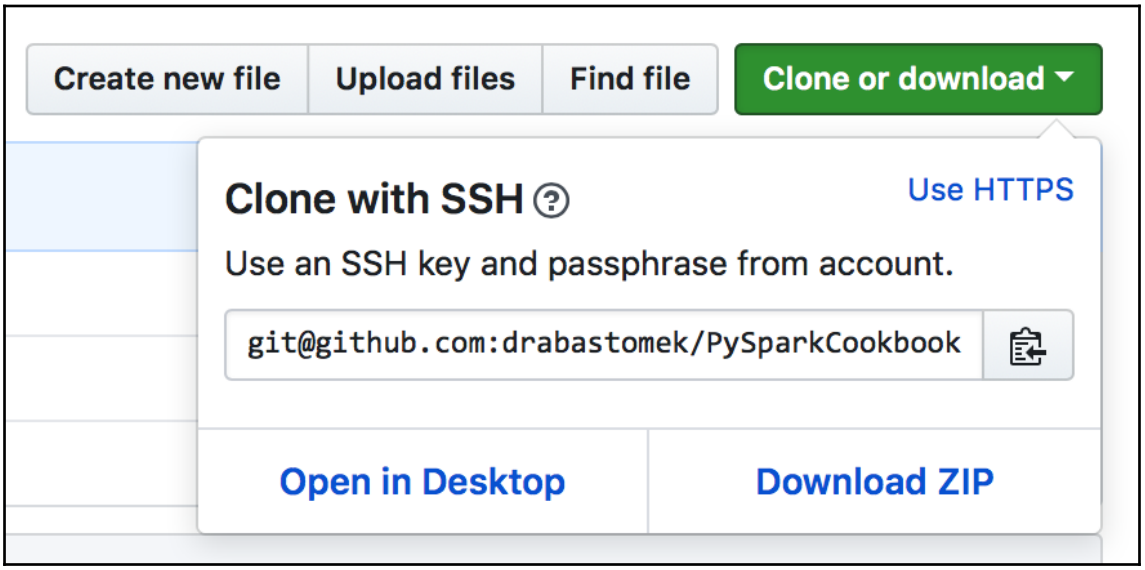
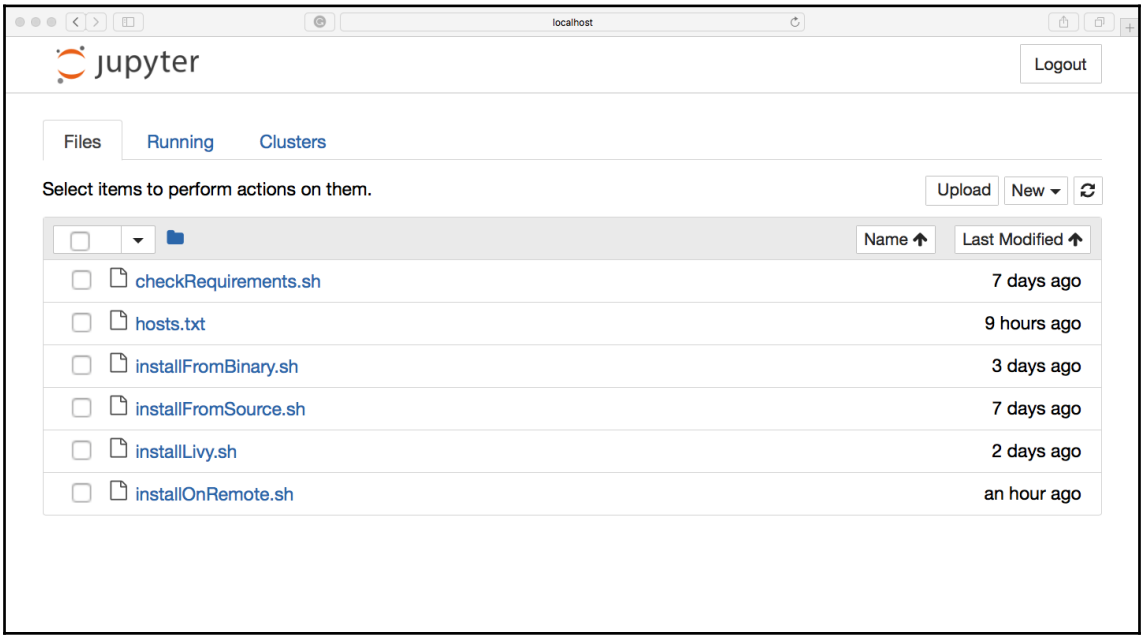
Chapter 1: Installing and Configuring Spark

