databricks4.1 The Spark UI

```
%run ../Includes/Classroom-Setup
```

Mounting course-specific datasets to **/mnt/training**...
Datasets are already mounted to **/mnt/training** from **s3a://databricks-corp-training/common**

```
res1: Boolean = false
```

```
res2: Boolean = false
```

```
DROP TABLE IF EXISTS People10M;
CREATE TABLE People10M
USING csv
OPTIONS (
path "/mnt/training/dataframes/people-10m.csv",
header "true");

DROP TABLE IF EXISTS ssaNames;
CREATE TABLE ssaNames USING parquet OPTIONS (
  path "/mnt/training/ssn/names.parquet",
  header "true"
);
```

OK

Catalog Error

```
SELECT * FROM People10M
```

	id	firstName	middleName	lastName	gender	birth
1	1	Pennie	Carry	Hirschmann	F	1955
2	2	An	Amira	Cowper	F	1992

3	3	Quyen	Marlen	Dome	F	1970
4	4	Coralie	Antonina	Marshal	F	1990
5	5	Terrie	Wava	Bonar	F	1980
6	6	Chassidy	Concepcion	Bourthouloume	F	1990
7	7	Geri	Tambra	Mosby	F	1970

Truncated results, showing first 1000 rows.

```
DESCRIBE People10M
```

	col_name	data_type	comment
1	id	string	null
2	firstName	string	null
3	middleName	string	null
4	lastName	string	null
5	gender	string	null
6	birthDate	string	null
7	ssn	string	null

Showing all 8 rows.

```
SELECT
  firstName,
  lastName,
  birthDate
FROM
  People10M
WHERE
  year(birthDate) > 1990
  AND gender = 'F'
```

	firstName	lastName	birthDate
1	An	Cowper	1992-02-08T05:00:00.000Z
2	Caroyln	Cardon	1994-05-15T04:00:00.000Z
3	Yesenia	Goldring	1997-07-09T04:00:00.000Z
4	Hedwig	Pendleberry	1998-12-02T05:00:00.000Z
5	Kala	Lyfe	1994-06-23T04:00:00.000Z
6	Gussie	McKeeman	1991-11-15T05:00:00.000Z
7	Pansy	Shrieves	1991-05-24T04:00:00.000Z

Truncated results, showing first 1000 rows.

Plan Optimization Example

```
CREATE OR REPLACE TEMPORARY VIEW joined AS
SELECT People10m.firstName,
  to_date(birthDate) AS date
FROM People10m
  JOIN ssaNames ON People10m.firstName = ssaNames.firstName;

CREATE OR REPLACE TEMPORARY VIEW filtered AS
SELECT firstName,count(firstName)
FROM joined
WHERE
  date >= "1990-01-01"
GROUP BY
  firstName, date;
```

OK

```
SELECT * FROM  filtered;
```

	firstName	count(firstName)
1	Virgen	38
2	Despina	99
3	Roberta	207
4	Jolie	68
5	Karena	63
6	Shiela	79
7	Emeline	130

Truncated results, showing first 1000 rows.

```
CACHE TABLE filtered;
```

OK

```
SELECT * FROM filtered;
```

	firstName	count(firstName)
1	Virgen	38
2	Despina	99
3		

	Roberta	207
4	Jolie	68
5	Karena	63
6	Shiela	79
7	Emeline	130

Truncated results, showing first 1000 rows.

```
SELECT * FROM filtered WHERE firstName = "Latisha";
```

	firstName	count(firstName)
1	Latisha	72
2	Latisha	72
3	Latisha	72
4	Latisha	72
5	Latisha	72
6	Latisha	72
7	Latisha	72

Showing all 260 rows.

```
UNCACHE TABLE IF EXISTS filtered;
```

OK

```
SELECT * FROM filtered WHERE firstName = "Latisha";
```

	firstName	count(firstName)
1	Latisha	72
2	Latisha	72
3	Latisha	72
4	Latisha	72
5	Latisha	72
6	Latisha	72
7	Latisha	72

Showing all 260 rows.

Set Partitions

```
DROP TABLE IF EXISTS bikeShare;
CREATE TABLE bikeShare
USING csv
OPTIONS (
  path "/mnt/training/bikeSharing/data-001/hour.csv",
  header "true")
```

OK

```
SELECT
  *
FROM
  bikeShare
WHERE
  hr = 10
```

	instant	dteday	season	yr	mnth	hr
1	11	2011-01-01	1	0	1	10
2	34	2011-01-02	1	0	1	10
3	56	2011-01-03	1	0	1	10
4	79	2011-01-04	1	0	1	10
5	102	2011-01-05	1	0	1	10
6	125	2011-01-06	1	0	1	10
7	148	2011-01-07	1	0	1	10

Showing all 727 rows.

```
DROP TABLE IF EXISTS bikeShare_partitioned;
CREATE TABLE bikeShare_partitioned
PARTITIONED BY (p_hr)
  AS
SELECT
  instant,
  dteday,
  season,
  yr,
  mnth,
  hr as p_hr,
  holiday,
  weekday,
  workingday,
  weathersit,
  temp
FROM
  bikeShare
```

Query returned no results

```
SELECT * FROM bikeShare_partitioned WHERE p_hr = 10
```

	instant	dteday	season	yr	mnth	p_hr
1	11	2011-01-01	1	0	1	10
2	34	2011-01-02	1	0	1	10
3	56	2011-01-03	1	0	1	10
4	79	2011-01-04	1	0	1	10
5	102	2011-01-05	1	0	1	10
6	125	2011-01-06	1	0	1	10
7	148	2011-01-07	1	0	1	10

Showing all 727 rows.

Beware of small files!

```
DROP TABLE IF EXISTS bikeShare_parquet;
CREATE TABLE bikeShare_instent
PARTITIONED BY (p_instant)
  AS
SELECT
  instant AS p_instant,
  dteday,
  season,
  yr,
  mnth,
  hr
  holiday,
  weekday,
  workingday,
  weathersit,
  temp
FROM
  bikeShare
```

```
  Cancelled
```

```
%run ../Includes/Classroom-Cleanup
```

Citations

Bike Sharing Data

[1] Fanaee-T, Hadi, and Gama, Joao, Event labeling combining ensemble detectors and background knowledge, Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.

@article{ year={2013}, issn={2192-6352}, journal={Progress in Artificial Intelligence}, doi={10.1007/s13748-013-0040-3}, title={Event labeling combining ensemble detectors and background knowledge}, url={http://dx.doi.org/10.1007/s13748-013-0040-3} (http://dx.doi.org/10.1007/s13748-013-0040-3}), publisher={Springer Berlin Heidelberg}, keywords={Event labeling; Event detection; Ensemble learning; Background knowledge}, author={Fanaee-T, Hadi and Gama, Joao}, pages={1-15} }