

Apache Drill Workshop

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





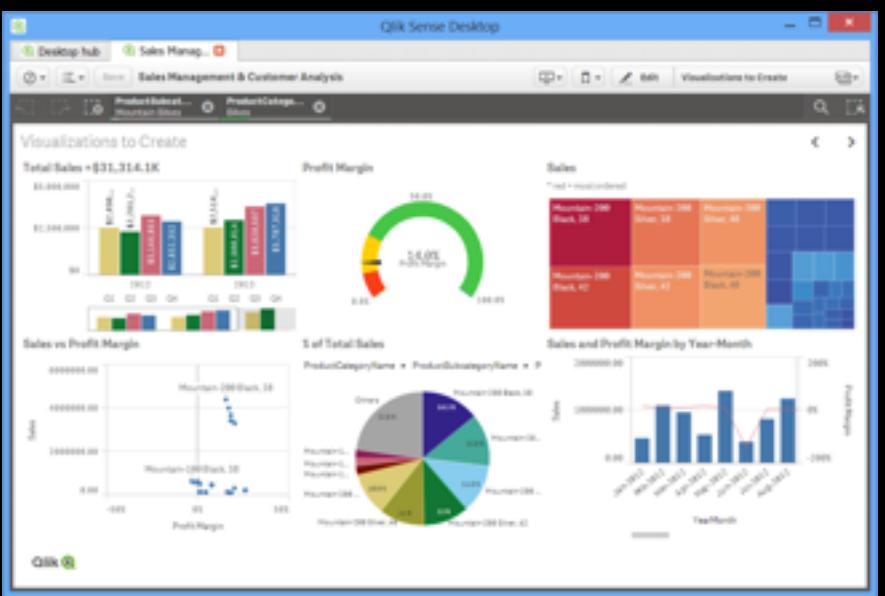
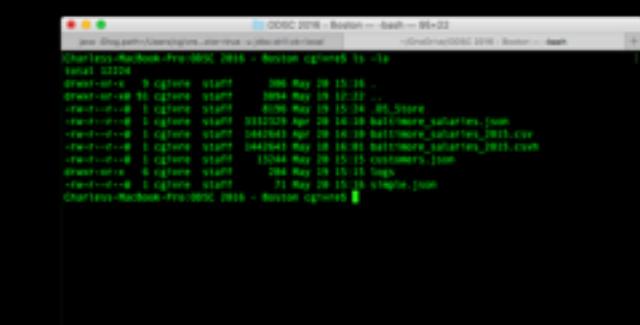
What is Drill?



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

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





ETL





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



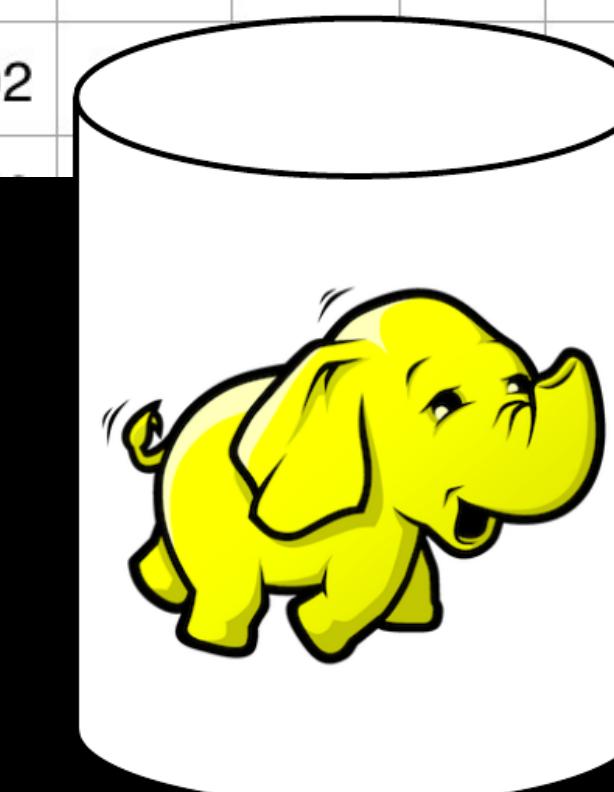
Quick Demo

Thank you Jair Aguirre!!

Quick Demo

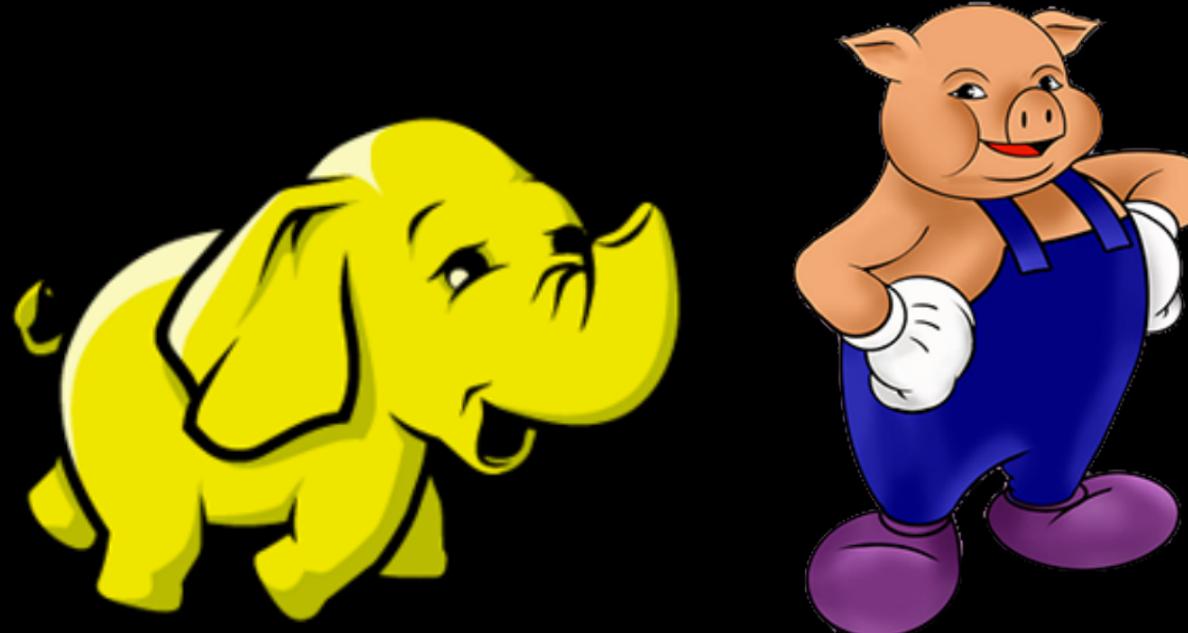
| yearID | IgID | teamID | franchID | divID | Rank | G | Ghome | W | L | DivWin | WCWin | LgWin | WSWin | R | AB | H | 2B | 3B | HR | E |
|--------|------|--------|----------|-------|------|----|-------|----|----|--------|-------|-------|-------|-----|------|-----|----|----|----|---|
| 1871 | NA | BS1 | 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 | |

mlahman.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;
```



Loading... Please Wait



Execution Time:
1 minute, 38 seconds

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:
0.89 seconds!!



NoSQL, No Problem



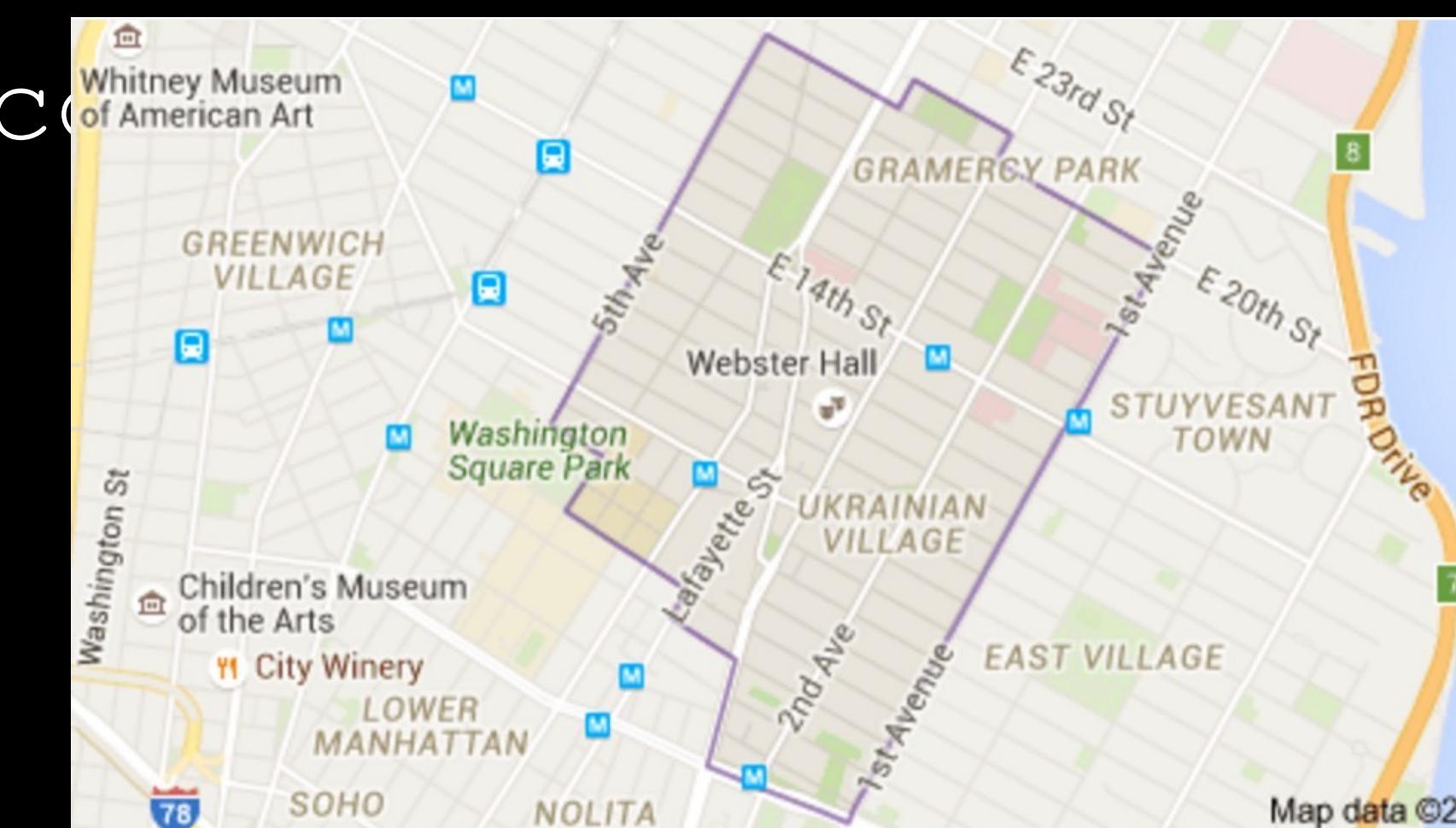
NoSQL, No Problem

