
hive	Integrates Drill with the Apache Hive metastore
kudu	Provides a connection to Apache Kudu
mongo	Connects to mongoDB
RDBMS	Provides a connection to relational databases such as MySQL, Postgres, Oracle and others.
S3	Provides a connection to an S3 cluster

DARK LABS

Querying Drill

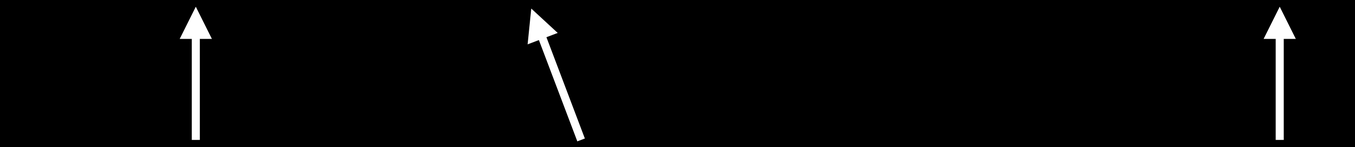


darklabs.bah.com

DARK LABS

Querying Drill

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



DARK LABS

Querying Drill

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


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

In Class Exercise: Create a Workspace

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

1. First, download all the files from https://github.com/cgivre/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

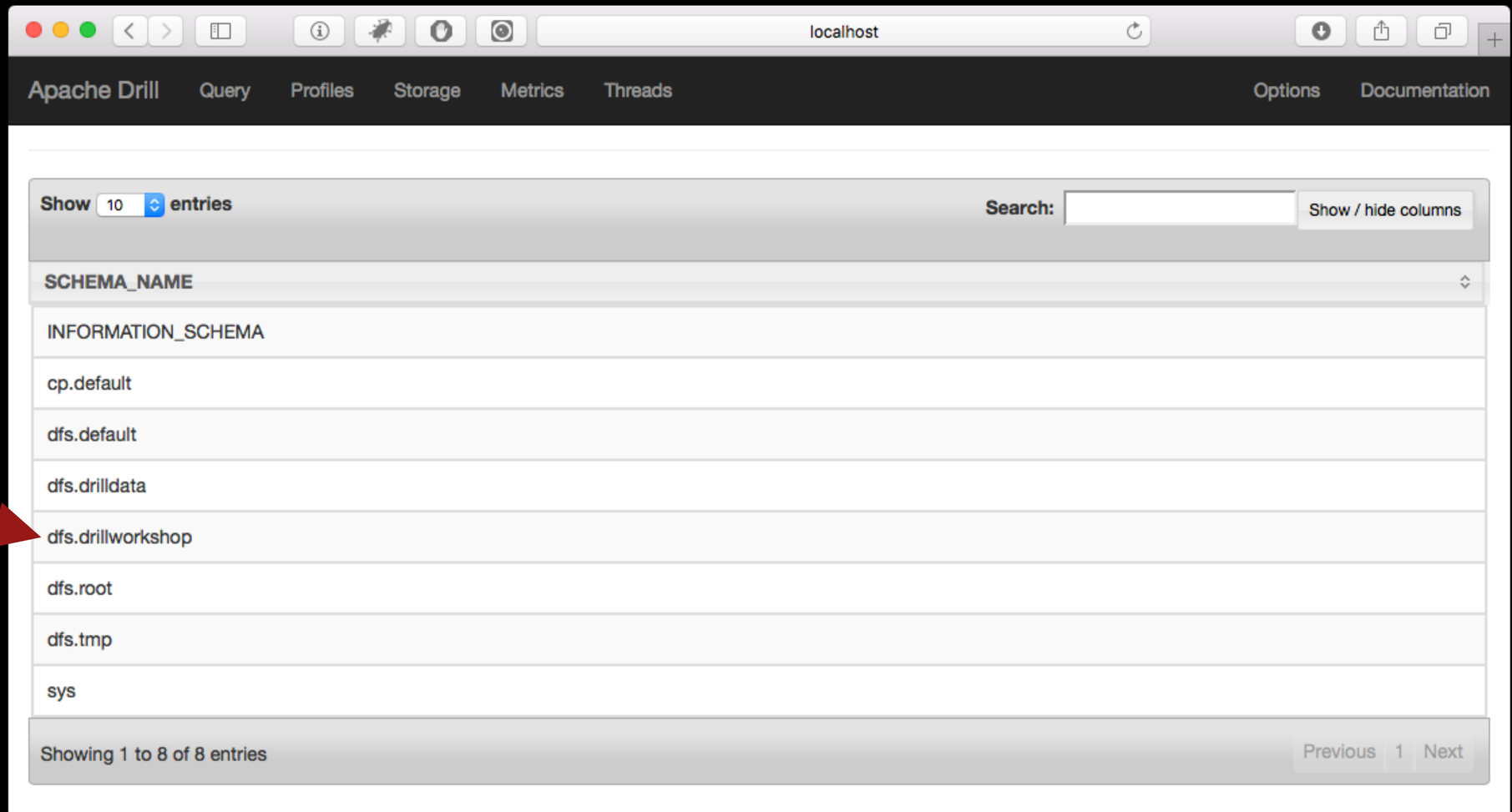
3. Paste the following into the 'workspaces' section and click update

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

DARK LABS

Querying Drill

```
SHOW databases;
```

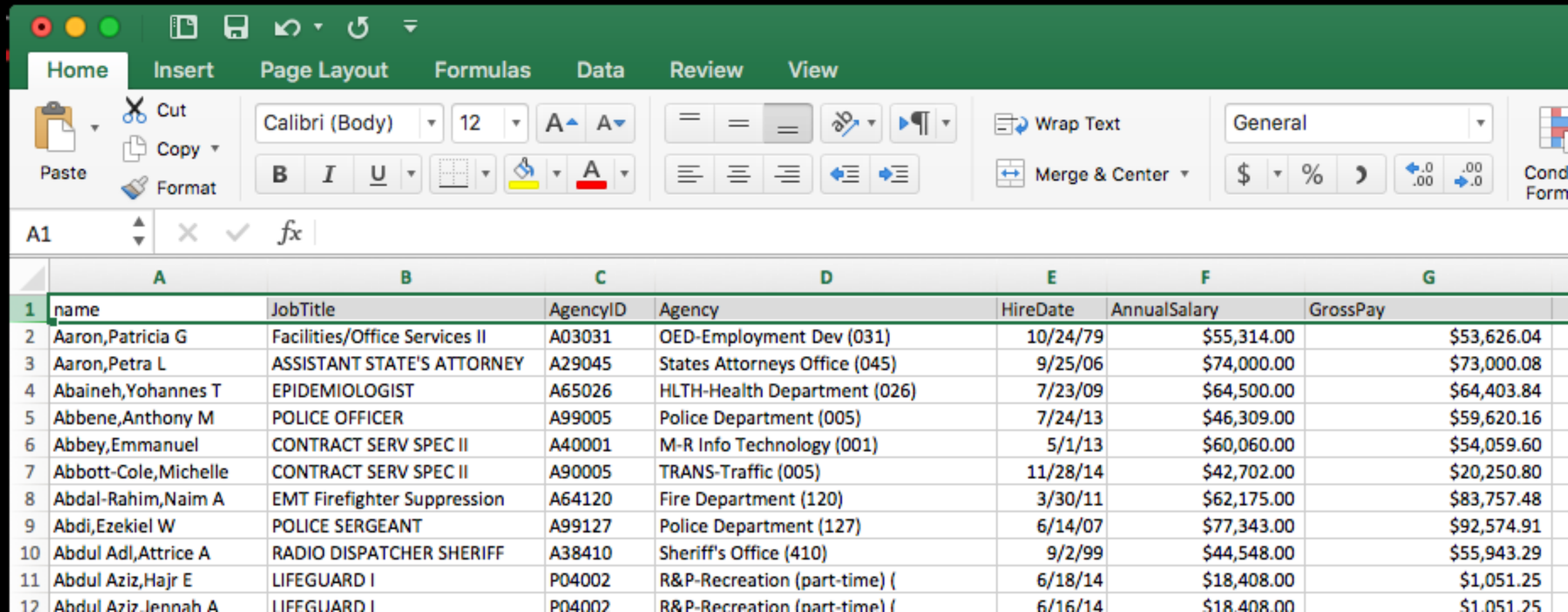


Success!!