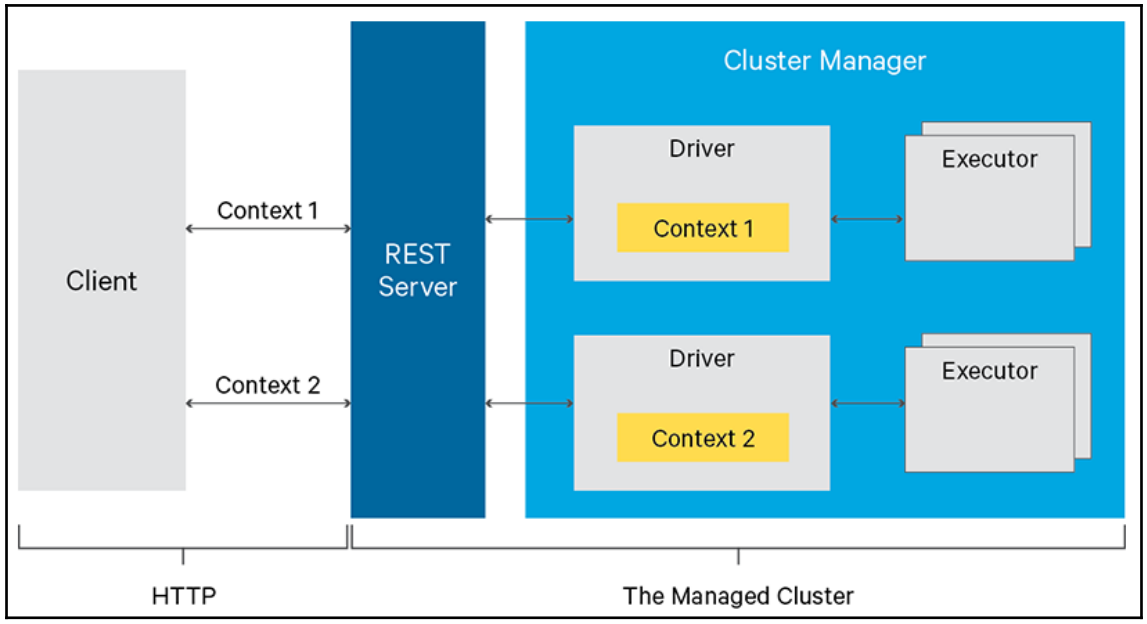


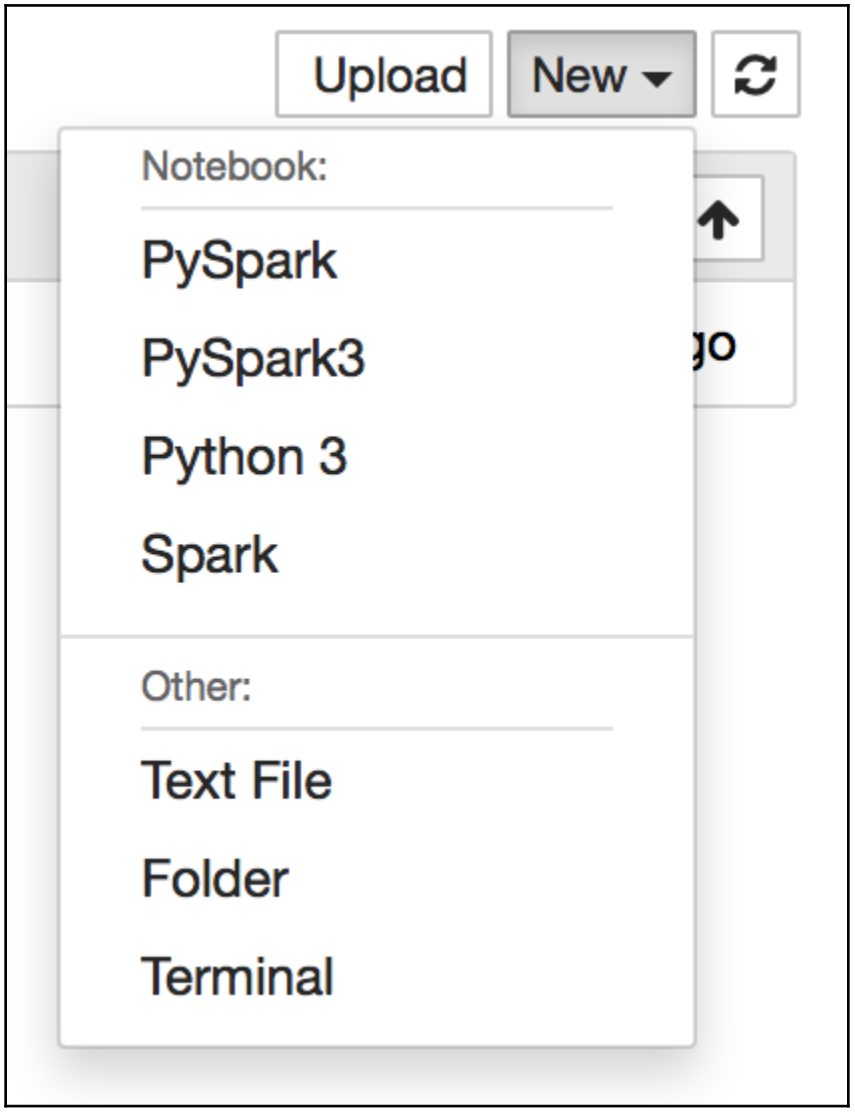

```
endeavor:Chapter01 drabast$ java -version
java version "1.8.0_25"
Java(TM) SE Runtime Environment (build 1.8.0_25-b17)
Java HotSpot(TM) 64-Bit Server VM (build 25.25-b02, mixed mode)
```

```
endeavor:Chapter03 drabast$ pyspark --version
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.3.0
      /_/

Using Scala version 2.11.8, Java HotSpot(TM) 64-Bit Server VM, 1.8.0_25
Branch master
Compiled by user sameera on 2018-02-22T19:24:29Z
Revision a0d7949896e70f427e7f3942ff340c9484ff0aab
Url git@github.com:sameeragarwal/spark.git
Type --help for more information.
```

```
endeavor:learningPySpark drabast$ pip install pyspark
Collecting pyspark
  Downloading pyspark-2.2.0.post0.tar.gz (188.3MB)
    100% |████████████████████████████████| 188.3MB 3.7kB/s
Requirement already satisfied: py4j==0.10.4 in /Users/drabast/anaconda/lib/python3.5/site-packages (from pyspark)
Building wheels for collected packages: pyspark
  Running setup.py bdist_wheel for pyspark ... done
  Stored in directory: /Users/drabast/Library/Caches/pip/wheels/5f/0b/b3/5cb16b15d28dcc32f8e7ec91a044829642874bb7586f6e6cbe
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-2.2.0
```




Client
Context 1
Context 2
REST Server
Cluster Manager
Driver
Context 1
Executor
Driver
Context 2
Executor
HTTP
The Managed Cluster


Upload
New
Notebook:
PySpark
PySpark3
Python 3
Spark
Other:
Text File
Folder
Terminal

```
In [1]: %%configure
{
    "executorCores" : 3
}
```

Current session configs: `{'kind': 'pyspark', 'executorCores': 3}`

No active sessions.