```
{  
  "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 },  

```

NoSQL, No Problem

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



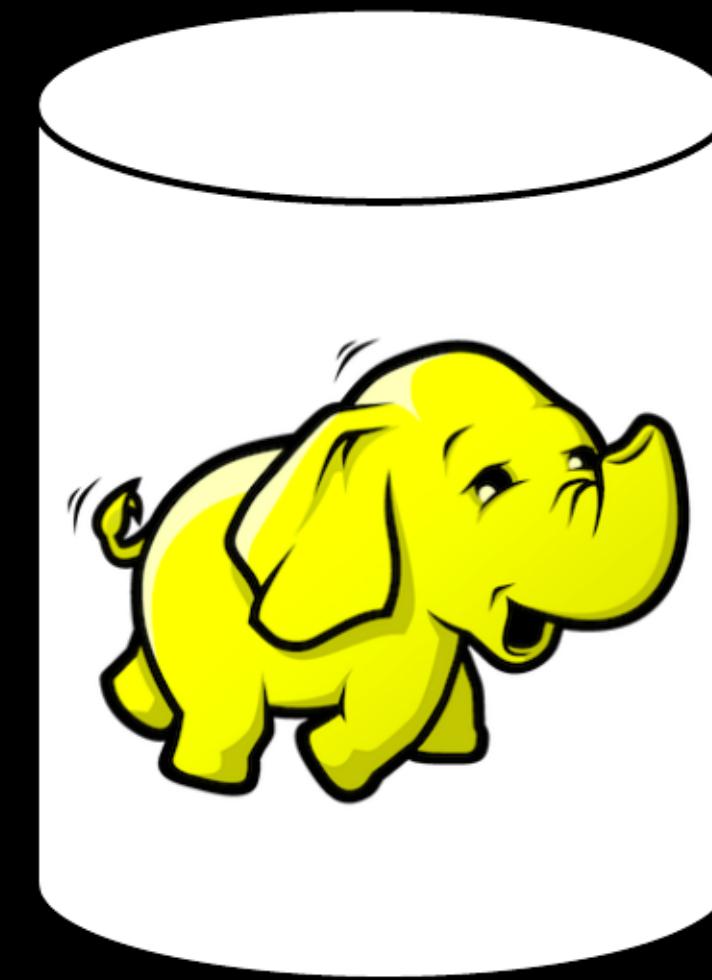
New York, NY 10003

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



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

WARNING
DANGER
DO NOT PULL
HANDLE

MARTIN-BAKER MARK 10A
EJECTION SEAT CAPACITY
8000' MIN. ON ROLLBACK





Installing Drill



Installing Drill

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



Starting Drill

Starting Drill

Embedded Mode: For use on a standalone system

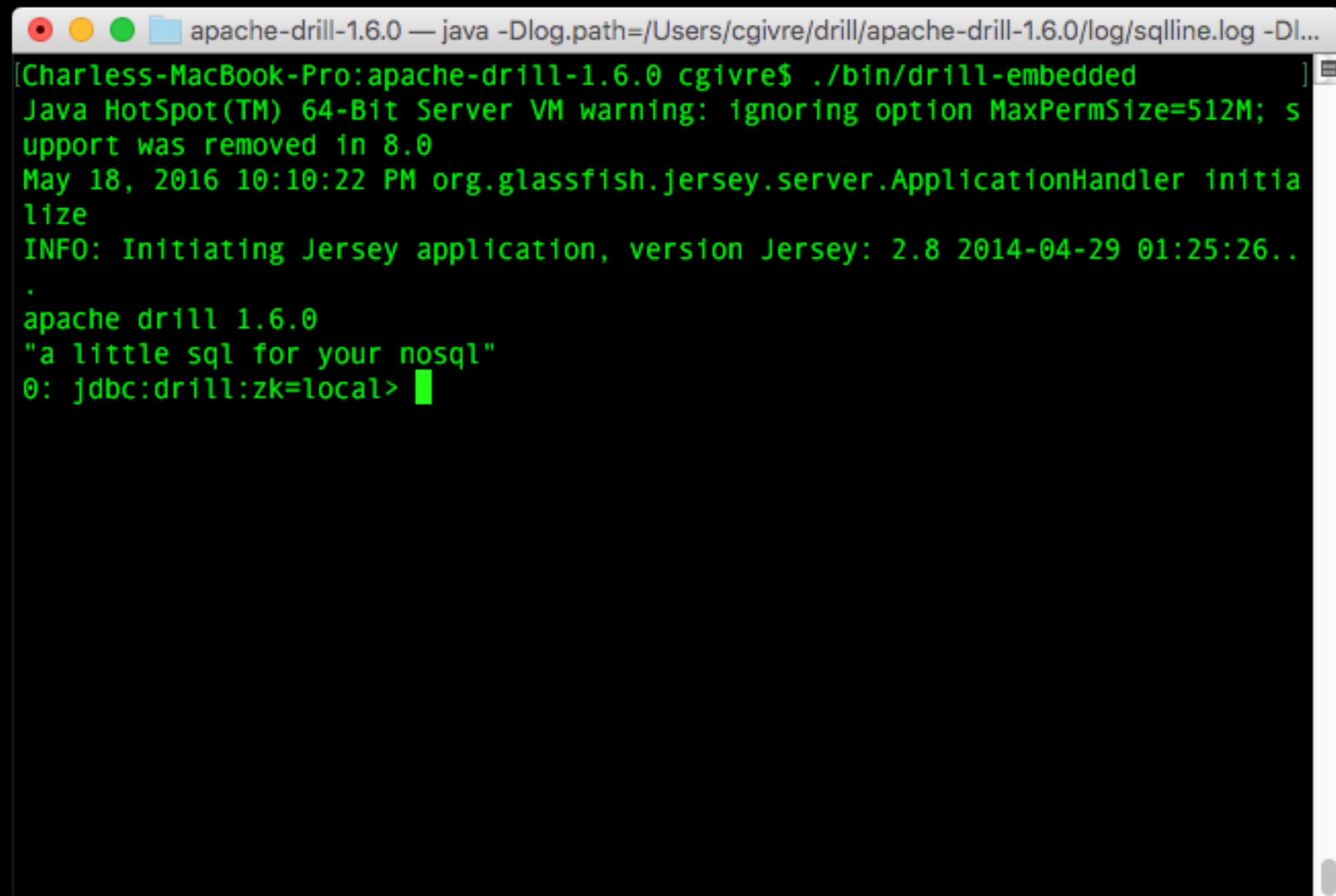
```
$ ./bin/drill-embedded
```



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



Querying Drill

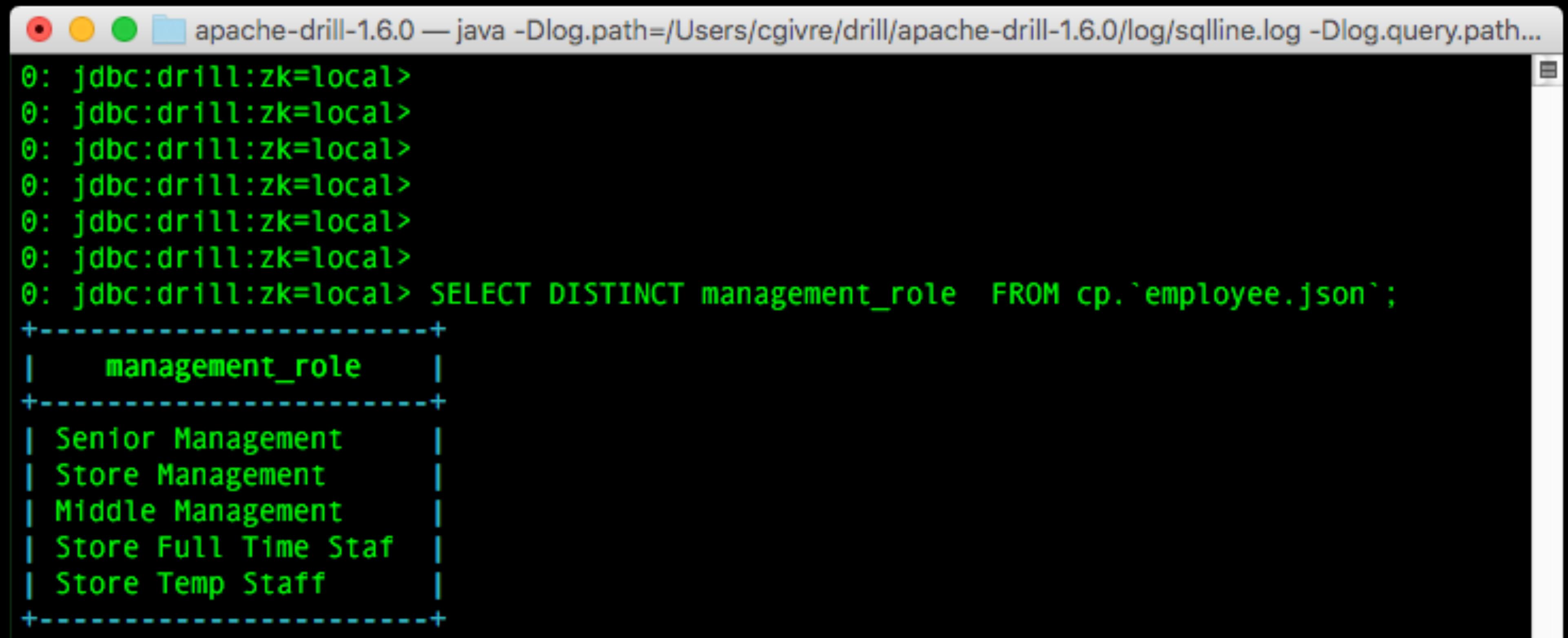


A screenshot of a terminal window on a Mac OS X system. The title bar reads "apache-drill-1.6.0 — java -Dlog.path=/Users/cgivre/drill/apache-drill-1.6.0/log/sqlline.log -Dl...". The terminal output shows the following:

```
[Charless-MacBook-Pro:apache-drill-1.6.0 cgivre$ ./bin/drill-embedded
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M; support was removed in 8.0
May 18, 2016 10:10:22 PM org.glassfish.jersey.server.ApplicationHandler initialize
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> █
```

Querying Drill

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

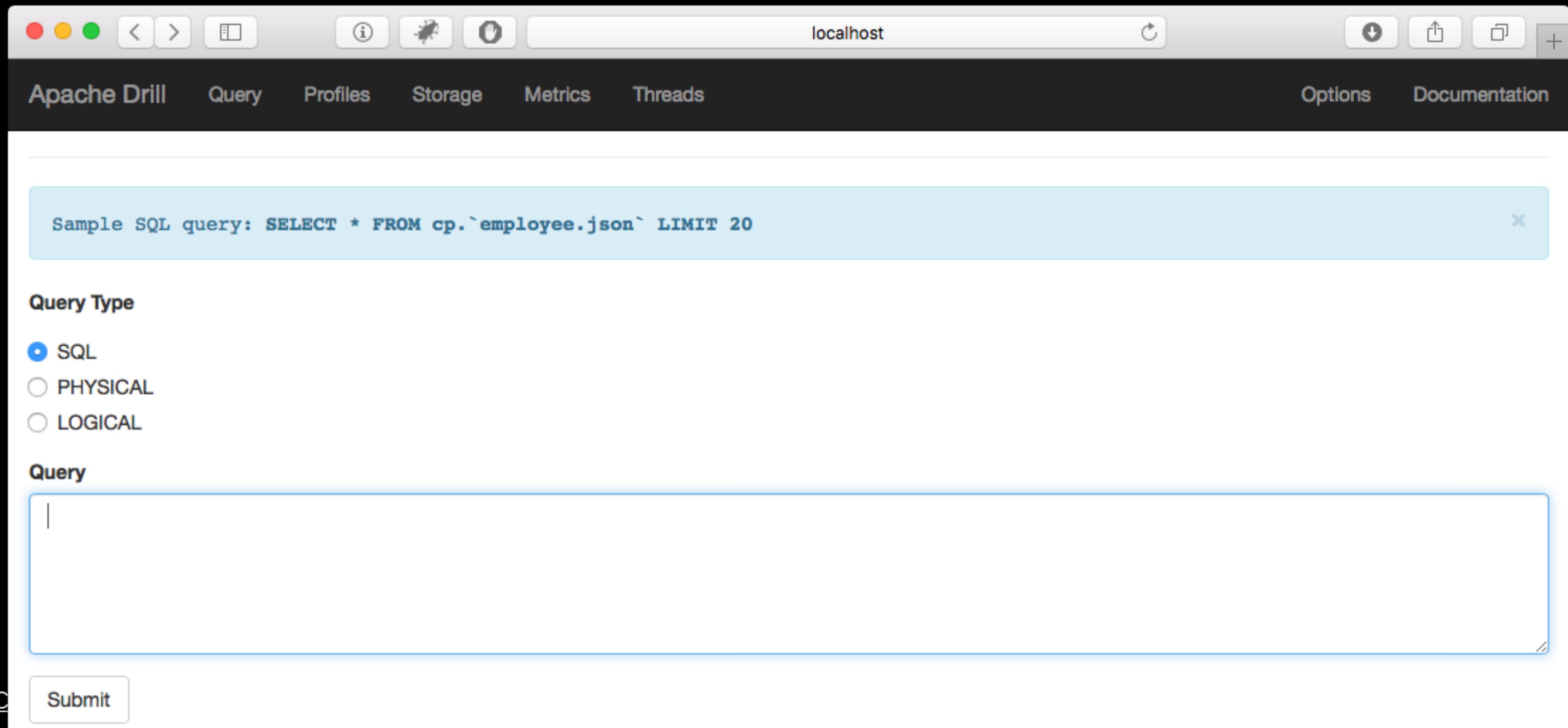


The screenshot shows a terminal window titled "apache-drill-1.6.0 — java -Dlog.path=/Users/cgivre/drill/apache-drill-1.6.0/log/sqlline.log -Dlog.query.path...". The window displays the execution of the following SQL query:

```
0: jdbc:drill:zk=local>
0: jdbc:drill:zk=local>
0: jdbc:drill:zk=local>
0: jdbc:drill:zk=local>
0: jdbc:drill:zk=local>
0: jdbc:drill:zk=local>
0: jdbc:drill:zk=local> SELECT DISTINCT management_role  FROM cp.`employee.json`;
+-----+
|   management_role   |
+-----+
| Senior Management |
| Store Management   |
| Middle Management |
| Store Full Time Sta|
| Store Temp Staff   |
+-----+
```

Querying Drill

<http://localhost:8047>



The screenshot shows the Apache Drill web interface running locally. The browser window has a title bar with the URL `localhost`. The main navigation bar includes links for `Apache Drill`, `Query`, `Profiles`, `Storage`, `Metrics`, `Threads`, `Options`, and `Documentation`.

A blue callout box displays a sample SQL query:

```
Sample SQL query: SELECT * FROM cp.`employee.json` LIMIT 20
```

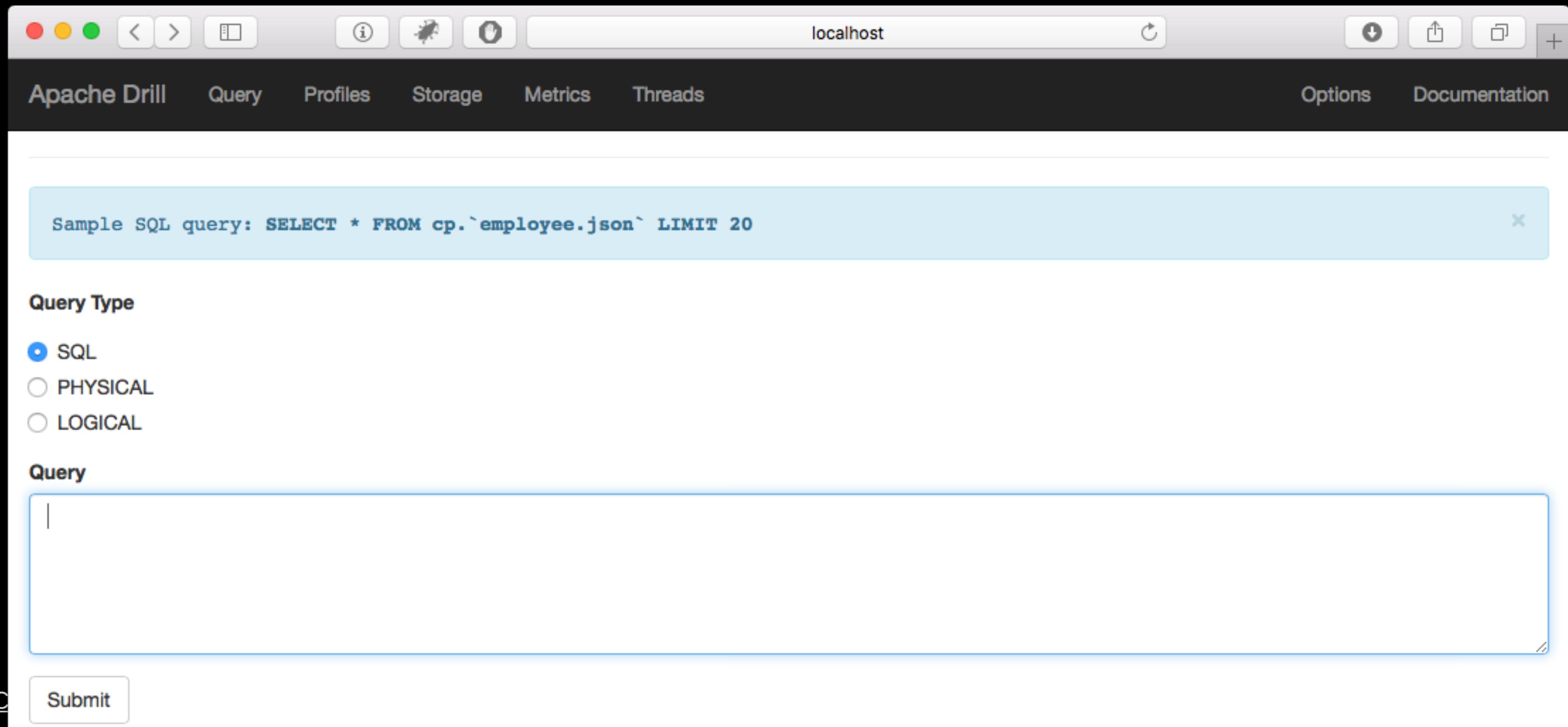
The interface includes a section for **Query Type** with three radio button options: `SQL` (selected), `PHYSICAL`, and `LOGICAL`.

The **Query** section contains a large text input field where a user can type their SQL query. A small vertical cursor is visible in the input field.

At the bottom left of the interface, there is a **Submit** button.

Querying Drill

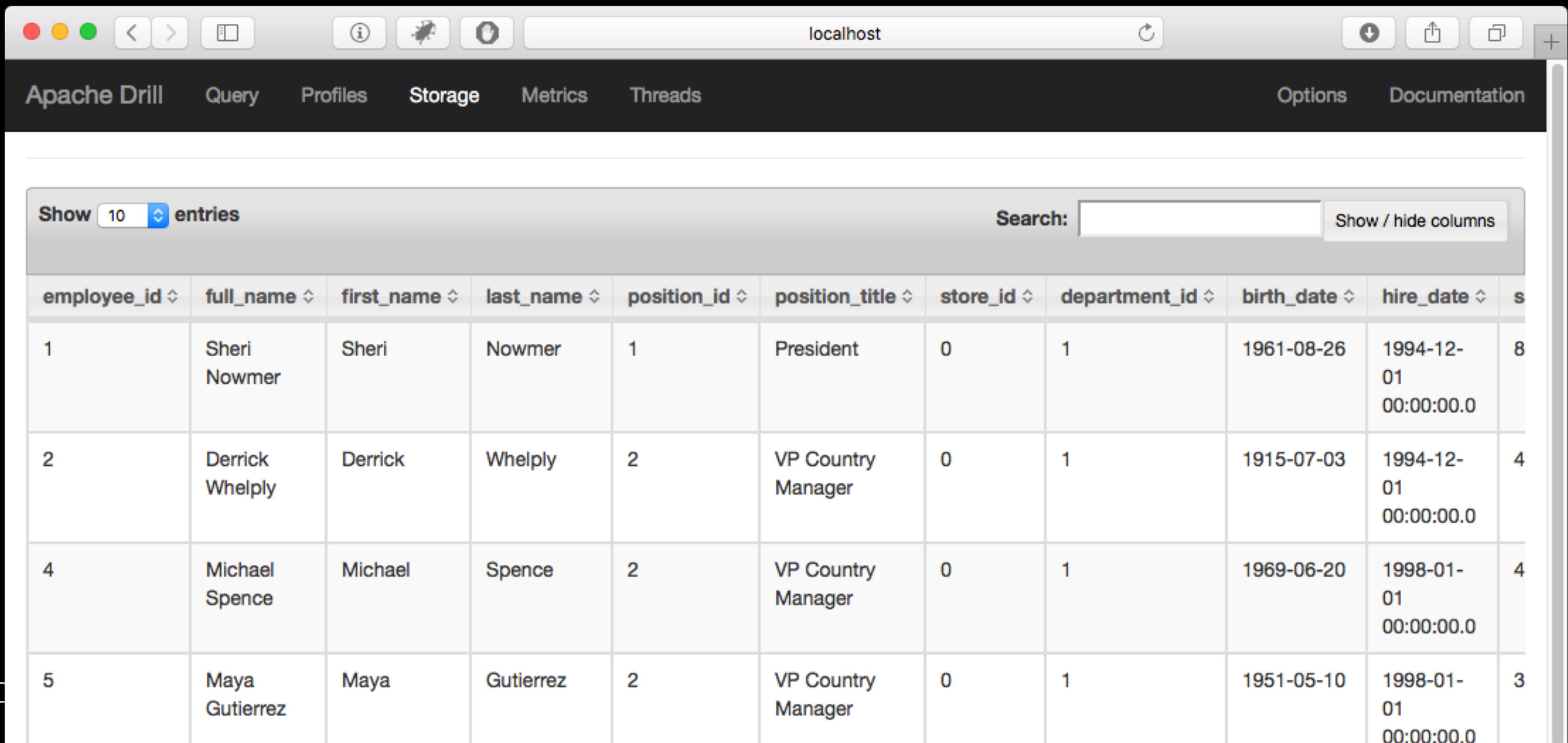
```
SELECT * FROM cp.`employee.json` LIMIT 20
```



The screenshot shows the Apache Drill web interface running on localhost. The top navigation bar includes links for Apache Drill, Query, Profiles, Storage, Metrics, Threads, Options, and Documentation. A sample SQL query is displayed in a blue box: `Sample SQL query: SELECT * FROM cp.`employee.json` LIMIT 20`. Below this, a "Query Type" section has "SQL" selected. The main "Query" input field contains the previously shown SQL statement. A "Submit" button is located at the bottom left of the query area.

Querying Drill

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



The screenshot shows the Apache Drill web interface running on localhost. The top navigation bar includes links for Apache Drill, Query, Profiles, Storage, Metrics, Threads, Options, and Documentation. Below the navigation is a search and filter panel with "Show 10 entries" and a "Search:" input field. The main area displays a table of employee data with the following columns: employee_id, full_name, first_name, last_name, position_id, position_title, store_id, department_id, birth_date, hire_date, and salary. The table contains four rows of data, corresponding to the results of the query.

| employee_id | full_name | first_name | last_name | position_id | position_title | store_id | department_id | birth_date | hire_date | salary |
|-------------|----------------|------------|-----------|-------------|--------------------|----------|---------------|------------|------------|----------|
| 1 | Sheri Nowmer | Sheri | Nowmer | 1 | President | 0 | 1 | 1961-08-26 | 1994-12-01 | 80000.00 |
| 2 | Derrick Whelby | Derrick | Whelby | 2 | VP Country Manager | 0 | 1 | 1915-07-03 | 1994-12-01 | 40000.00 |
| 4 | Michael Spence | Michael | Spence | 2 | VP Country Manager | 0 | 1 | 1969-06-20 | 1998-01-01 | 40000.00 |
| 5 | Maya Gutierrez | Maya | Gutierrez | 2 | VP Country Manager | 0 | 1 | 1951-05-10 | 1998-01-01 | 30000.00 |



Querying Drill

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



Querying Drill

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

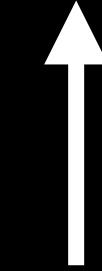


Querying Drill

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

Querying Drill

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



Storage Plugin



Workspace



Table

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 |

Querying Drill

Apache Drill Query Profiles **Storage** Metrics Threads Options Documentation

[Click here to go to view Storage Plugins](#)

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 |

New Storage Plugin

| | |
|---|------------------------|
| <input type="text" value="Storage Name"/> | Create |
|---|------------------------|

Querying Drill

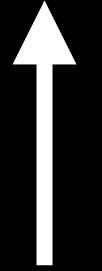
```
FROM dfs.logs.`/data/customers.csv`
```



Storage Plugin



Workspace



Table

Querying Drill

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



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



In Class Exercise: Create a Workspace

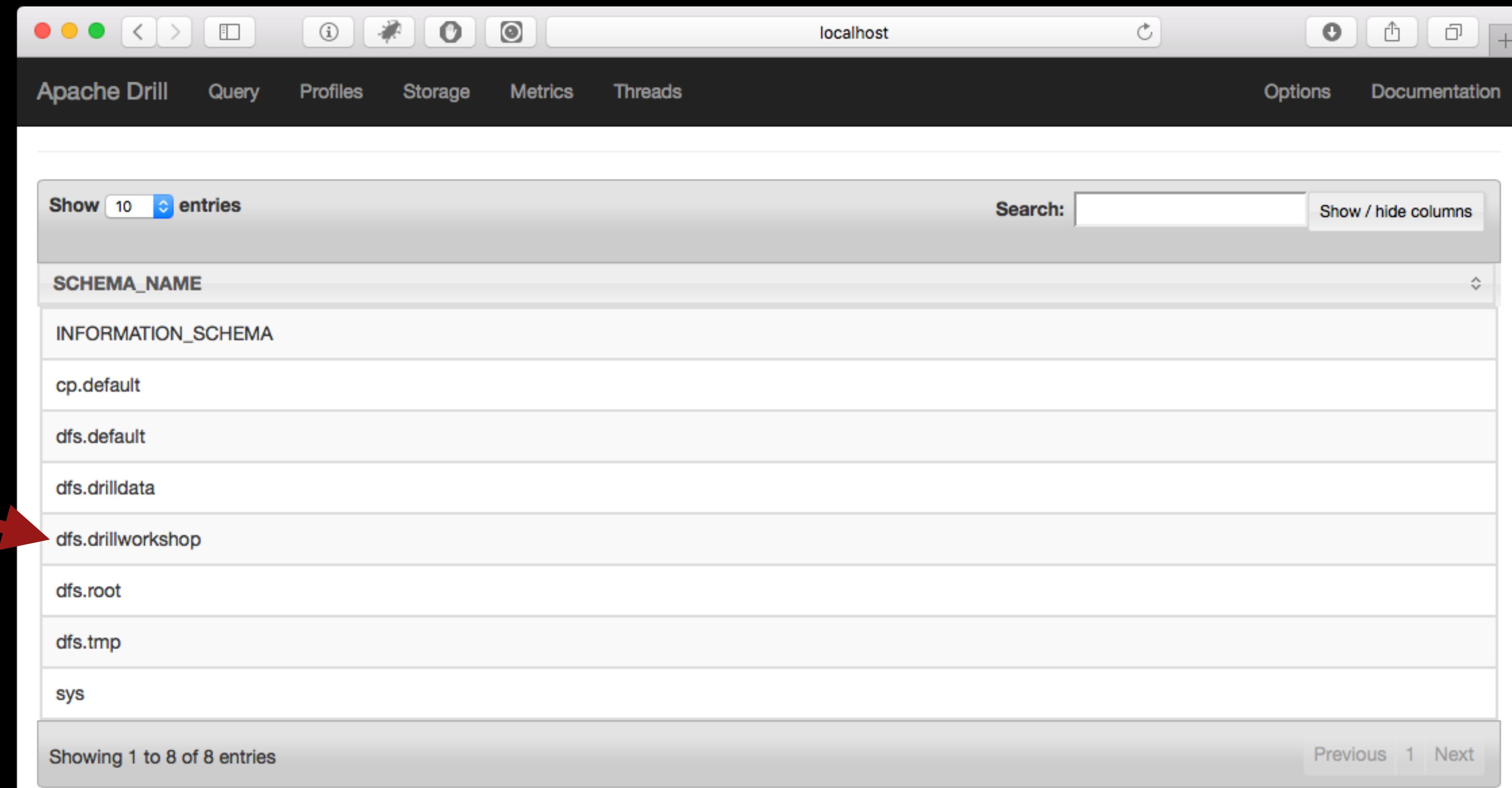
In this exercise we are going to create a workspace called 'drillworkshop', which we will use for future exercises.

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  
}
```