darklabs.bah.com

DARK LABS

Querying Drill



	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

darklabs.bah.com

DARK LABS

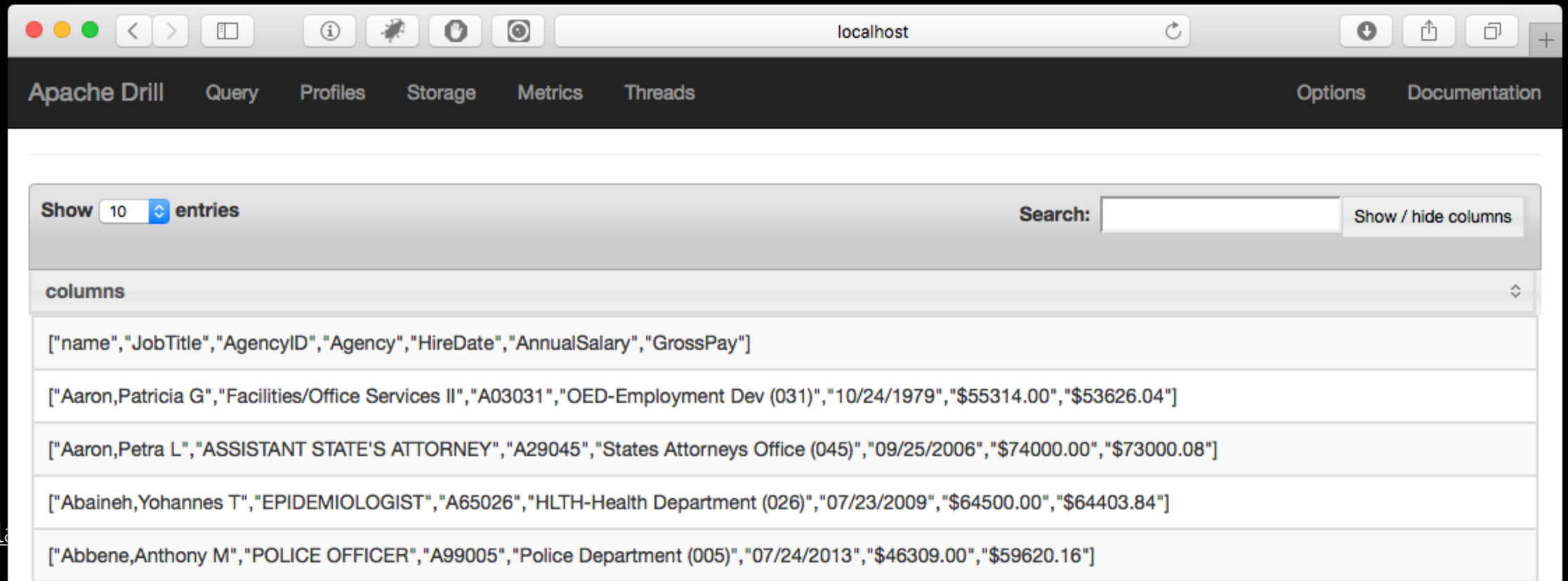
Querying Drill

```
SELECT *
FROM dfs.drillworkshop.`baltimore_salaries_2015.csv`
LIMIT 10
```

DARK LABS

Querying Drill

```
SELECT *
FROM dfs.drillworkshop.`baltimore_salaries_2015.csv`
LIMIT 10
```



DARK LABS

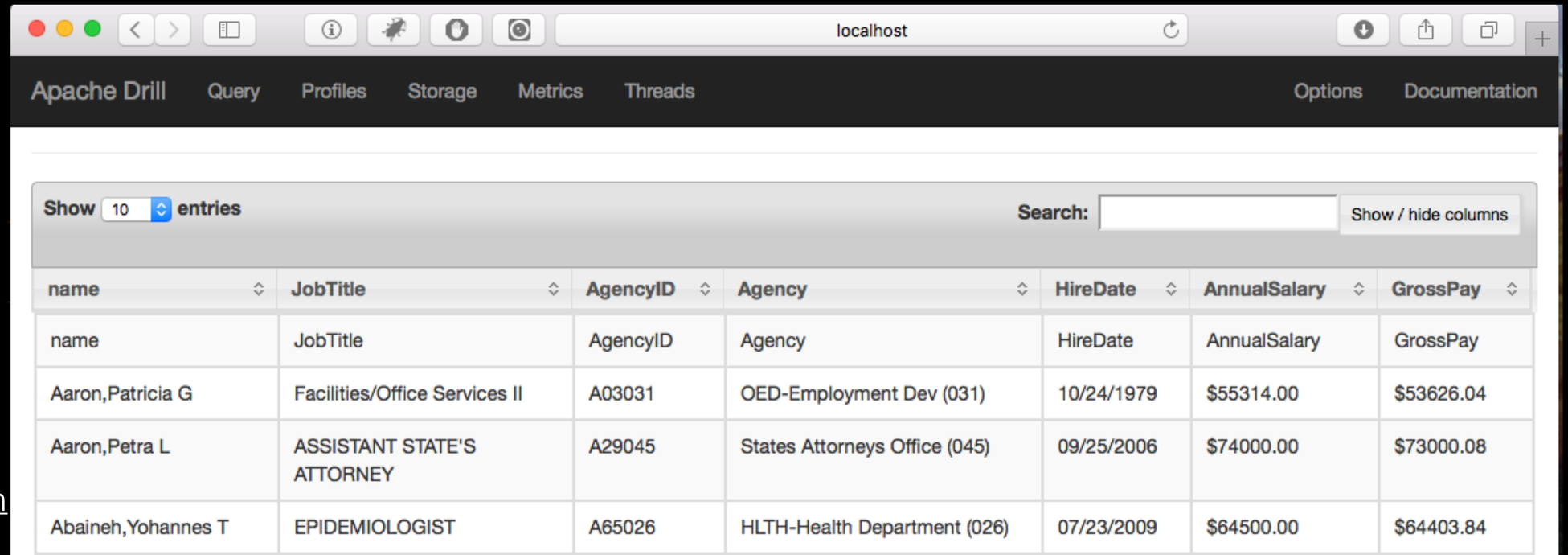
Querying Drill

```
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.drillworkshop.`baltimore_salaries_2015.csv`
LIMIT 10
```

DARK LABS

Querying Drill

```
SELECT columns[0] AS name,
columns[1] AS JobTitle,
. . .
FROM dfs.drillworkshop.`baltimore_salaries_2015.csv`
LIMIT 10
```

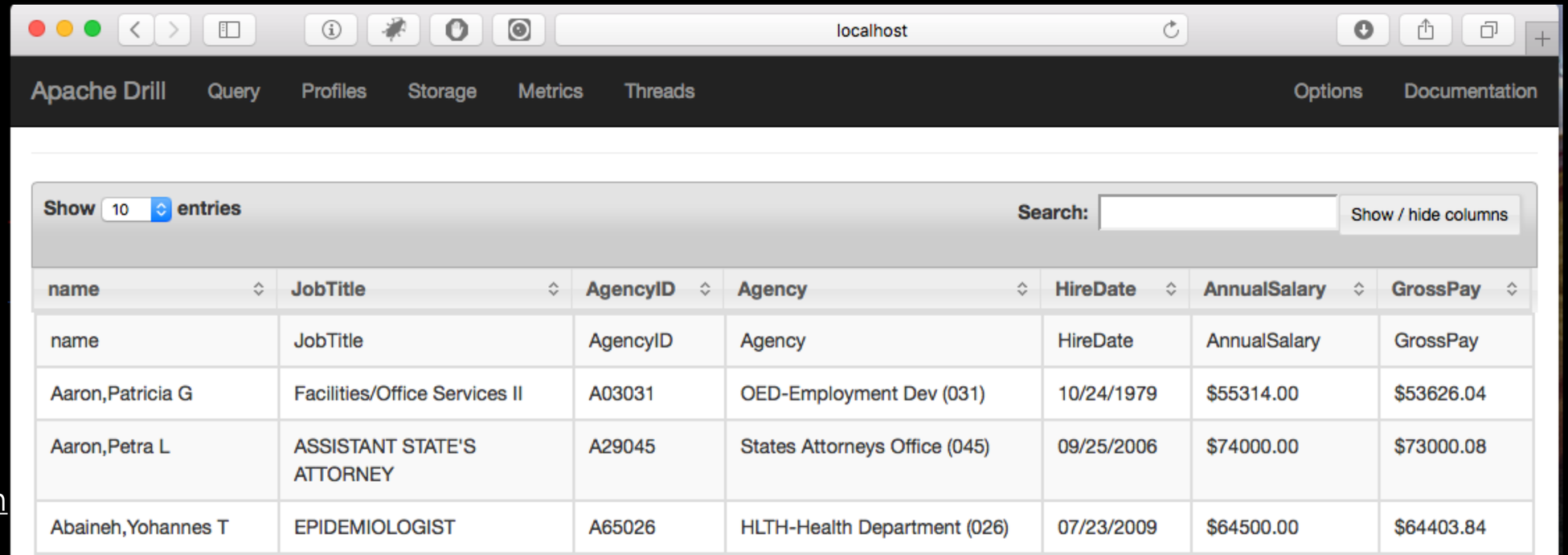


name	JobTitle	AgencyID	Agency	HireDate	AnnualSalary	GrossPay
name	JobTitle	AgencyID	Agency	HireDate	AnnualSalary	GrossPay
Aaron,Patricia G	Facilities/Office Services II	A03031	OED-Employment Dev (031)	10/24/1979	$55314.00	$53626.04
Aaron,Petra L	ASSISTANT STATE'S ATTORNEY	A29045	States Attorneys Office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