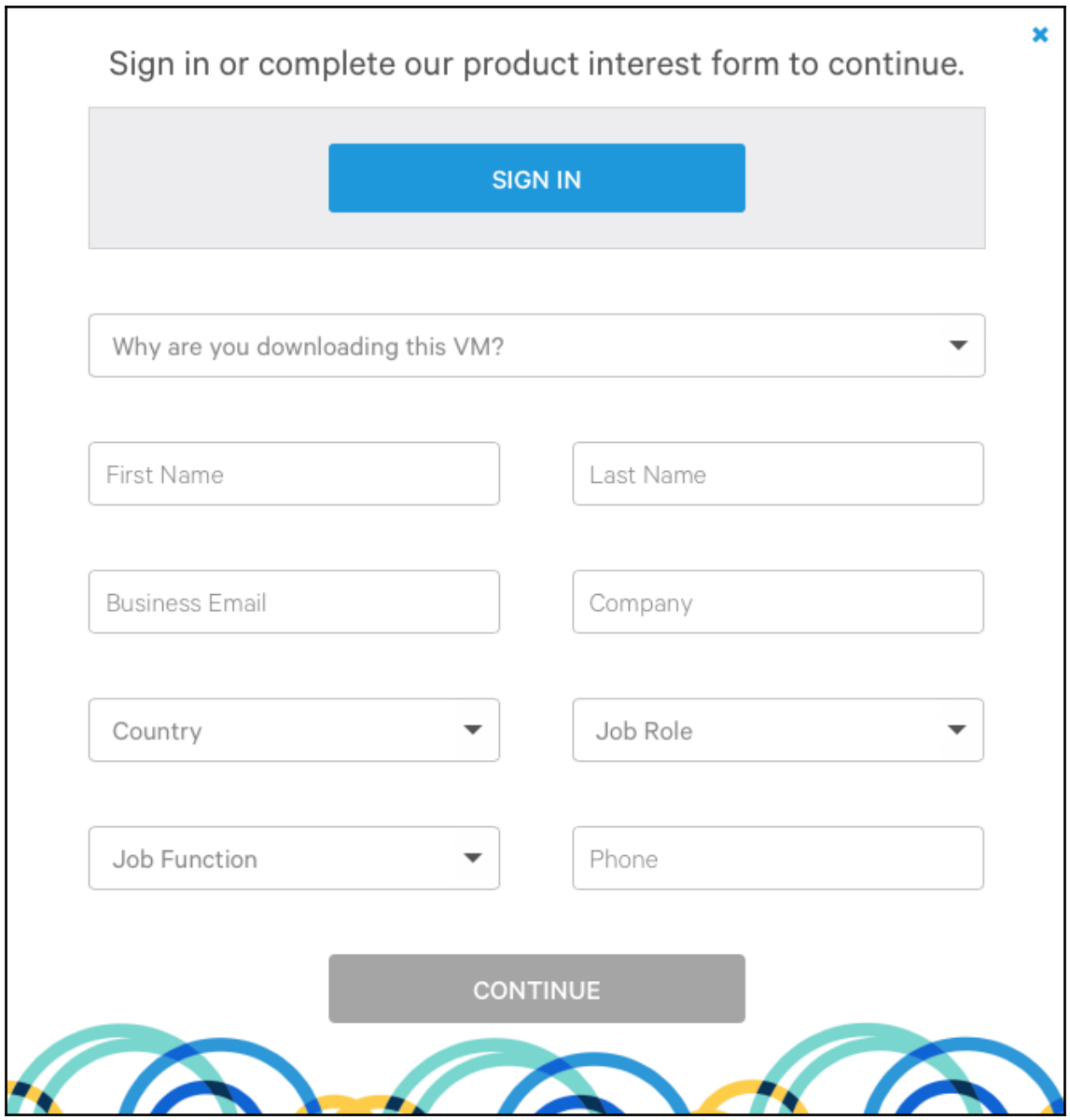
```
Starting Spark application
```

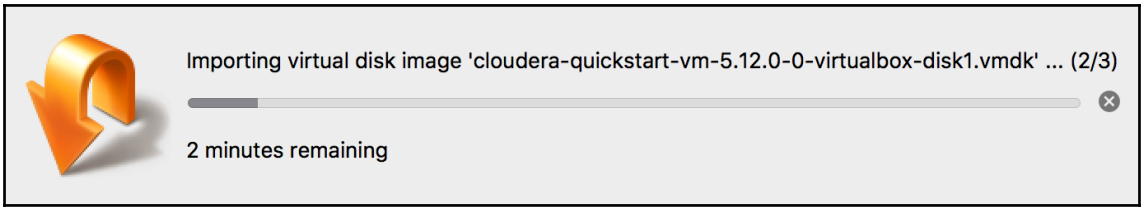
ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
1	None	pyspark	idle			✔

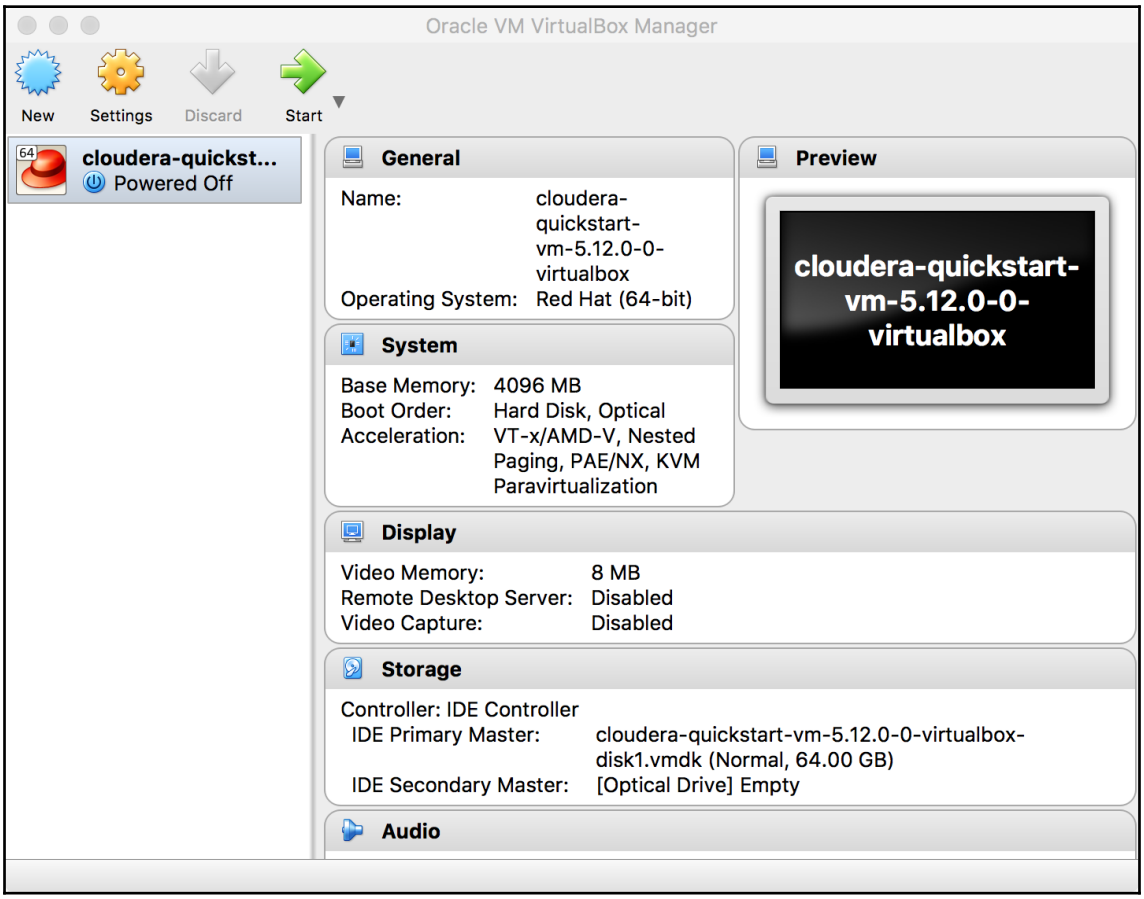
```
SparkSession available as 'spark'.
```