Querying Drill

```
SHOW databases;
```

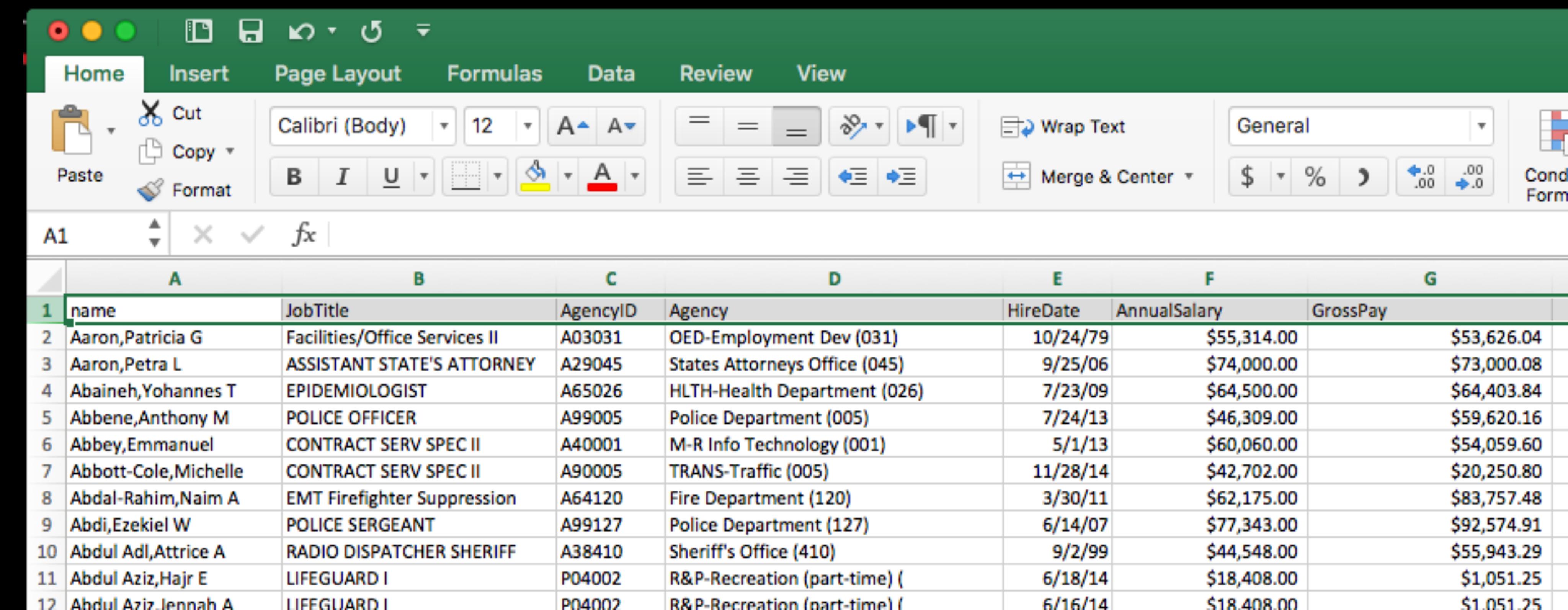


A screenshot of the Apache Drill web interface. The title bar shows "localhost" and the menu bar includes "Apache Drill", "Query", "Profiles", "Storage", "Metrics", "Threads", "Options", and "Documentation". The main content area displays a table with one column labeled "SCHEMA_NAME". The table lists eight entries: INFORMATION_SCHEMA, cp.default, dfs.default, dfs.drilldata, dfs.drillworkshop, dfs.root, dfs.tmp, and sys. A red arrow points from the word "Success!!" on the left to the "dfs.drillworkshop" entry in the table.

| SCHEMA_NAME |
|--------------------|
| INFORMATION_SCHEMA |
| cp.default |
| dfs.default |
| dfs.drilldata |
| dfs.drillworkshop |
| dfs.root |
| dfs.tmp |
| sys |

Showing 1 to 8 of 8 entries

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 |

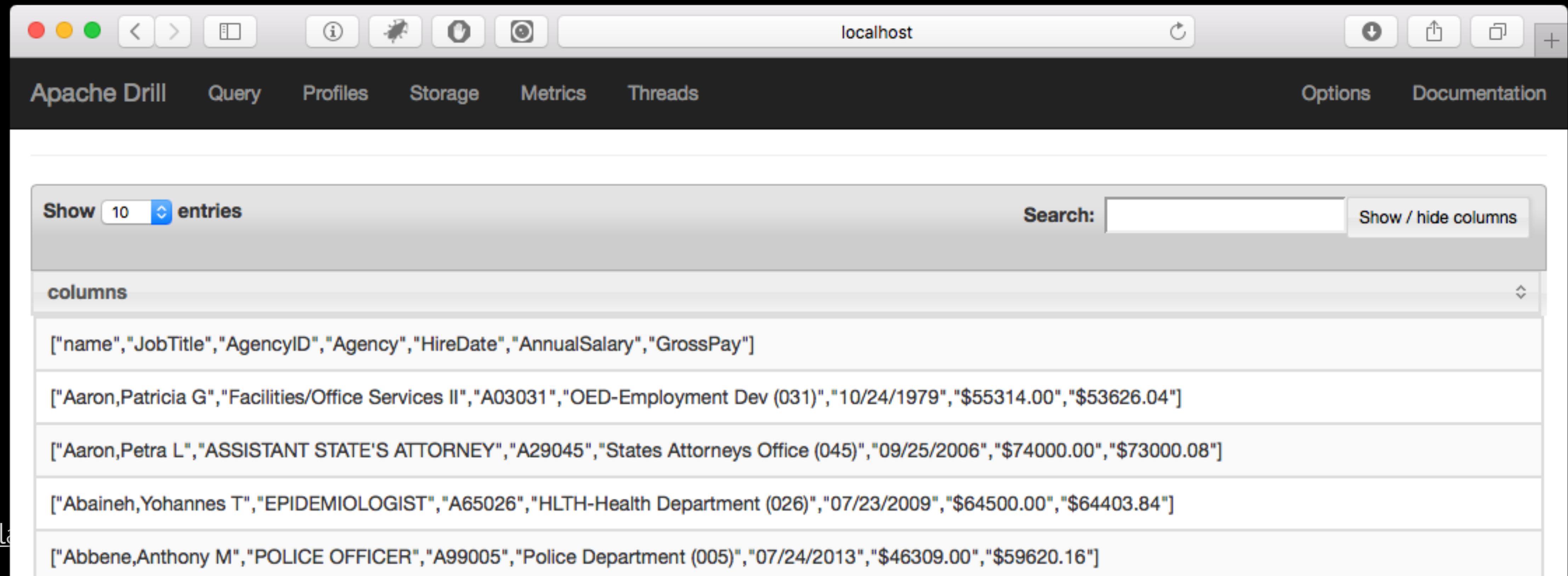


Querying Drill

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

Querying Drill

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



The screenshot shows the Apache Drill web interface running on localhost. The top navigation bar includes links for Apache Drill, Query, Profiles, Storage, Metrics, Threads, Options, and Documentation. Below the navigation is a search bar with 'Search:' and a 'Show / hide columns' button. On the left, there's a 'columns' section with a dropdown set to '10 entries'. The main content area displays five rows of data from the 'baltimore_salaries_2015.csv' file.

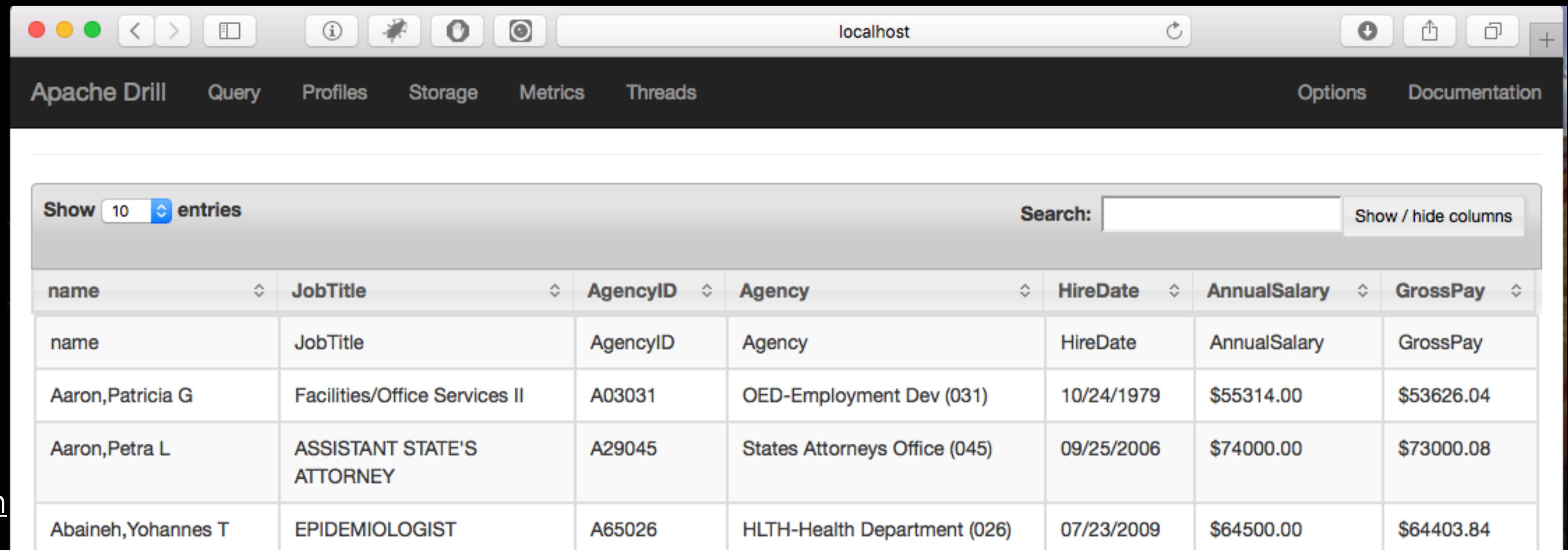
| name | JobTitle | AgencyID | Agency | HireDate | AnnualSalary | GrossPay |
|--------------------|-------------------------------|----------|-------------------------------|------------|--------------|------------|
| Aaron,Patricia G | Facilities/Office Services II | A03031 | OED-Employment Dev (031) | 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 |
| Abbene,Anthony M | POLICE OFFICER | A99005 | Police Department (005) | 07/24/2013 | \$46309.00 | \$59620.16 |

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
```

Querying Drill

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

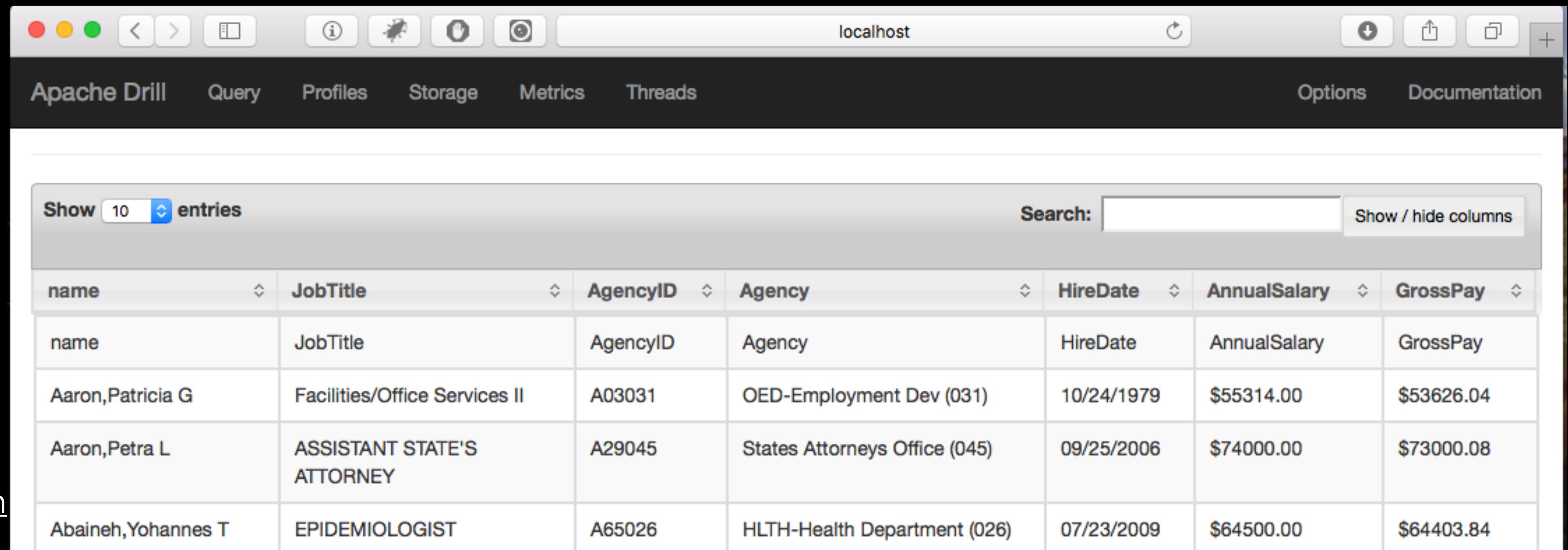


The screenshot shows the Apache Drill web interface running on localhost. The top navigation bar includes links for Apache Drill, Query, Profiles, Storage, Metrics, Threads, Options, and Documentation. Below the header is a search bar with dropdowns for 'Show' (set to 10) and 'entries', and a 'Search:' input field. A 'Show / hide columns' button is also present. The main content area displays a table with the following data:

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



The screenshot shows the Apache Drill web interface running on localhost. The top navigation bar includes links for Apache Drill, Query, Profiles, Storage, Metrics, Threads, Options, and Documentation. Below the header is a search bar with dropdowns for 'Show' (set to 10) and 'entries', and a 'Search:' input field. A 'Show / hide columns' button is also present. The main content area displays a table with the following data:

| 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    "delimiter": ", "  
}
```