DARK LABS

Querying Drill

```
SELECT columns[0] AS name,
columns[1] AS JobTitle,
. . .
FROM dfs.drillworkshop.`baltimore_salaries_2015.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

DARK LABS

Querying Drill

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

DARK LABS

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

DARK LABS

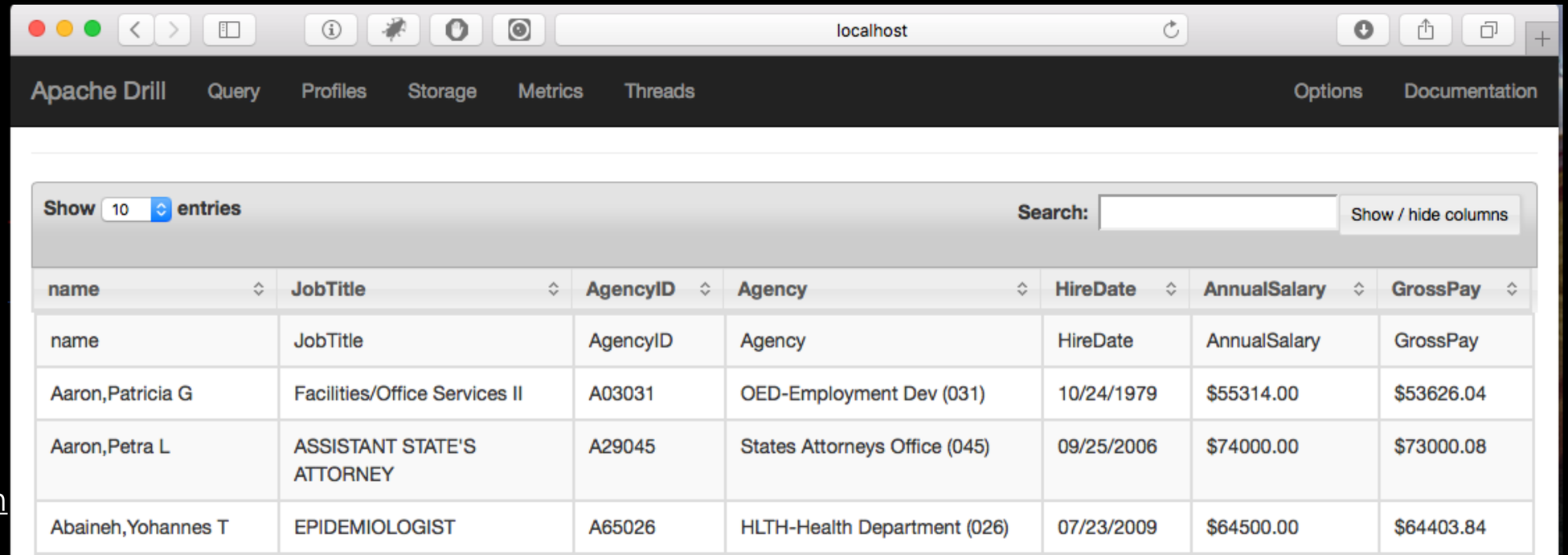
Querying Drill

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

DARK LABS

Querying Drill

```
SELECT *
FROM
dfs.drillworkshop.`baltimore_salaries_2015.csvh`
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

DARK LABS

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

DARK LABS

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

DARK LABS

Querying Drill

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

DARK LABS

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]

DARK LABS

Querying Drill

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

DARK LABS

Querying Drill

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

DARK LABS

AnnualPay has extra characters

AnnualPay is a string

DARK LABS

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
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/

```
SELECT LTRIM( AnnualPay, ‘$’ ) AS annualPay
FROM dfs.drillworkshop.`*.csvh`
```

DARK LABS

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:

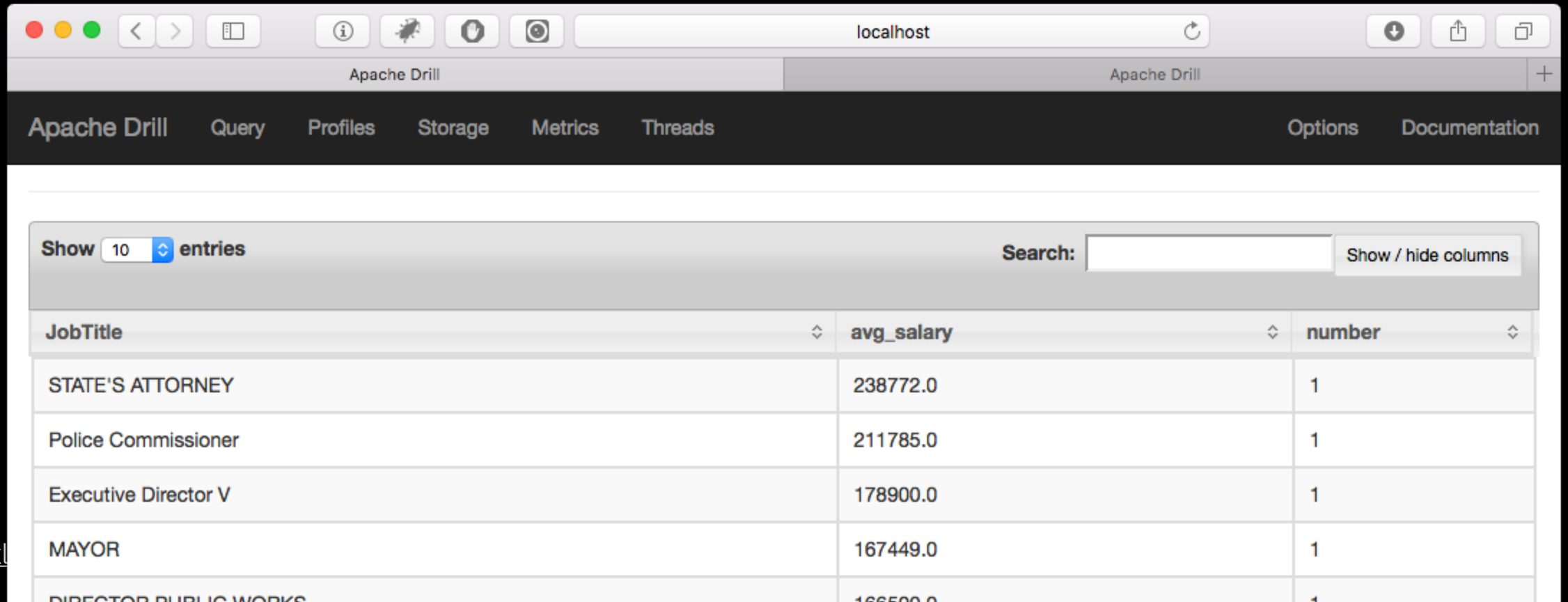
https://drill.apache.org/docs/data-type-conversion/#cast

```
SELECT CAST( LTRIM( AnnualPay, '$' ) AS FLOAT ) AS
annualPay
FROM dfs.drillworkshop.`*.csvh`
```


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

DARK LABS

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



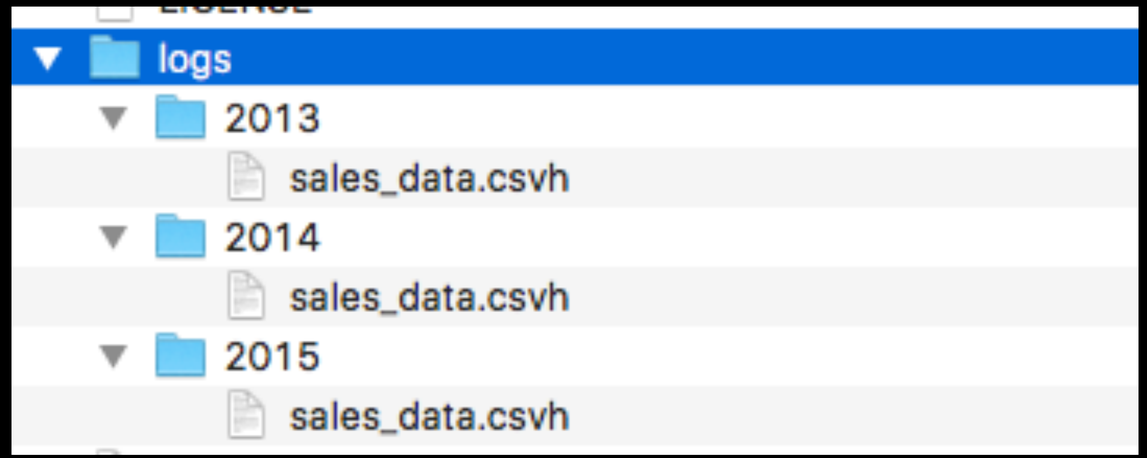
DARK LABS

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

DARK LABS

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

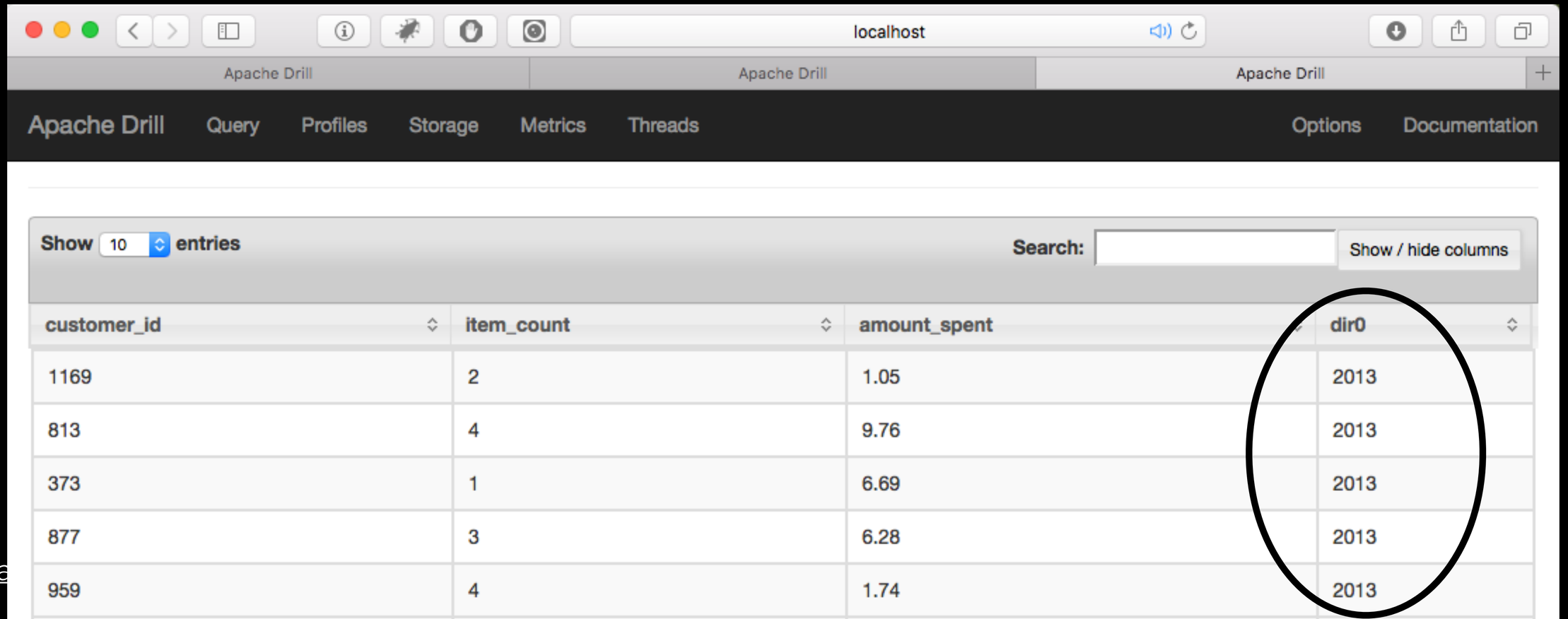
- In the sample data files, there is a folder called ‘logs’ which contains the following structure:




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

DARK LABS

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



DARK LABS

dir*n* accesses the subdirectories

DARK LABS

dir*n* accesses the subdirectories

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

DARK LABS

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>')
```

DARK LABS

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.drillworkshop.`logs/`
GROUP BY dir0
```

Let’s look at JSON data

DARK LABS

Let's look at JSON data

```
[
   {
      "name": "Farley, Colette L.",
      "email": "iaculis@atarcu.ca",
      "DOB": "2011-08-14",
      "phone": "1-758-453-3833"
   },
   {
      "name": "Kelley, Cherokee R.",
      "email": "ante.blandit@malesuadafringilla.edu",
      "DOB": "1992-09-01",
      "phone": "1-595-478-7825"
   }
…
]
```

DARK LABS

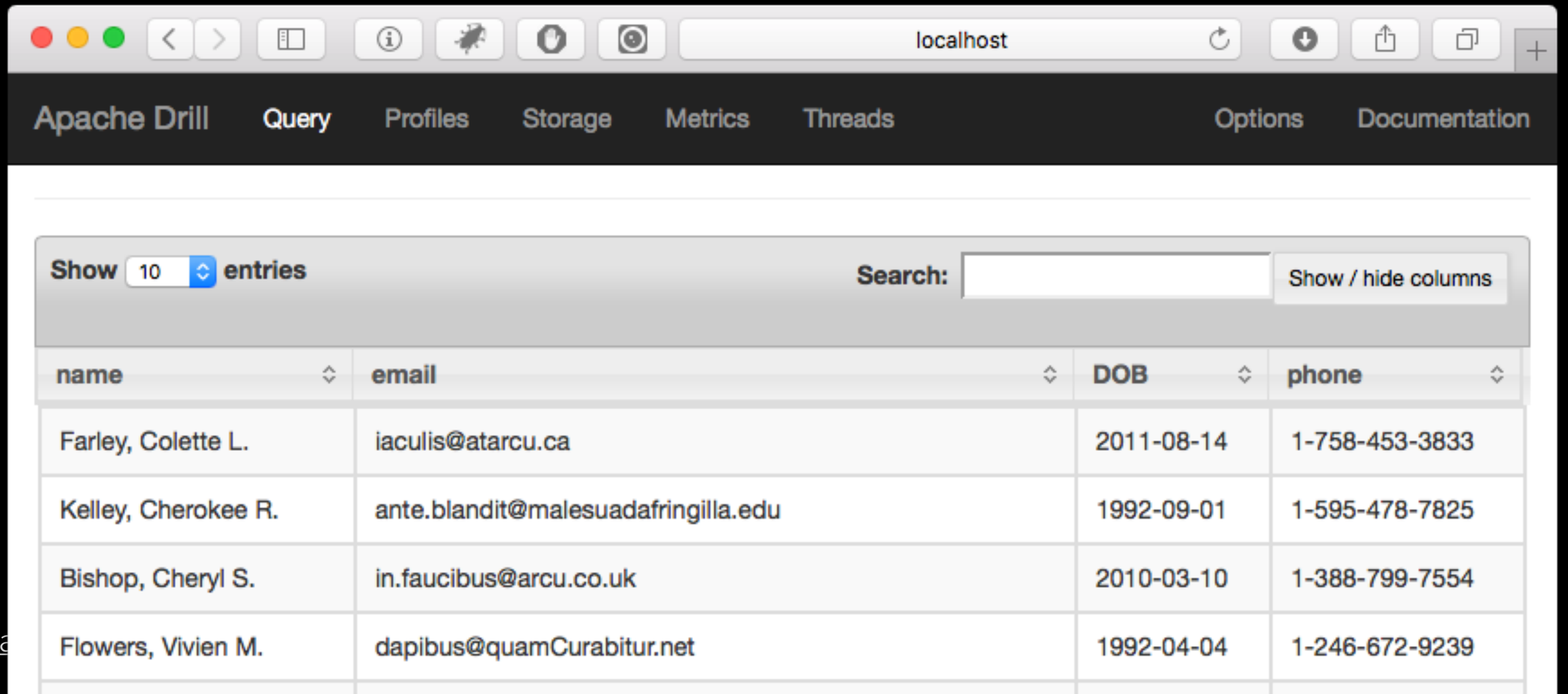
Let's look at JSON data

```
SELECT *
FROM dfs.drillworkshop.`customers.json`
```

DARK LABS

Let's look at JSON data

```
SELECT *
FROM dfs.drillworkshop.`customers.json`
```



DARK LABS

Let's look at JSON data

```
SELECT *
FROM dfs.drillworkshop.`customers.json`
```

What about nested data?

DARK LABS

Please open
`baltimore_salaries.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" ]
```

DARK LABS


```
{
  "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"
    ]
```

DARK LABS

Drill has a series of functions for nested data

Please run

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

in Drill

DARK LABS

Let's look at this data in Drill

DARK LABS

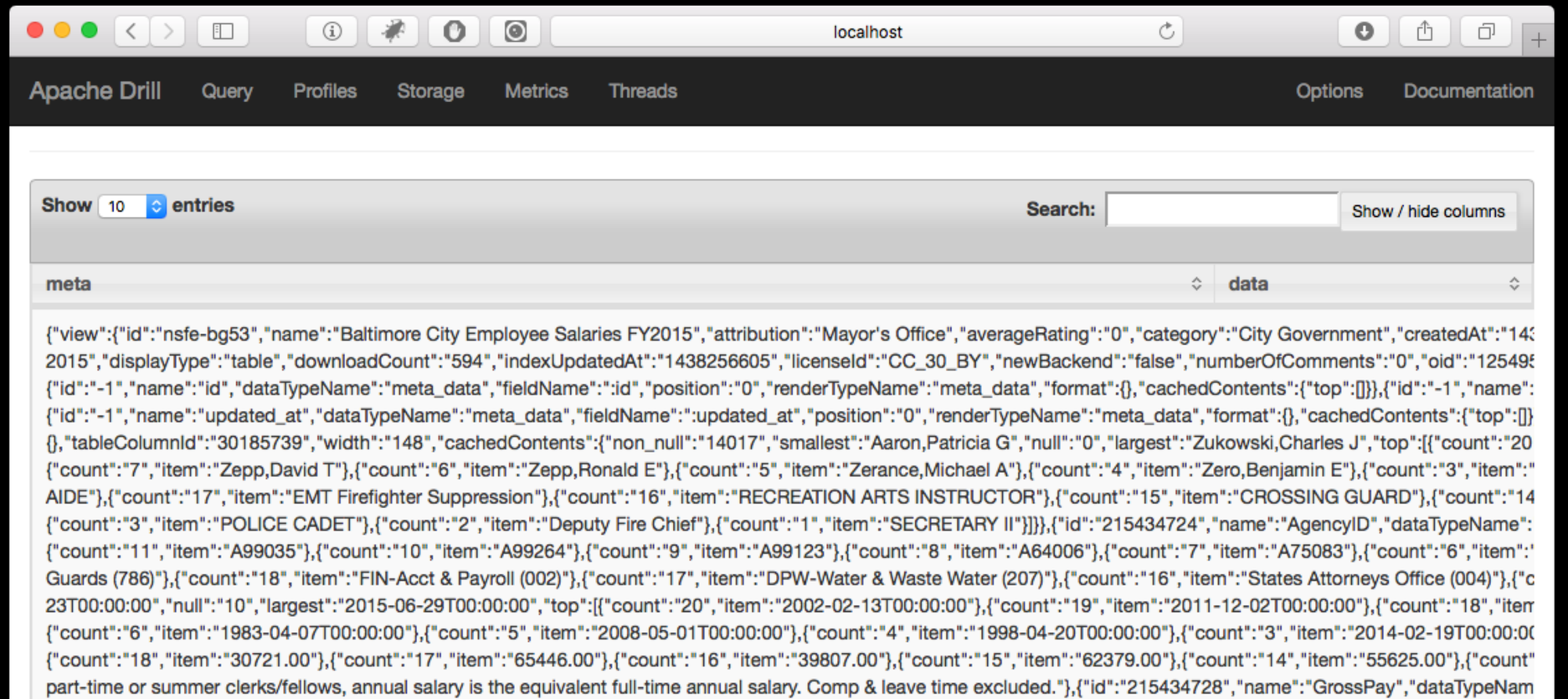
Let's look at this data in Drill