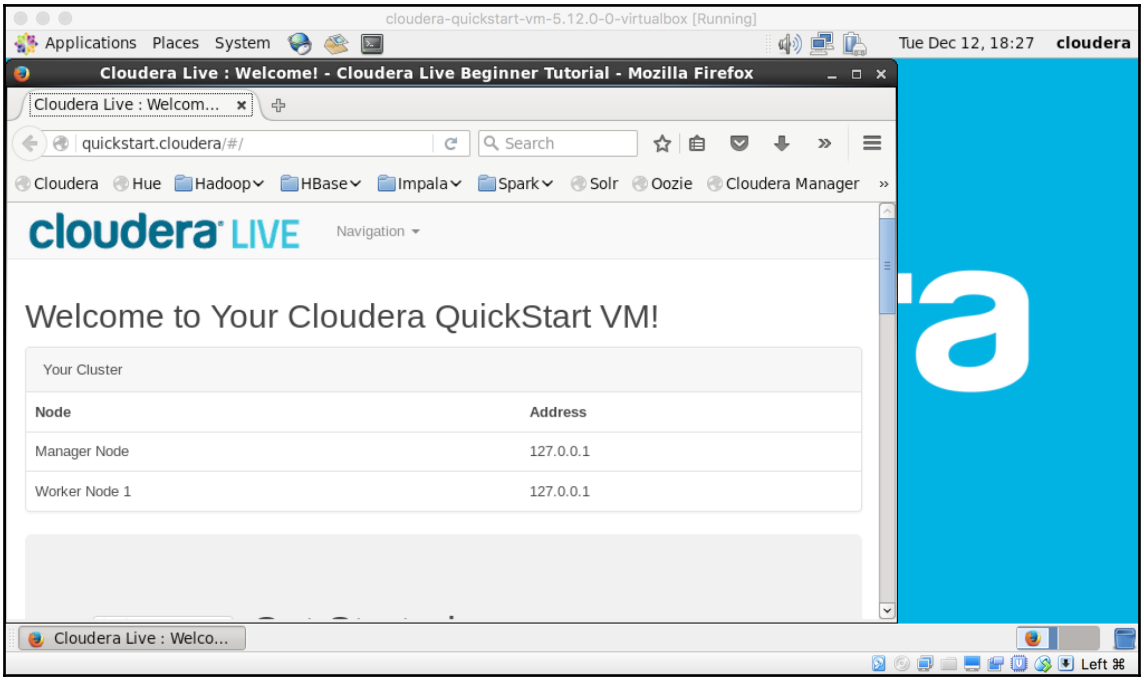
```
In [5]: %%sql
        SELECT * FROM swimmers
```