Querying Drill

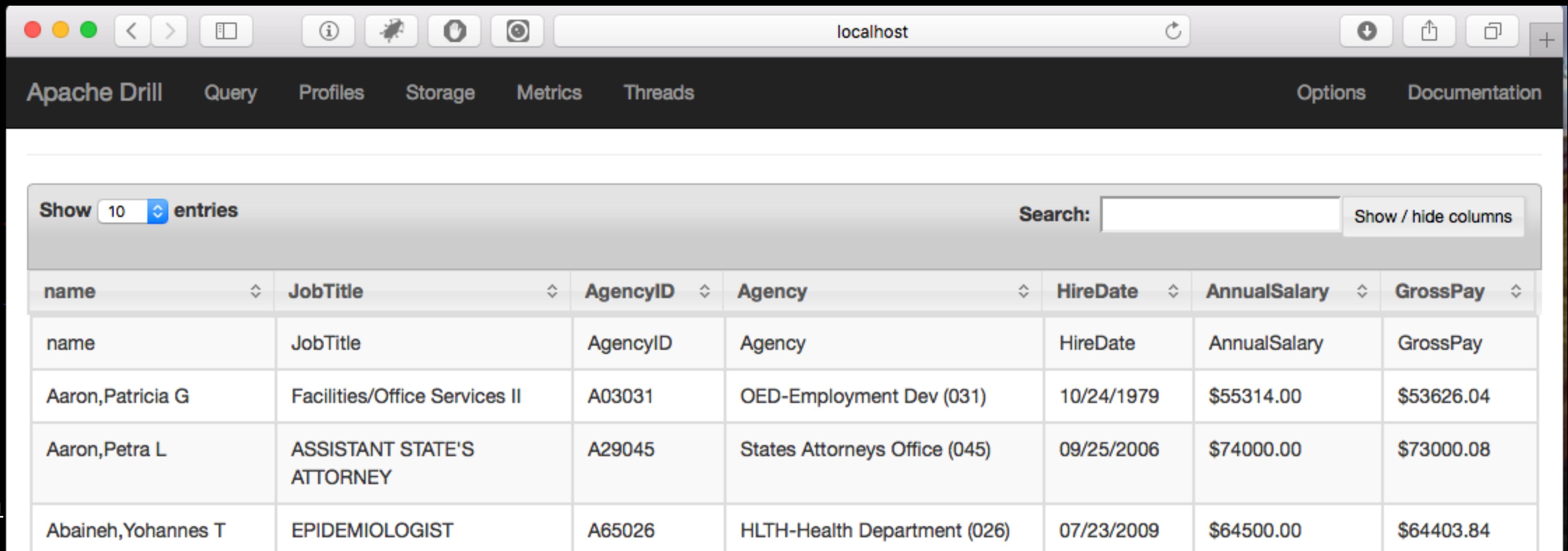
| File Extension | File Type |
|----------------|-----------------------------------|
| .psv | Pipe separated values |
| .csv | Comma separated value files |
| .csvh | Comma separated value with header |
| .tsv | Tab separated values |
| .json | JavaScript Object Notation files |
| .avro | Avro files (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 |

Querying Drill

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



The screenshot shows the Apache Drill web interface running on localhost. The top navigation bar includes links for Apache Drill, Query, Profiles, Storage, Metrics, Threads, Options, and Documentation. Below the navigation is a search bar with dropdowns for 'Show' (set to 10) and 'entries', and a 'Search:' input field. A 'Show / hide columns' button is also present. The main content area displays a table with the following data:

| 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 |



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 dfs.drillworkshop.`*`.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.drillworkshop.`*`.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.drillworkshop.`*.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 |
| 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`
```

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
```



```
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
```

The screenshot shows the Apache Drill web interface running on localhost. The top navigation bar includes links for Apache Drill, Query, Profiles, Storage, Metrics, Threads, Options, and Documentation. The main content area displays a table of query results.

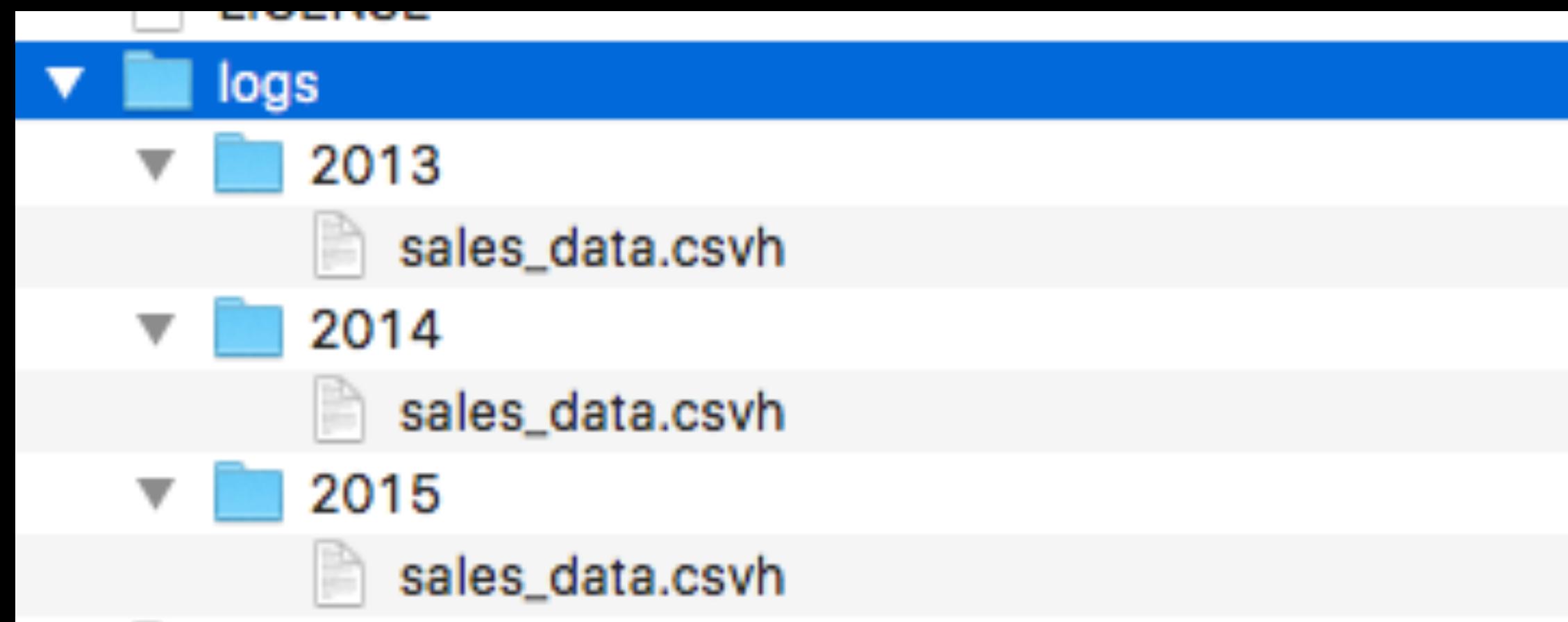
| JobTitle | avg_salary | number |
|-----------------------|------------|--------|
| STATE'S ATTORNEY | 238772.0 | 1 |
| Police Commissioner | 211785.0 | 1 |
| Executive Director V | 178900.0 | 1 |
| MAYOR | 167449.0 | 1 |
| DIRECTOR PUBLIC WORKS | 166500.0 | 1 |



Problem: You have multiple log files
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.drillworkshop.`logs/`
LIMIT 10
```



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

The screenshot shows the Apache Drill interface running on localhost. The top navigation bar includes tabs for Apache Drill, Query, Profiles, Storage, Metrics, Threads, Options, and Documentation. Below the navigation is a search bar with 'Show 10 entries' and a 'Search:' field. A 'Show / hide columns' button is also present. The main area displays a table with the following data:

| 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

`dirn` 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

```
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

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"  
  }  
]
```

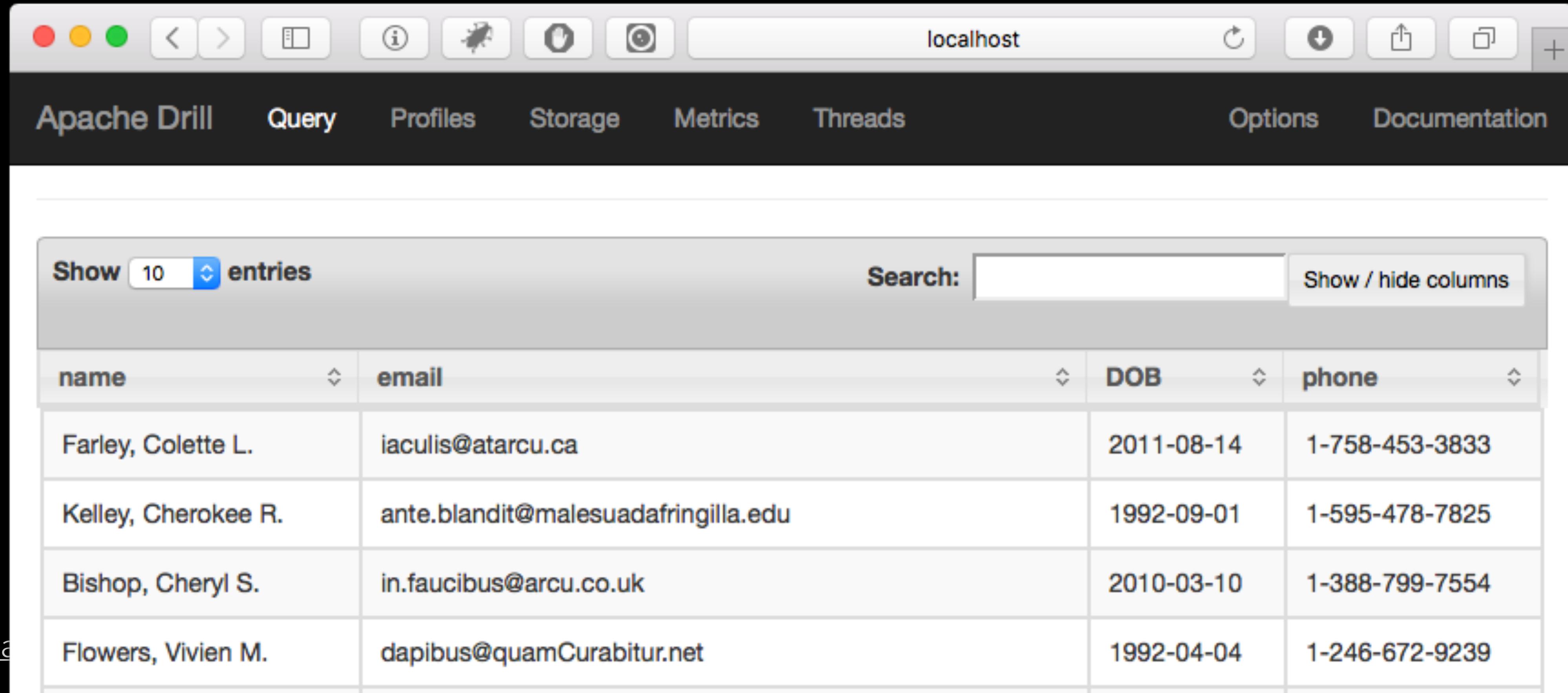


Let's look at JSON data

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

Let's look at JSON data

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



The screenshot shows the Apache Drill web interface running on localhost. The top navigation bar includes links for Apache Drill, Query, Profiles, Storage, Metrics, Threads, Options, and Documentation. Below the navigation is a search and filter section with "Show 10 entries" and a "Search:" input field. The main content area displays a table with four columns: name, email, DOB, and phone. The table contains four rows of customer data.

| name | email | DOB | phone |
|---------------------|-------------------------------------|------------|----------------|
| Farley, Colette L. | iaculis@atarcu.ca | 2011-08-14 | 1-758-453-3833 |
| Kelley, Cherokee R. | ante.blandit@malesuadafringilla.edu | 1992-09-01 | 1-595-478-7825 |
| Bishop, Cheryl S. | in.faucibus@arcu.co.uk | 2010-03-10 | 1-388-799-7554 |
| Flowers, Vivien M. | dapibus@quamCurabitur.net | 1992-04-04 | 1-246-672-9239 |

Let's look at JSON data

```
SELECT *\nFROM dfs.drillworkshop.`customers.json`
```





What about nested data?



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" ]
```



```
"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"  
]
```