```
SELECT *
FROM dfs.drillworkshop.`baltimore_salaries.json`
```

DARK LABS

Let's look at this data in Drill

```
SELECT *
FROM dfs.drillworkshop.`baltimore_salaries.json`
```

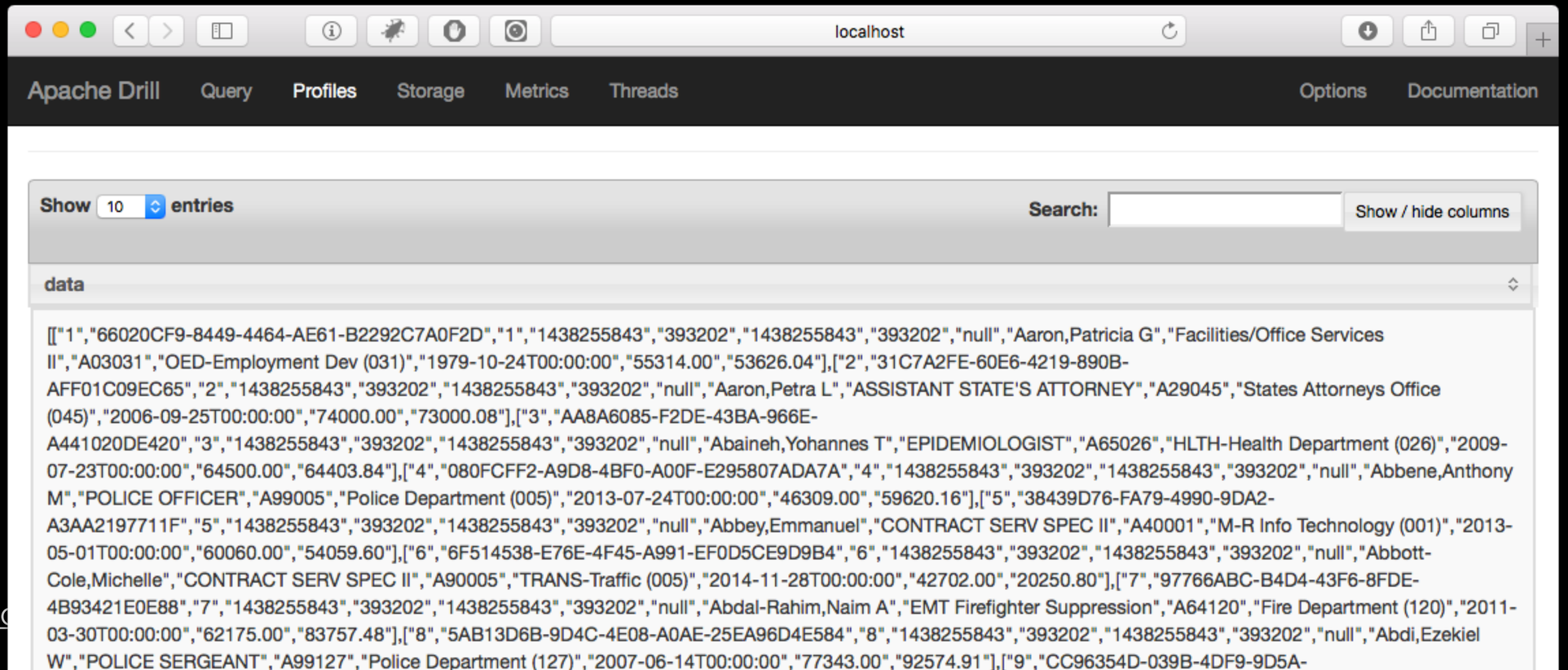


darklabs.bah.com

DARK LABS

Let's look at this data in Drill

```
SELECT data
FROM dfs.drillworkshop.`baltimore_salaries.json`
```



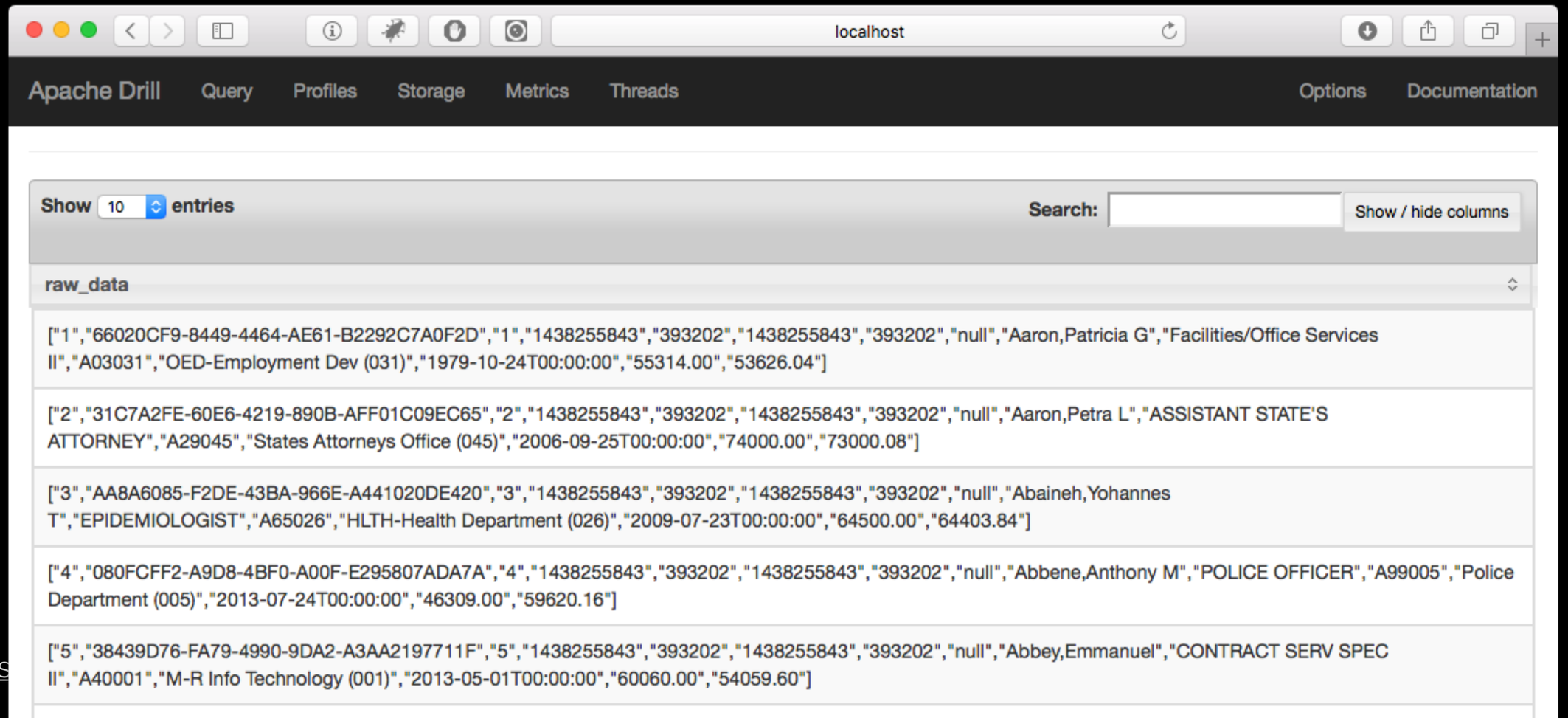
DARK LABS

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


```
SELECT FLATTEN( data ) AS raw_data
FROM dfs.drillworkshop.`baltimore_salaries.json`
```

DARK LABS