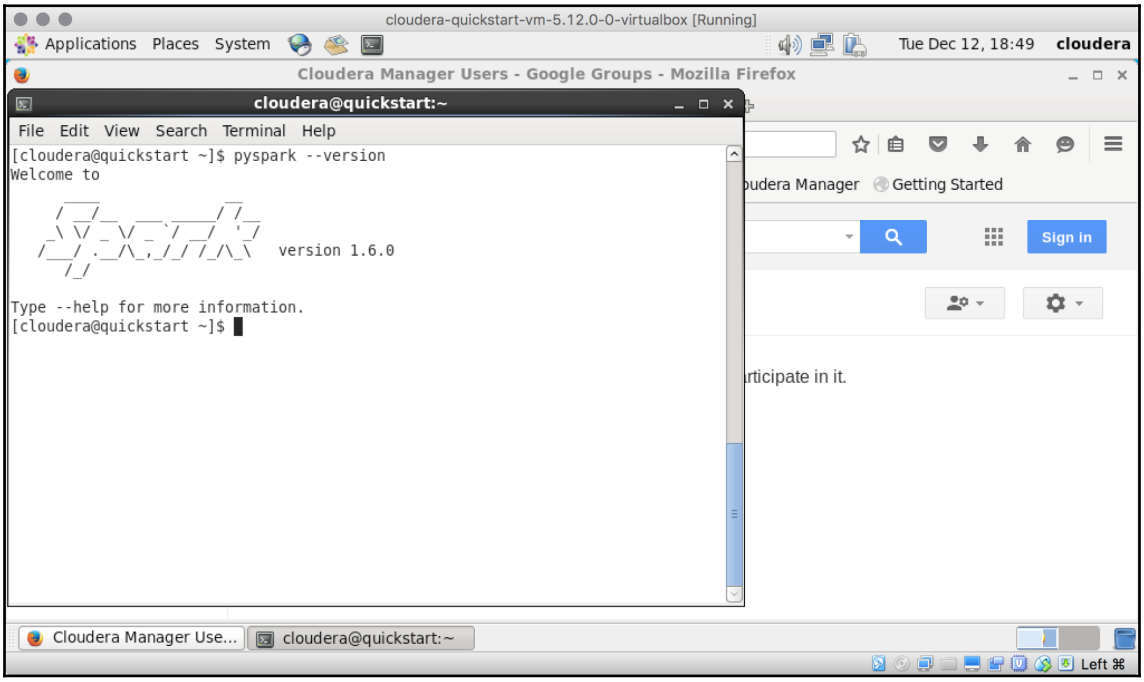
id	name	age	eyeColor
123	Katie	19	brown
234	Michael	22	green
345	Simone	23	blue


Sign in or complete our product interest form to continue.
SIGN IN
Why are you downloading this VM?
First Name
Last Name
Business Email
Company
Country
Job Role
Job Function
Phone
CONTINUE