Drill has a series of functions
for nested data



Please run

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

in Drill



Let's look at this data in Drill

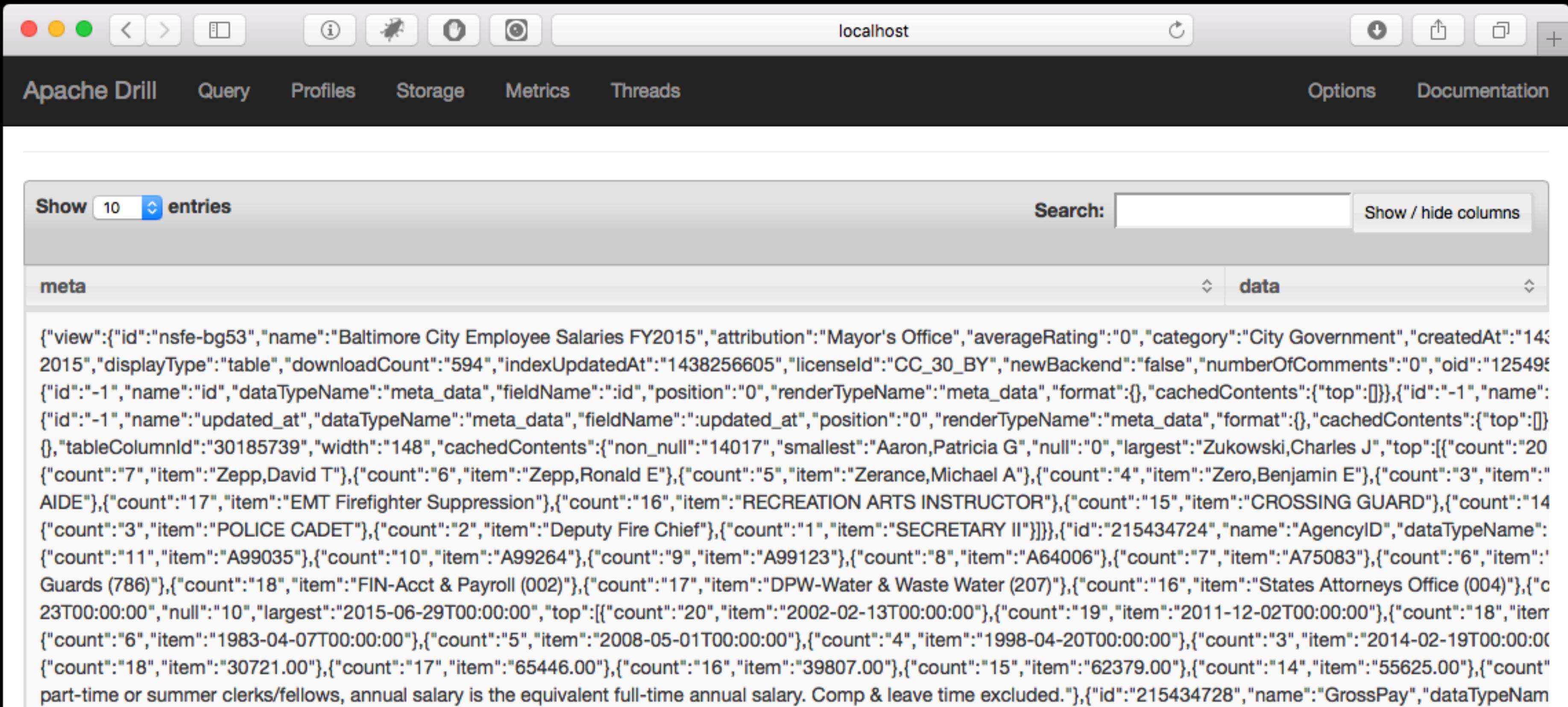


Let's look at this data in Drill

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

Let's look at this data in Drill

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

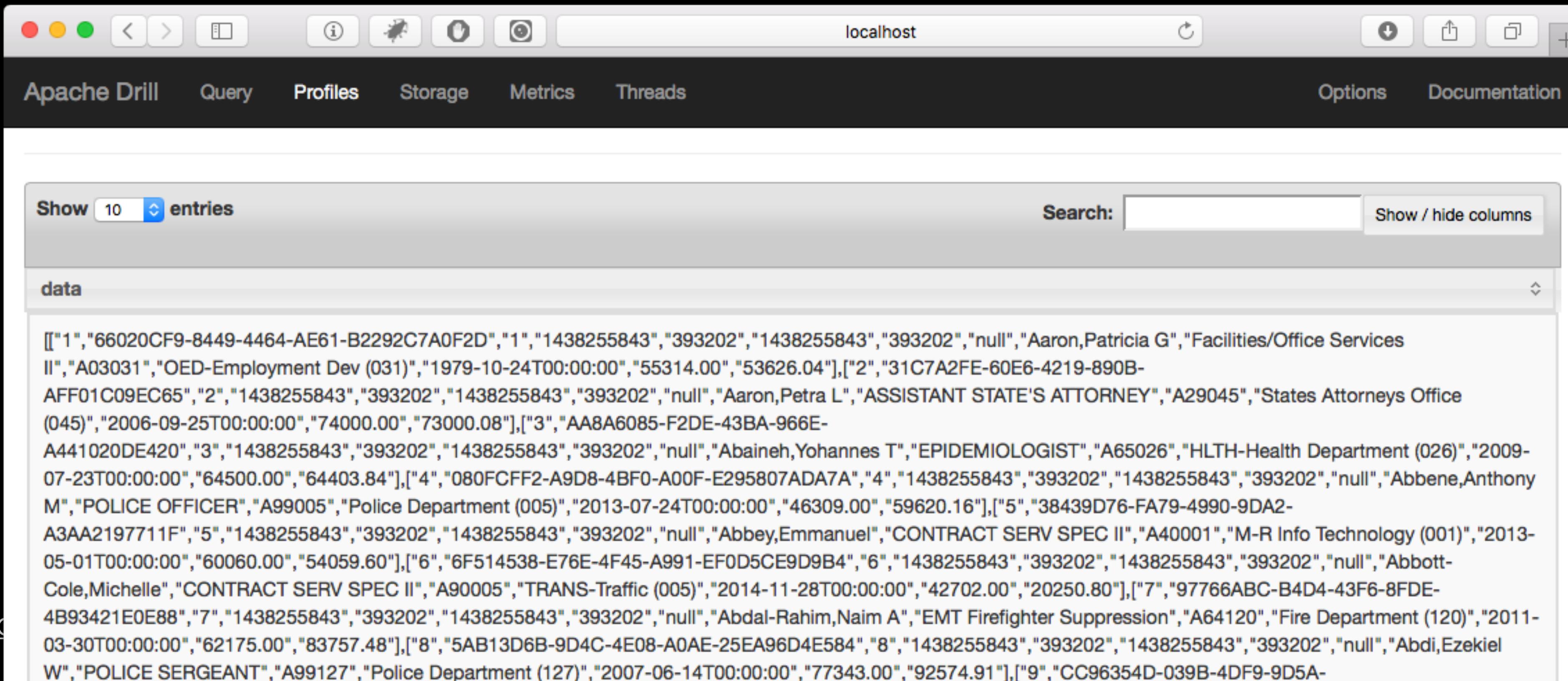


The screenshot shows the Apache Drill web interface running on localhost. The top navigation bar includes links for Apache Drill, Query, Profiles, Storage, Metrics, Threads, Options, and Documentation. Below the header is a search bar and a "Show / hide columns" button. The main content area displays a JSON object under the "meta" tab, with a "data" tab visible on the right. The JSON object represents a view of Baltimore City Employee Salaries FY2015, containing various fields like id, name, and count.

```
{"view": {"id": "nsfe-bg53", "name": "Baltimore City Employee Salaries FY2015", "attribution": "Mayor's Office", "averageRating": "0", "category": "City Government", "createdAt": "142015", "displayType": "table", "downloadCount": "594", "indexUpdatedAt": "1438256605", "licenseId": "CC_30_BY", "newBackend": "false", "numberOfComments": "0", "oid": "12549"}, {"id": "-1", "name": "id", "dataTypeName": "meta_data", "fieldName": ":id", "position": "0", "renderTypeName": "meta_data", "format": {}, "cachedContents": {"top": []}}, {"id": "-1", "name": "updated_at", "dataTypeName": "meta_data", "fieldName": ":updated_at", "position": "0", "renderTypeName": "meta_data", "format": {}, "cachedContents": {"top": []}}, {"tableColumnId": "30185739", "width": "148", "cachedContents": {"non_null": "14017", "smallest": "Aaron,Patricia G", "null": "0", "largest": "Zukowski,Charles J", "top": [{"count": "20", "item": "Zepp,David T"}, {"count": "6", "item": "Zepp,Ronald E"}, {"count": "5", "item": "Zerance,Michael A"}, {"count": "4", "item": "Zero,Benjamin E"}, {"count": "3", "item": "AIDE"}, {"count": "17", "item": "EMT Firefighter Suppression"}, {"count": "16", "item": "RECREATION ARTS INSTRUCTOR"}, {"count": "15", "item": "CROSSING GUARD"}, {"count": "14", "item": "POLICE CADET"}, {"count": "2", "item": "Deputy Fire Chief"}, {"count": "1", "item": "SECRETARY II"}]}, {"id": "215434724", "name": "AgencyID", "dataTypeName": "meta_data", "fieldName": ":agency_id", "position": "0", "renderTypeName": "meta_data", "format": {}, "cachedContents": {"top": [{"count": "11", "item": "A99035"}, {"count": "10", "item": "A99264"}, {"count": "9", "item": "A99123"}, {"count": "8", "item": "A64006"}, {"count": "7", "item": "A75083"}, {"count": "6", "item": "Guards (786)"}, {"count": "18", "item": "FIN-Acct & Payroll (002)"}, {"count": "17", "item": "DPW-Water & Waste Water (207)"}, {"count": "16", "item": "States Attorneys Office (004)"}, {"count": "23T00:00:00", "item": null}, {"count": "10", "item": "largest"}, {"count": "2015-06-29T00:00:00", "item": "top"}]}], "top": [{"count": "20", "item": "2002-02-13T00:00:00"}, {"count": "19", "item": "2011-12-02T00:00:00"}, {"count": "18", "item": "1983-04-07T00:00:00"}, {"count": "5", "item": "2008-05-01T00:00:00"}, {"count": "4", "item": "1998-04-20T00:00:00"}, {"count": "3", "item": "2014-02-19T00:00:00"}, {"count": "18", "item": "30721.00"}, {"count": "17", "item": "65446.00"}, {"count": "16", "item": "39807.00"}, {"count": "15", "item": "62379.00"}, {"count": "14", "item": "55625.00"}, {"count": "part-time or summer clerks/fellows, annual salary is the equivalent full-time annual salary. Comp & leave time excluded."}, {"id": "215434728", "name": "GrossPay", "dataTypeName": "meta_data", "fieldName": ":gross_pay", "position": "0", "renderTypeName": "meta_data", "format": {}, "cachedContents": {"top": [{"count": "14", "item": "55625.00"}, {"count": "13", "item": "62379.00"}, {"count": "12", "item": "39807.00"}, {"count": "11", "item": "65446.00"}, {"count": "10", "item": "30721.00"}, {"count": "9", "item": "1983-04-07T00:00:00"}, {"count": "8", "item": "2014-02-19T00:00:00"}, {"count": "7", "item": "17"}, {"count": "6", "item": "18"}, {"count": "5", "item": "19"}, {"count": "4", "item": "20"}, {"count": "3", "item": "21"}, {"count": "2", "item": "22"}, {"count": "1", "item": "23"}]}]
```

Let's look at this data in Drill

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



The screenshot shows the Apache Drill web interface running on localhost. The top navigation bar includes links for Apache Drill, Query, Profiles, Storage, Metrics, Threads, Options, and Documentation. Below the header is a search bar with 'Show 10 entries' and a 'Search:' field. The main content area displays a JSON array of salary data for Baltimore employees. The data includes fields such as employee ID, name, department, and salary information.

| Employee ID | Name | Department | Salary |
|---|-----------------------|-----------------------------|---|
| "1", "66020CF9-8449-4464-AE61-B2292C7A0F2D" | Aaron, Patricia G | Facilities/Office Services | "null", "1438255843", "393202", "1438255843", "393202", "1438255843", "393202", "null", "Aaron, Patricia G", "Facilities/Office Services |
| "2", "31C7A2FE-60E6-4219-890B-AFF01C09EC65" | Aaron, Petra L | ASSISTANT STATE'S ATTORNEY | "null", "1438255843", "393202", "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"] |
| "3", "AA8A6085-F2DE-43BA-966E-A441020DE420" | Abaineh, Yohannes T | EPIDEMIOLOGIST | "null", "1438255843", "393202", "1438255843", "393202", "1438255843", "393202", "null", "Abaineh, Yohannes T", "EPIDEMIOLOGIST", "A65026", "HLTH-Health Department (026)", "2009-07-23T00:00:00", "64500.00", "64403.84"] |
| "4", "080FCFF2-A9D8-4BF0-A00F-E295807ADA7A" | Abbene, Anthony M | POLICE OFFICER | "null", "1438255843", "393202", "1438255843", "393202", "1438255843", "393202", "null", "Abbene, Anthony M", "POLICE OFFICER", "A99005", "Police Department (005)", "2013-07-24T00:00:00", "46309.00", "59620.16"] |
| "5", "38439D76-FA79-4990-9DA2-A3AA2197711F" | Abbey, Emmanuel | CONTRACT SERV SPEC II | "null", "1438255843", "393202", "1438255843", "393202", "1438255843", "393202", "null", "Abbey, Emmanuel", "CONTRACT SERV SPEC II", "A40001", "M-R Info Technology (001)", "2013-05-01T00:00:00", "60060.00", "54059.60"] |
| "6", "6F514538-E76E-4F45-A991-EF0D5CE9D9B4" | Abbott-Cole, Michelle | CONTRACT SERV SPEC II | "null", "1438255843", "393202", "1438255843", "393202", "1438255843", "393202", "null", "Abbott-Cole, Michelle", "CONTRACT SERV SPEC II", "A90005", "TRANS-Traffic (005)", "2014-11-28T00:00:00", "42702.00", "20250.80"] |
| "7", "97766ABC-B4D4-43F6-8FDE-4B93421E0E88" | Abdal-Rahim, Naim A | EMT Firefighter Suppression | "null", "1438255843", "393202", "1438255843", "393202", "1438255843", "393202", "null", "Abdal-Rahim, Naim A", "EMT Firefighter Suppression", "A64120", "Fire Department (120)", "2011-03-30T00:00:00", "62175.00", "83757.48"] |
| "8", "5AB13D6B-9D4C-4E08-A0AE-25EA96D4E584" | Abdi, Ezekiel W | POLICE SERGEANT | "null", "1438255843", "393202", "1438255843", "393202", "1438255843", "393202", "null", "Abdi, Ezekiel W", "POLICE SERGEANT", "A99127", "Police Department (127)", "2007-06-14T00:00:00", "77343.00", "92574.91"] |
| "9", "CC96354D-039B-4DF9-9D5A- | | | |



FLATTEN(<json array>)

separates elements in a repeated field into individual records.