```
SELECT FLATTEN( data ) AS raw_data
FROM dfs.drillworkshop.`baltimore_salaries.json`
```

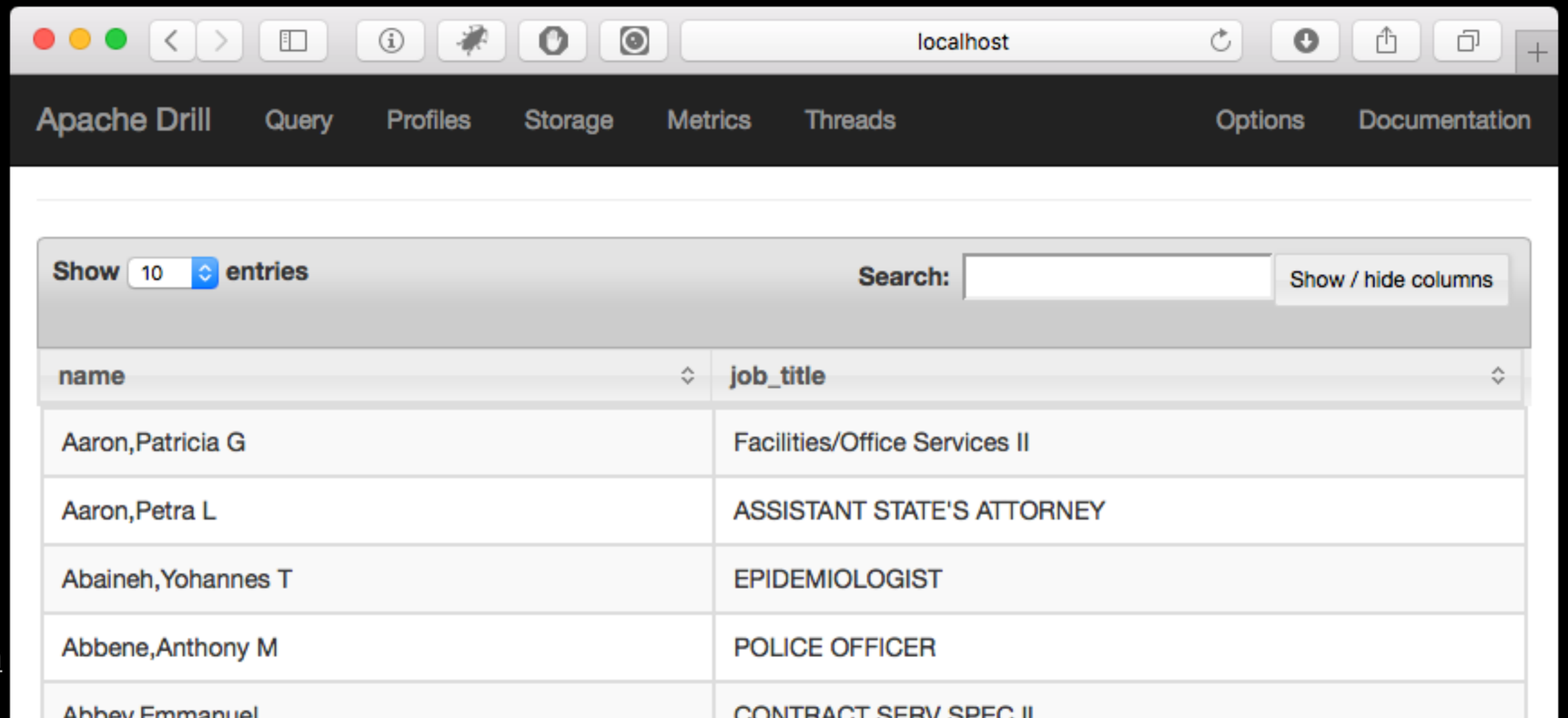



```
SELECT FLATTEN( data ) AS raw_data
FROM dfs.drillworkshop.`baltimore_salaries.json`
```


```
SELECT raw_data[8] AS name …
FROM
(
SELECT FLATTEN( data ) AS raw_data
FROM dfs.drillworkshop.`baltimore_salaries.json`
)
```

DARK LABS

```
SELECT raw_data[8] AS name, raw_data[9] AS job_title
FROM
(
SELECT FLATTEN( data ) AS raw_data
FROM dfs.drillworkshop.`baltimore_salaries.json`
)
```



DARK LABS

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.

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

HINT 2: GROUP BY does NOT support aliases.

DARK LABS

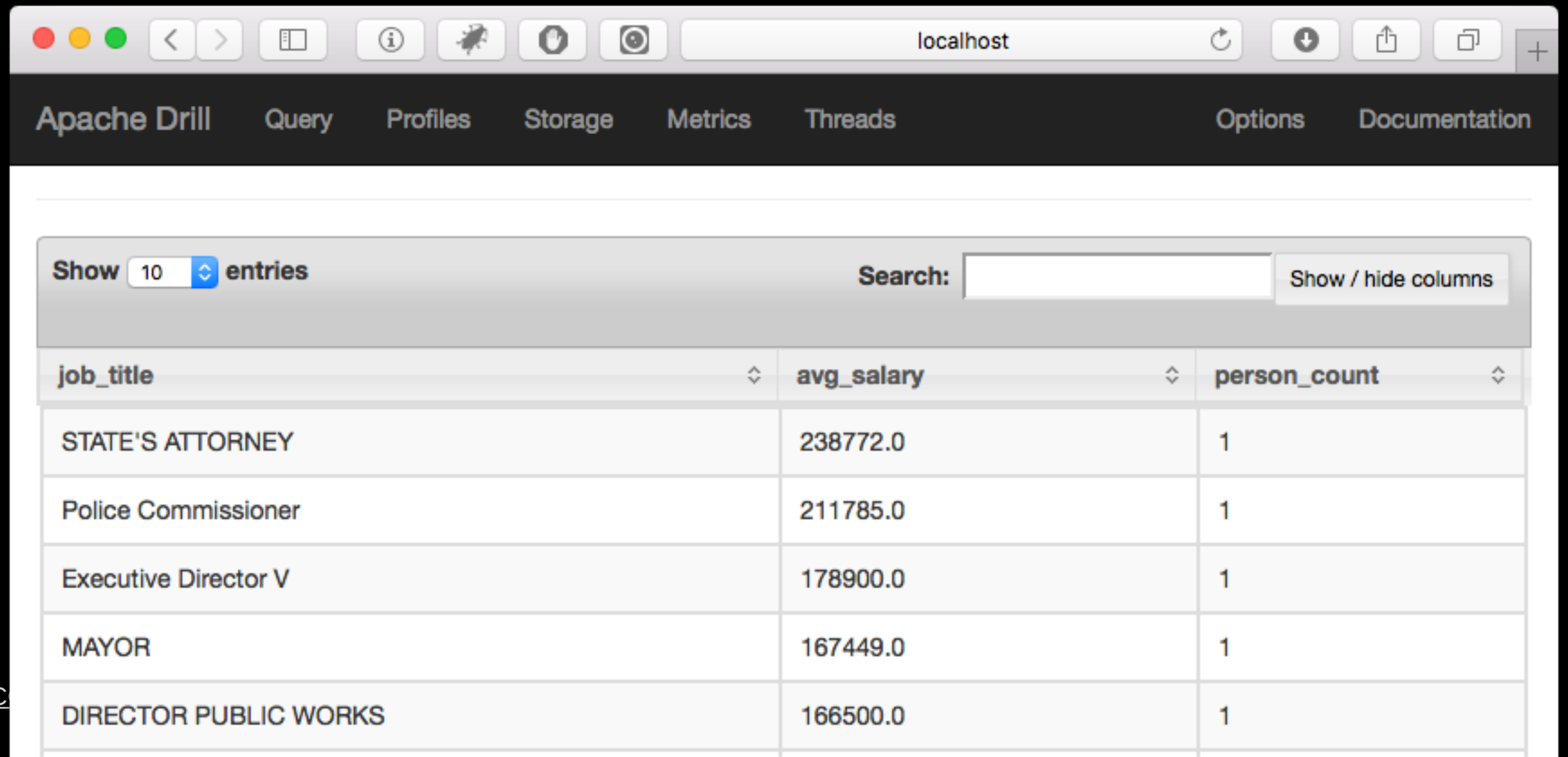
In Class Exercise

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.drillworkshop.`baltimore_salaries.json`
)
GROUP BY raw_data[9]
ORDER BY avg_salary DESC
```

DARK LABS

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



DARK LABS

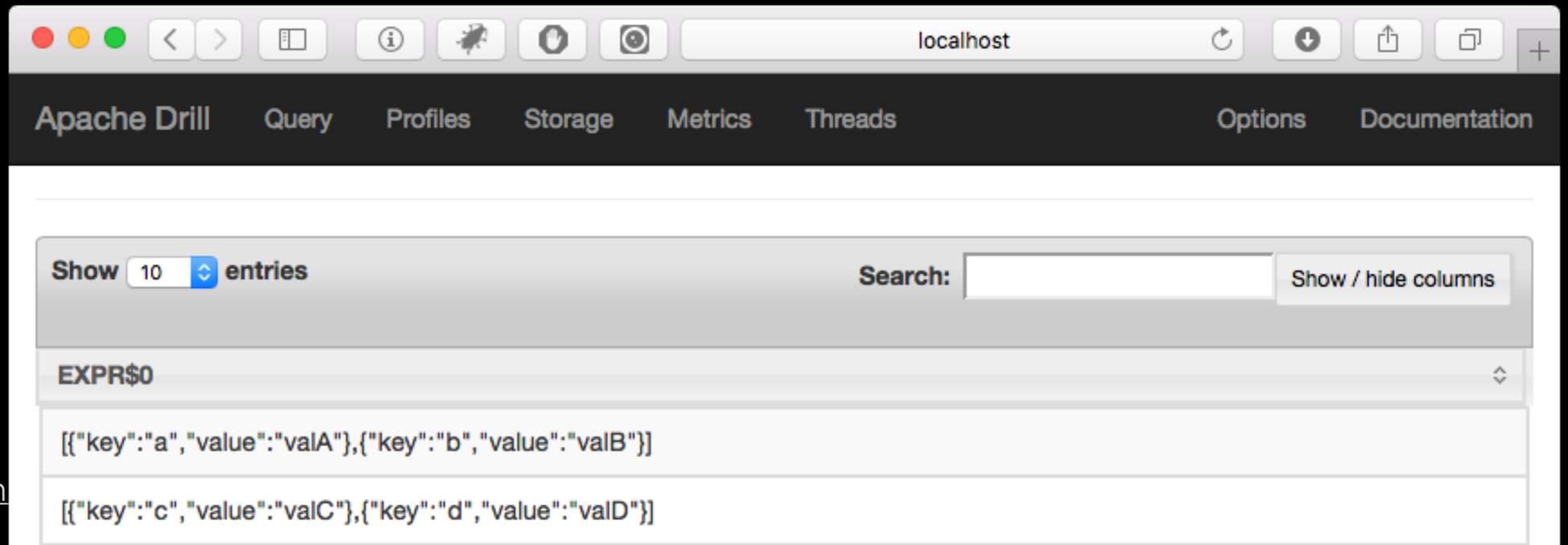
KVGEN(<map>) returns a list of keys and values in a map


```
{"rec1":{"a": "valA", "b": "valB"}}
{"rec1":{"c": "valC", "d": "valD"}}
```