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



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

The screenshot shows the Apache Drill web interface running on localhost. The top navigation bar includes links for Apache Drill, Query, Profiles, Storage, Metrics, Threads, Options, and Documentation. Below the header is a search and filter section with "Show 10 entries" and a "Search:" input field. The main content area displays five rows of JSON data under the heading "raw_data".

| raw_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)", "date": "1979-10-24T00:00:00", "salary": "55314.00", "id": "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)", "date": "2006-09-25T00:00:00", "salary": "74000.00", "id": "73000.08"}] |
| [{"3": "AA8A6085-F2DE-43BA-966E-A441020DE420", "3": "1438255843", "393202": "1438255843", "393202": "null", "Abaineh, Yohannes T": "EPIDEMIOLOGIST", "A65026": "HLTH-Health Department (026)", "date": "2009-07-23T00:00:00", "salary": "64500.00", "id": "64403.84"}] |
| [{"4": "080FCFF2-A9D8-4BF0-A00F-E295807ADA7A", "4": "1438255843", "393202": "1438255843", "393202": "null", "Abbene, Anthony M": "POLICE OFFICER", "A99005": "Police Department (005)", "date": "2013-07-24T00:00:00", "salary": "46309.00", "id": "59620.16"}] |
| [{"5": "38439D76-FA79-4990-9DA2-A3AA2197711F", "5": "1438255843", "393202": "1438255843", "393202": "null", "Abbey, Emmanuel": "CONTRACT SERV SPEC II", "A40001": "M-R Info Technology (001)", "date": "2013-05-01T00:00:00", "salary": "60060.00", "id": "54059.60"}] |

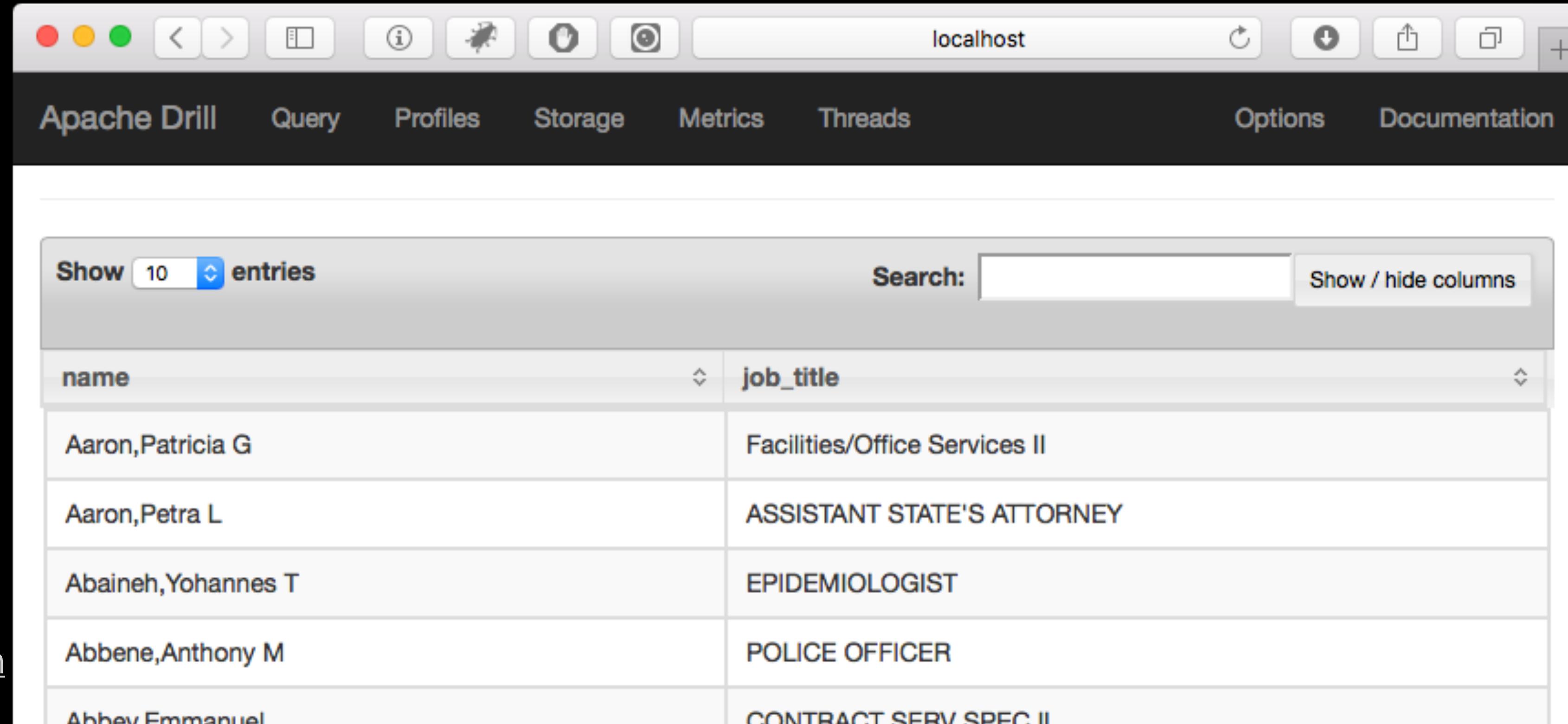


```
SELECT FLATTEN( data ) AS raw_data
FROM dfs.drillworkshop.`baltimore_salaries.json`
```



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

```
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`  
)
```



The screenshot shows the Apache Drill web interface running on localhost. The top navigation bar includes links for Apache Drill, Query, Profiles, Storage, Metrics, Threads, Options, and Documentation. Below the navigation is a search bar with filters for 'Show 10 entries' and a 'Search:' field. A 'Show / hide columns' button is also present. The main content area displays a table with two columns: 'name' and 'job_title'. The data rows are:

| name | job_title |
|--------------------|-------------------------------|
| Aaron,Patricia G | Facilities/Office Services II |
| Aaron,Petra L | ASSISTANT STATE'S ATTORNEY |
| Abaineh,Yohannes T | EPIDEMIOLOGIST |
| Abbene,Anthony M | POLICE OFFICER |
| Abbey Emmanuel | CONTRACT SERV SPEC II |



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...

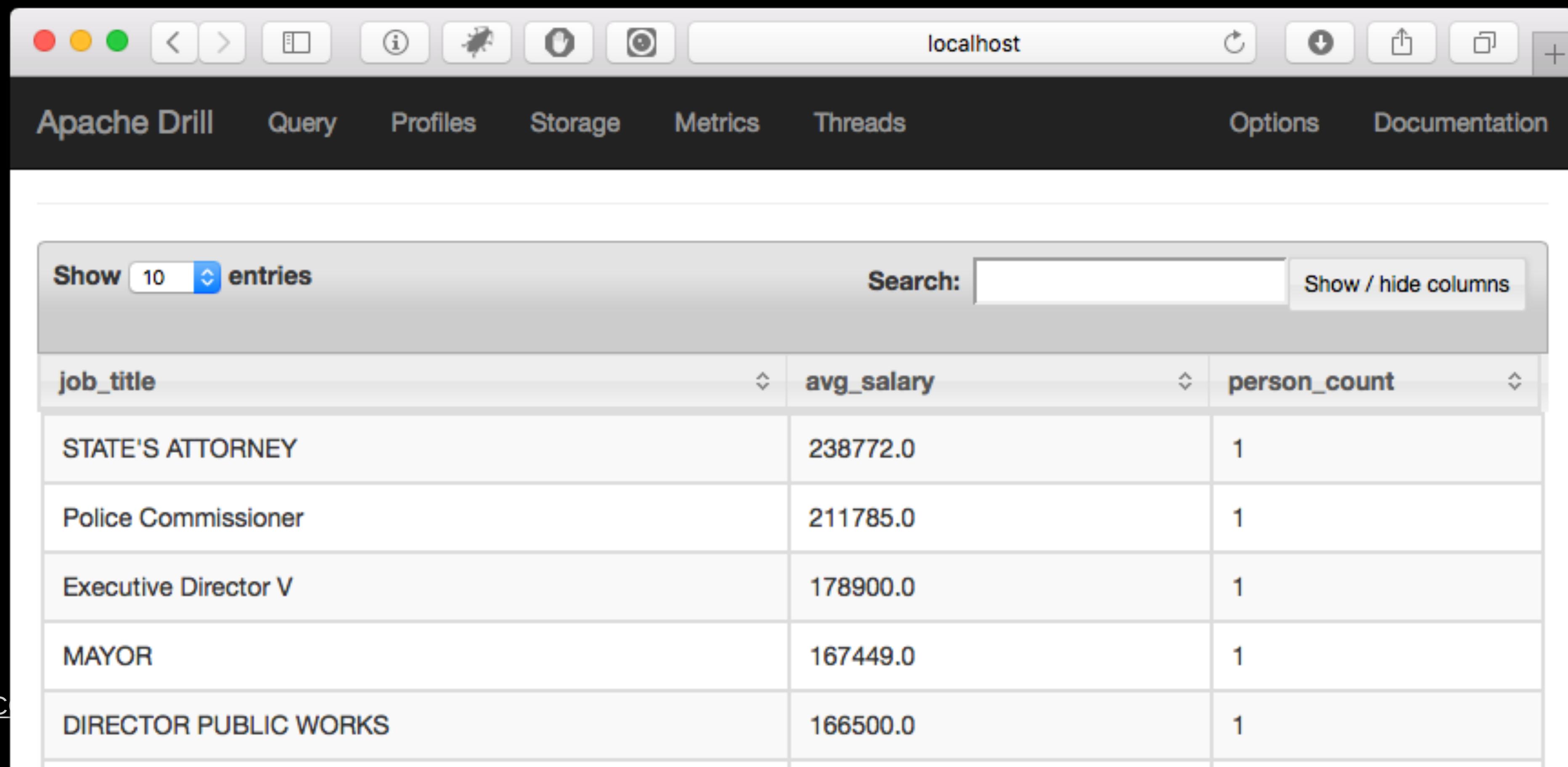
HINT 2: GROUP BY does NOT support aliases.

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
```

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



The screenshot shows the Apache Drill interface running on localhost. The top navigation bar includes links for Apache Drill, Query, Profiles, Storage, Metrics, Threads, Options, and Documentation. Below the navigation is a search and filter panel with "Show 10 entries" and a "Search:" field. The main content area displays a table with three columns: job_title, avg_salary, and person_count. The data is as follows:

| job_title | avg_salary | person_count |
|-----------------------|------------|--------------|
| 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 |



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

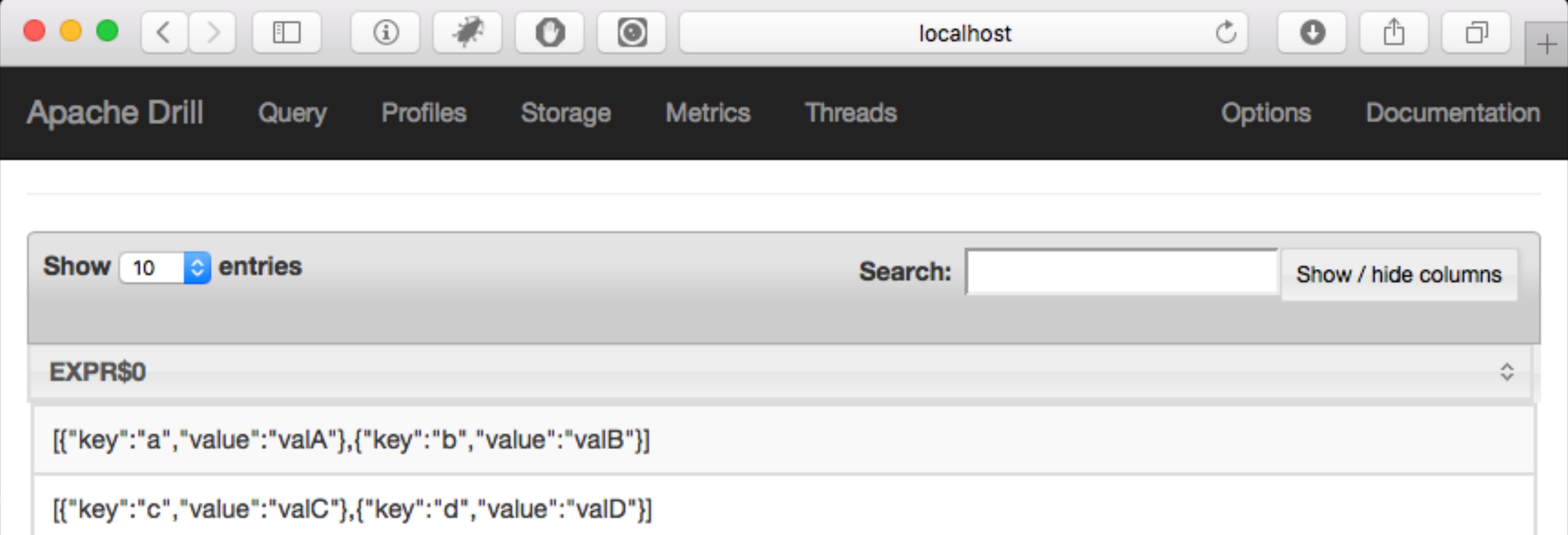


```
{ "rec1": { "a": "valA", "b": "valB" } }
{ "rec1": { "c": "valC", "d": "valD" } }
```



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

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



A screenshot of the Apache Drill web interface. The top navigation bar includes links for Apache Drill, Query, Profiles, Storage, Metrics, Threads, Options, and Documentation. The main content area shows a table with two rows of data. The first row has the header 'EXPR\$0' and contains the JSON object: [{"key": "a", "value": "valA"}, {"key": "b", "value": "valB"}]. The second row contains the JSON object: [{"key": "c", "value": "valC"}, {"key": "d", "value": "valD"}]. The interface includes standard table controls like 'Show 10 entries' and a search bar.

| EXPR\$0 |
|--|
| [{"key": "a", "value": "valA"}, {"key": "b", "value": "valB"}] |
| [{"key": "c", "value": "valC"}, {"key": "d", "value": "valD"}] |



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

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

The screenshot shows the Apache Drill web interface running on localhost. The top navigation bar includes links for Apache Drill, Query, Profiles, Storage, Metrics, Threads, Options, and Documentation. Below the header is a search bar with 'Search:' and a 'Show / hide columns' button. On the left, there's a 'Show 10 entries' dropdown and a 'Search:' input field. The main content area displays a table with one column labeled 'EXPR\$0'. The table contains four rows of JSON objects:

| EXPR\$0 |
|-------------------------------|
| {"key": "a", "value": "valA"} |
| {"key": "b", "value": "valB"} |
| {"key": "c", "value": "valC"} |



Saving Data

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
```



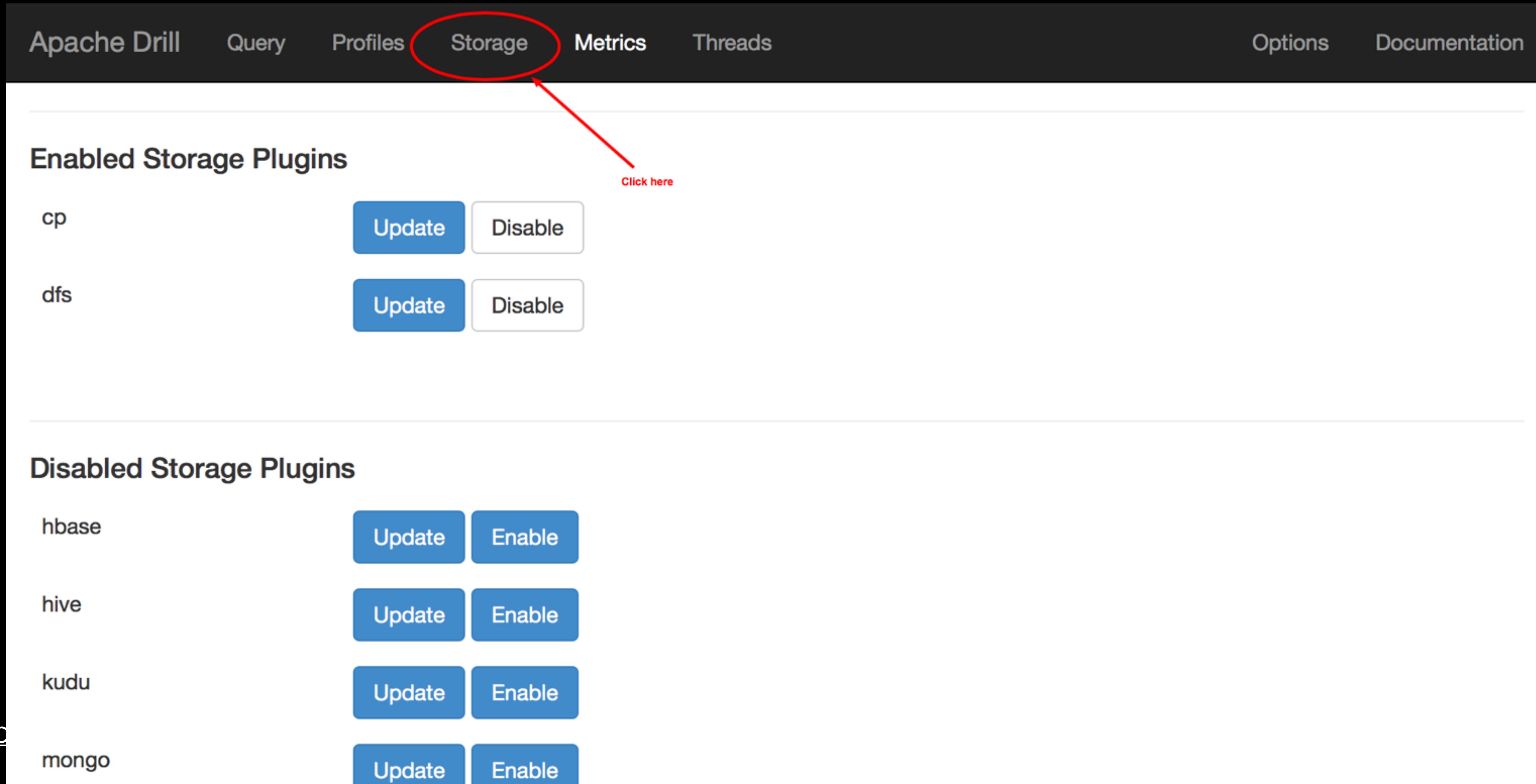
Connecting other Data Sources



Connecting other Data Sources



Connecting other Data Sources



The screenshot shows the Apache Drill web interface with the 'Storage' tab selected. A red circle highlights the 'Storage' tab, and a red arrow points to the 'Click here' link below it.

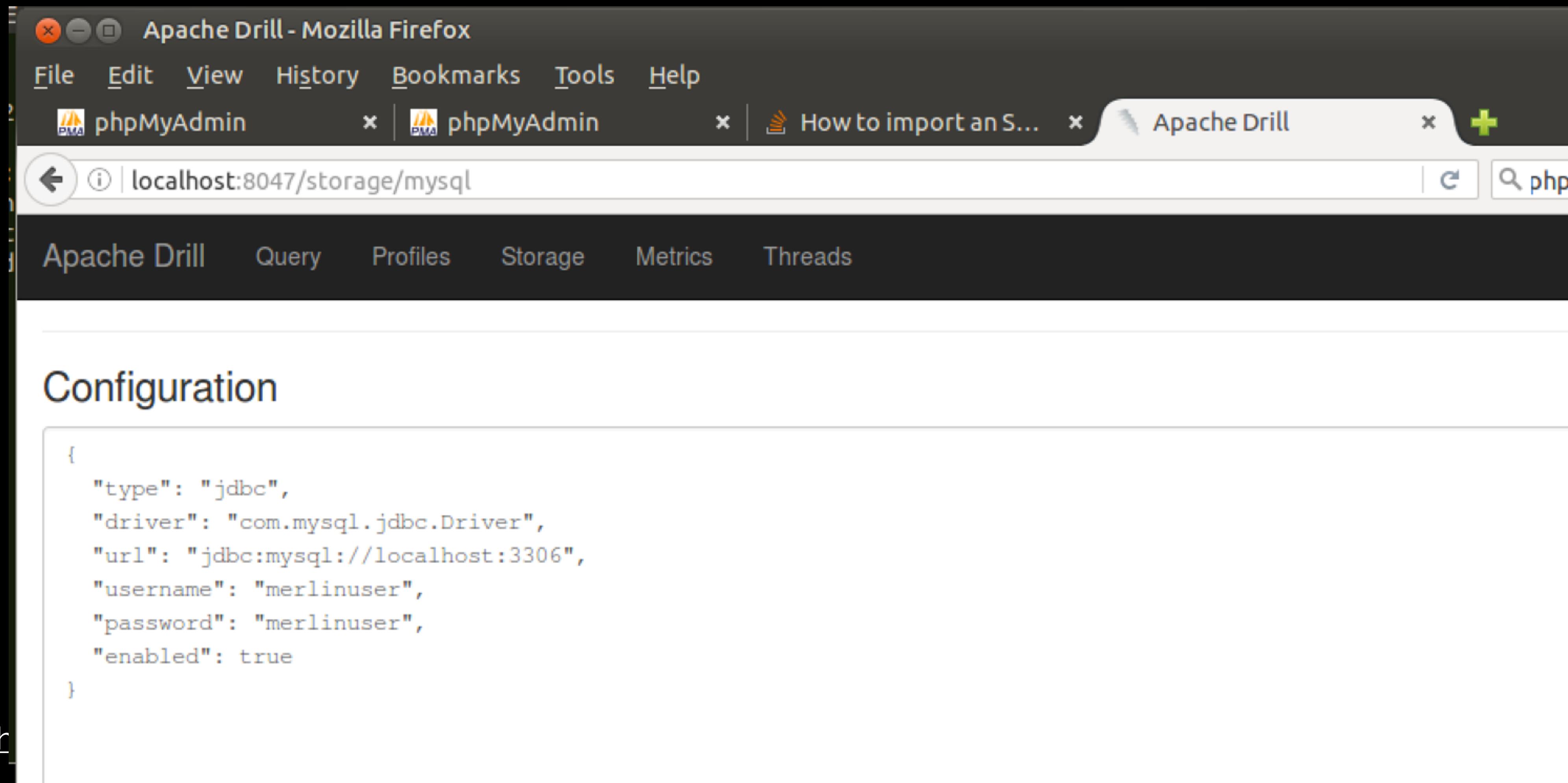
Enabled Storage Plugins

| | | |
|-----|------------------------|-------------------------|
| cp | Update | Disable |
| dfs | Update | Disable |

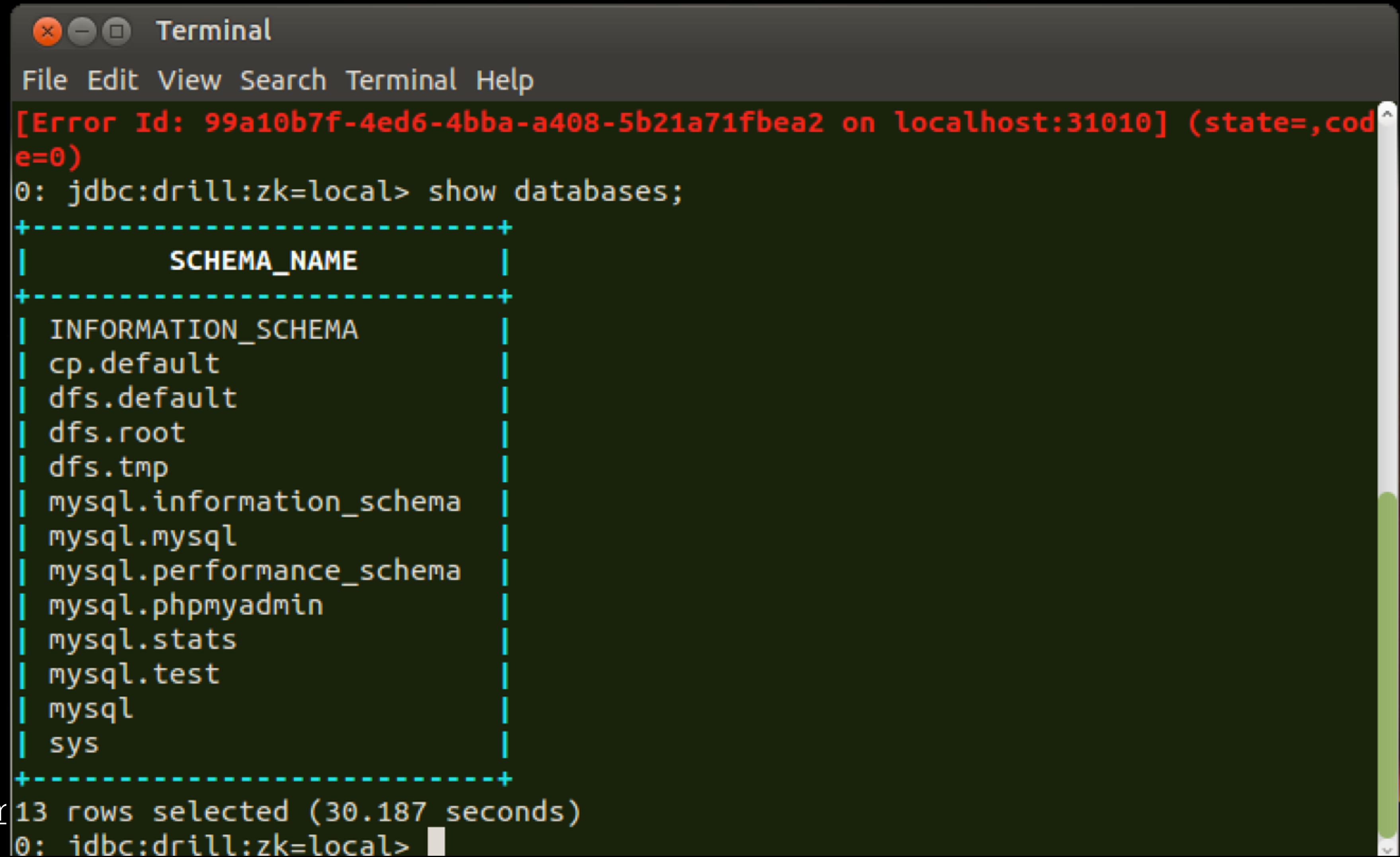
Disabled Storage Plugins

| | | |
|-------|------------------------|------------------------|
| hbase | Update | Enable |
| hive | Update | Enable |
| kudu | Update | Enable |
| mongo | Update | Enable |

Connecting other Data Sources



Connecting other Data Sources



The screenshot shows a terminal window titled "Terminal" with a dark background. The window contains the following text:

```
[Error Id: 99a10b7f-4ed6-4bba-a408-5b21a71fbea2 on localhost:31010] (state=, code=0)
0: jdbc:drill:zk=local> show databases;
+-----+
| SCHEMA_NAME |
+-----+
| INFORMATION_SCHEMA |
| cp.default |
| dfs.default |
| dfs.root |
| dfs.tmp |
| mysql.information_schema |
| mysql.mysql |
| mysql.performance_schema |
| mysql.phpmyadmin |
| mysql.stats |
| mysql.test |
| mysql |
| sys |
+-----+
13 rows selected (30.187 seconds)
0: jdbc:drill:zk=local>
```



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
```



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



Connecting to Drill

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



Connecting to Drill

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



Questions?



Thank you!

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