DARK LABS

```
{"rec1":{"a": "valA", "b": "valB"}}
{"rec1":{"c": "valC", "d": "valD"}}
```

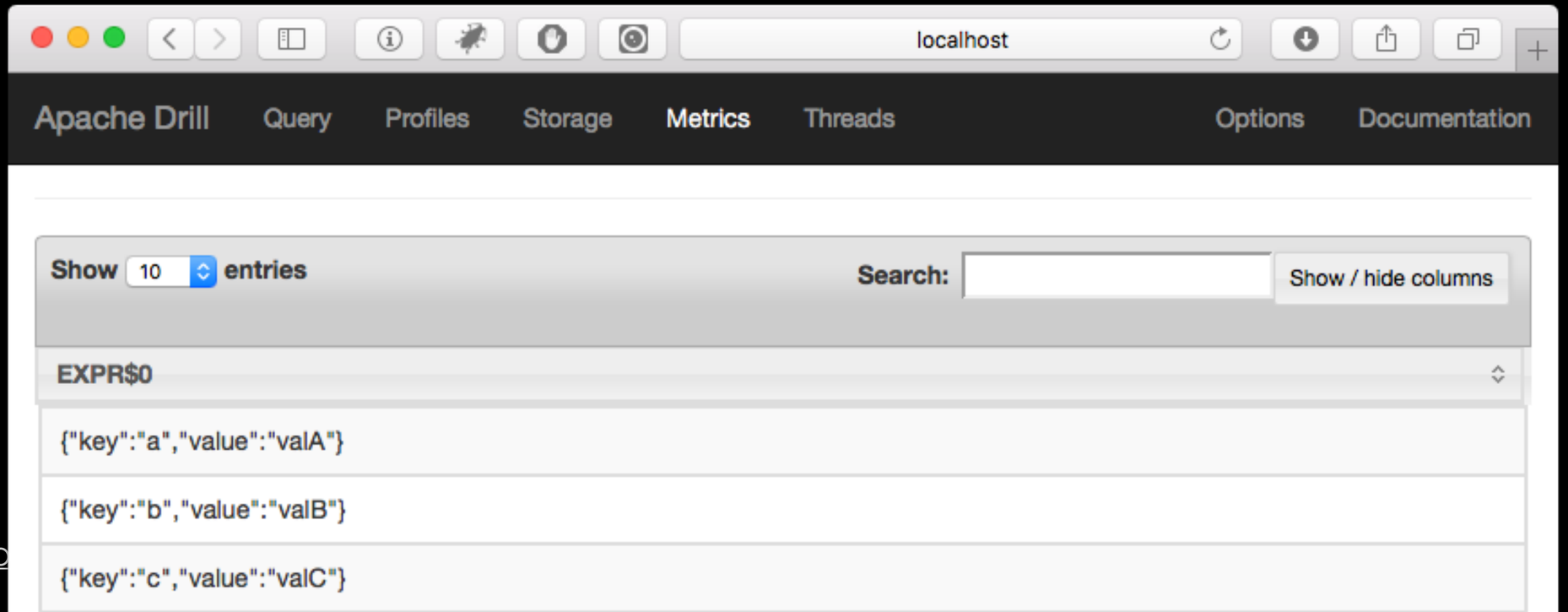
```
SELECT KVGEN( rec1 ) FROM dfs.drillworkshop.`simple.json`
```



DARK LABS

```
{"rec1":{"a": "valA", "b": "valB"}}
{"rec1":{"c": "valC", "d": "valD"}}
```

```
SELECT FLATTEN( KVGEN( rec1 ) )
FROM dfs.drillworkshop.`simple.json`
```



Saving Data

DARK LABS

Saving Data

Drill supports:

- CSV, TSV, PSV
- Parquet (default)
- JSON

Saving Data

```
ALTER SESSION SET `store.format` = '<format>';
```


```
CREATE TABLE <file_name> AS <query>
```


```
CREATE TABLE <file_name> AS <query>
```

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

DARK LABS

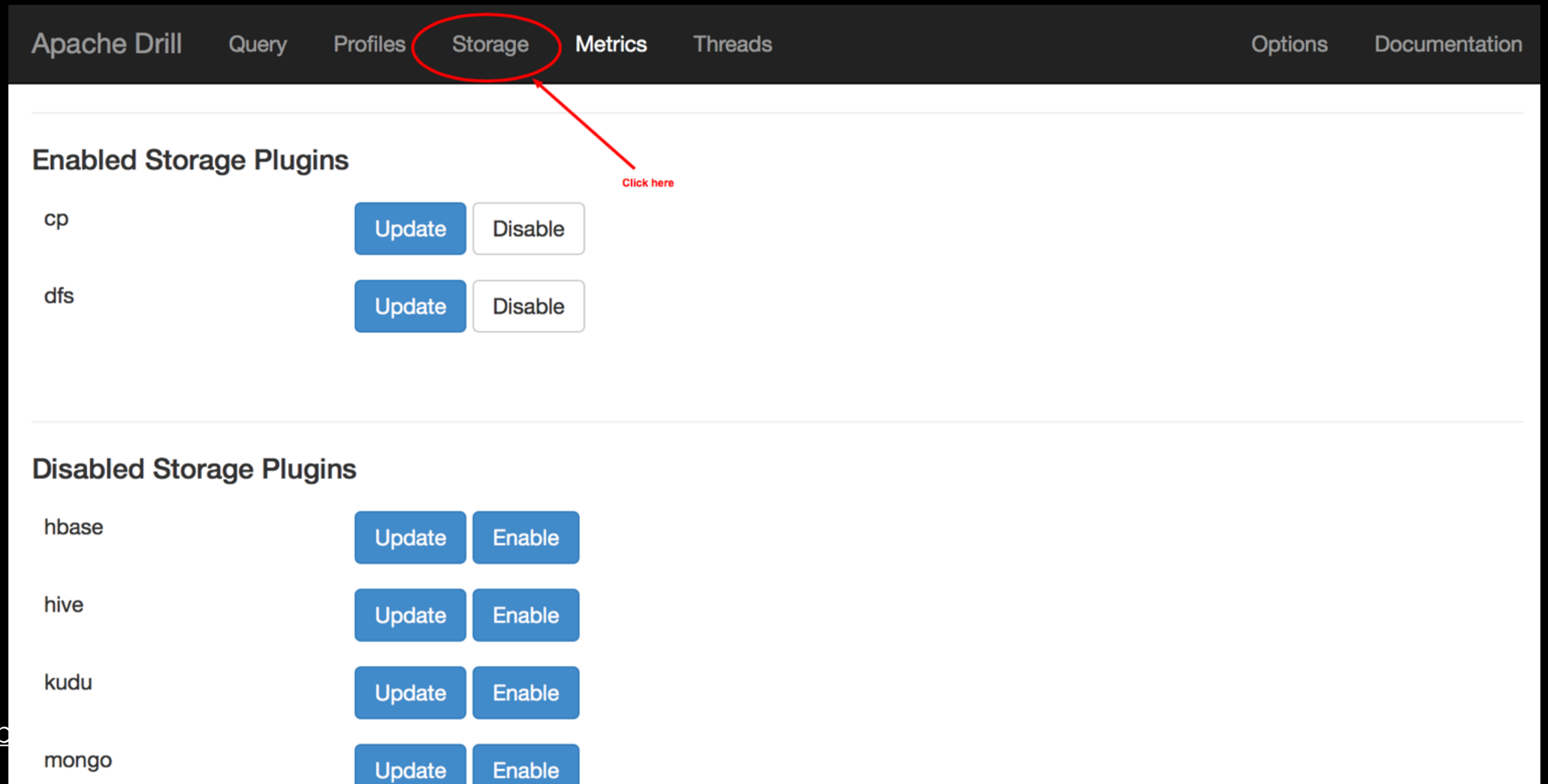
Connecting other Data Sources

DARK LABS

Connecting other Data Sources

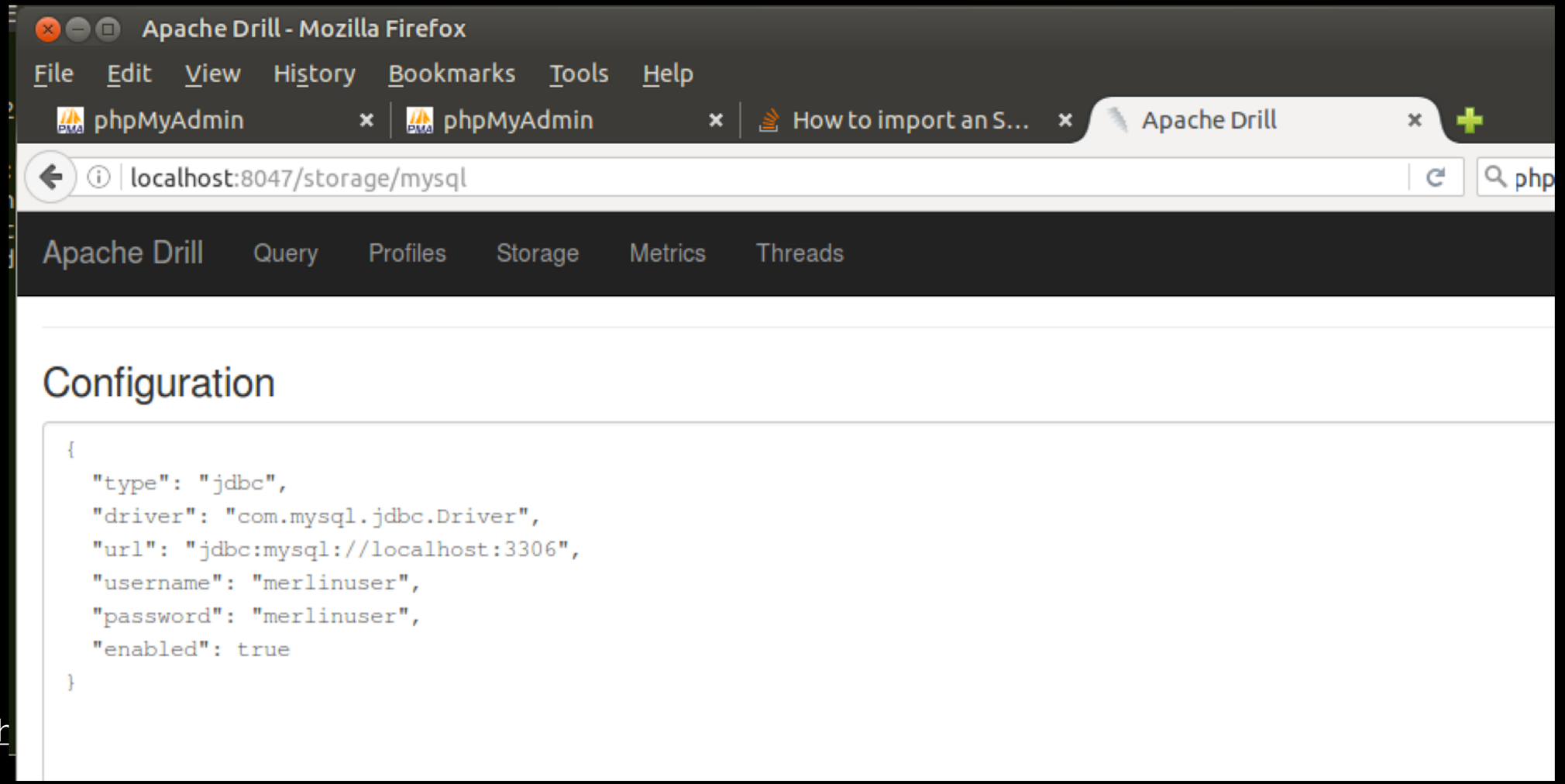
DARK LABS

Connecting other Data Sources



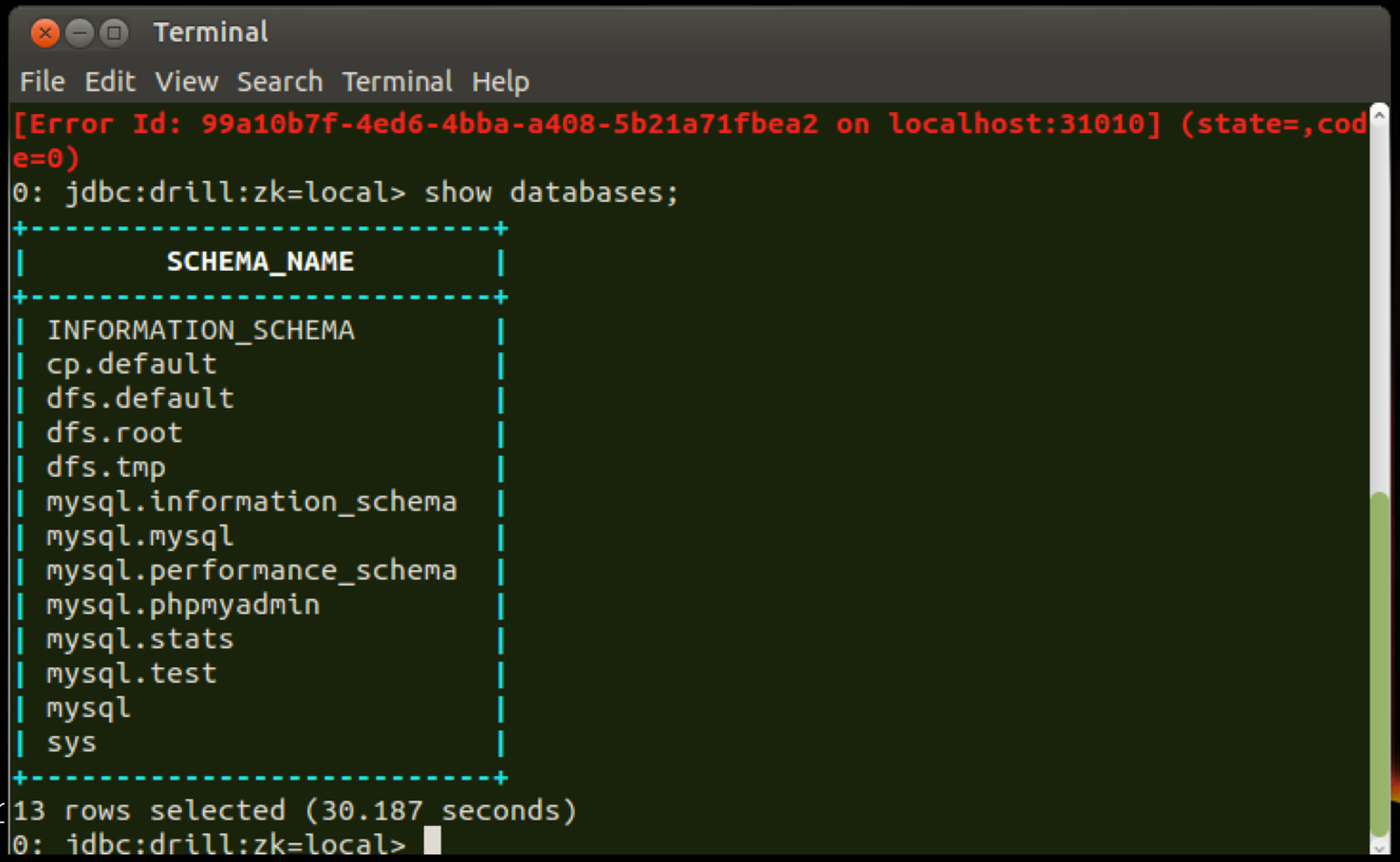
DARK LABS

Connecting other Data Sources



DARK LABS

Connecting other Data Sources



DARK LABS

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
```

DARK LABS

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
```

DARK LABS

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

DARK LABS

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
```

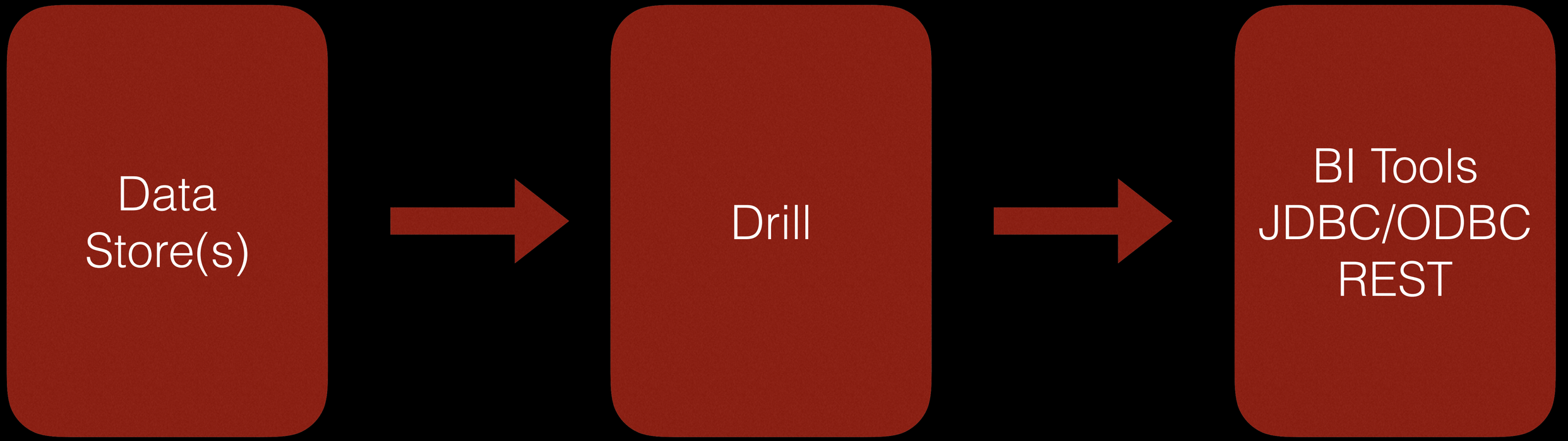
MySQL: 0.047 seconds

Drill: 0.366 seconds

Connecting to Drill

DARK LABS

Connecting to Drill



DARK LABS

Connecting to Drill

```
pip install pydrill
```

Connecting to Drill

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

DARK LABS

Connecting to Drill

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

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

DARK LABS

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.drillworkshop.`*.csvh`
GROUP BY JobTitle
Order By avg_salary DESC
LIMIT 10
''')
```

DARK LABS

Connecting to Drill

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

Questions?

DARK LABS

Thank you!

Charles Givre
@cgivre
givre_charles@bah.com
thedataist.com