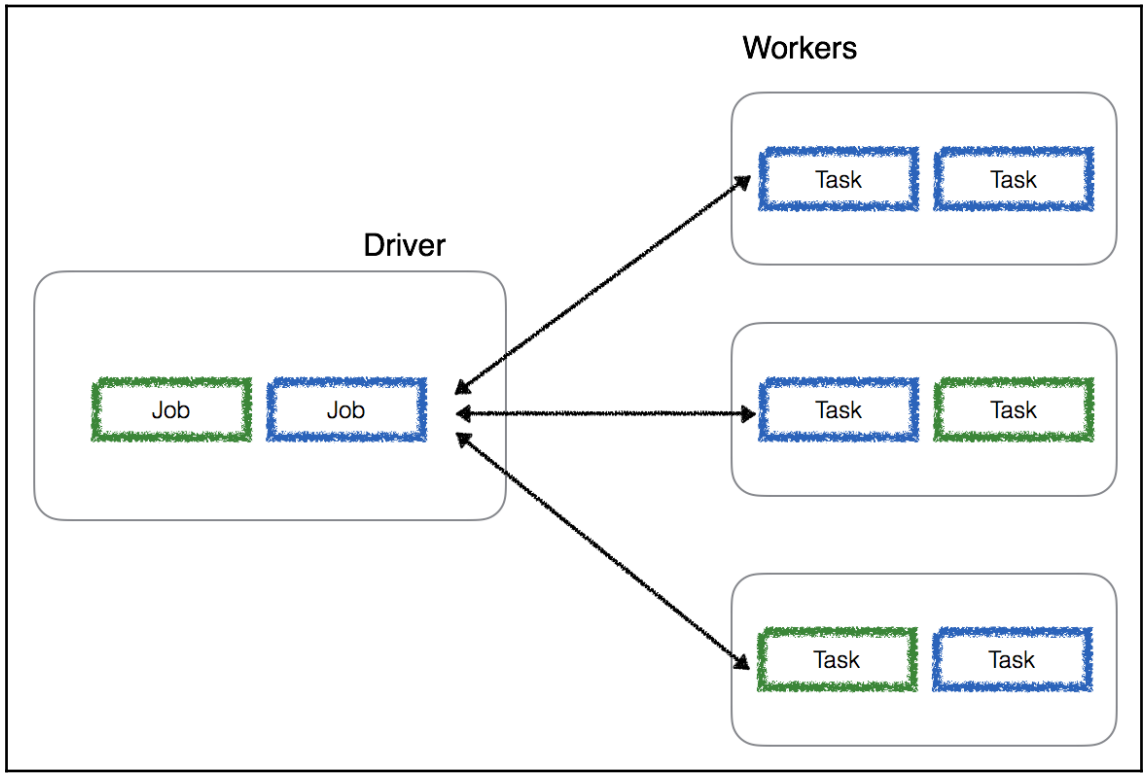

Importing virtual disk image 'cloudera-quickstart-vm-5.12.0-0-virtualbox-disk1.vmdk' ... (2/3)
2 minutes remaining


Oracle VM VirtualBox Manager
New
Settings
Discard
Start
cloudera-quickst...
Powered Off
General
Name: cloudera-quickstart-vm-5.12.0-0-virtualbox
Operating System: Red Hat (64-bit)
System
Base Memory: 4096 MB
Boot Order: Hard Disk, Optical
Acceleration: VT-x/AMD-V, Nested Paging, PAE/NX, KVM Paravirtualization
Preview
cloudera-quickstart-vm-5.12.0-0-virtualbox
Display
Video Memory: 8 MB
Remote Desktop Server: Disabled
Video Capture: Disabled
Storage
Controller: IDE Controller
IDE Primary Master: cloudera-quickstart-vm-5.12.0-0-virtualbox-disk1.vmdk (Normal, 64.00 GB)
IDE Secondary Master: [Optical Drive] Empty
Audio


cloudera-quickstart-vm-5.12.0-0-virtualbox [Running]
Applications Places System
Tue Dec 12, 18:27 cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Cloudera Live : Welcom...
quickstart.cloudera/#/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager
cloudera LIVE
Navigation
Welcome to Your Cloudera QuickStart VM!
Your Cluster
Node
Address
Manager Node
127.0.0.1
Worker Node 1
127.0.0.1
Cloudera Live : Welco...
Left ⌘


cloudera-quickstart-vm-5.12.0-0-virtualbox [Running]
Applications Places System
Tue Dec 12, 18:49 cloudera
Cloudera Manager Users - Google Groups - Mozilla Firefox
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ pyspark --version
Welcome to
version 1.6.0
Type --help for more information.
[cloudera@quickstart ~]$
Cloudera Manager
Getting Started
Sign in
Cloudera Manager Use...
cloudera@quickstart:~
Left ⌘

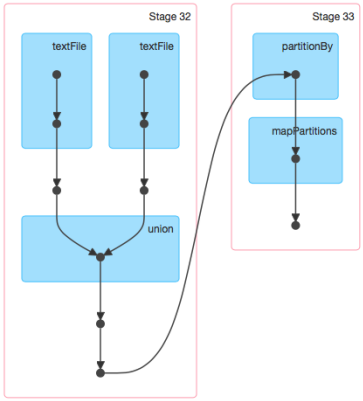
Chapter 2: Abstracting Data with RDDs



Details for Job 24

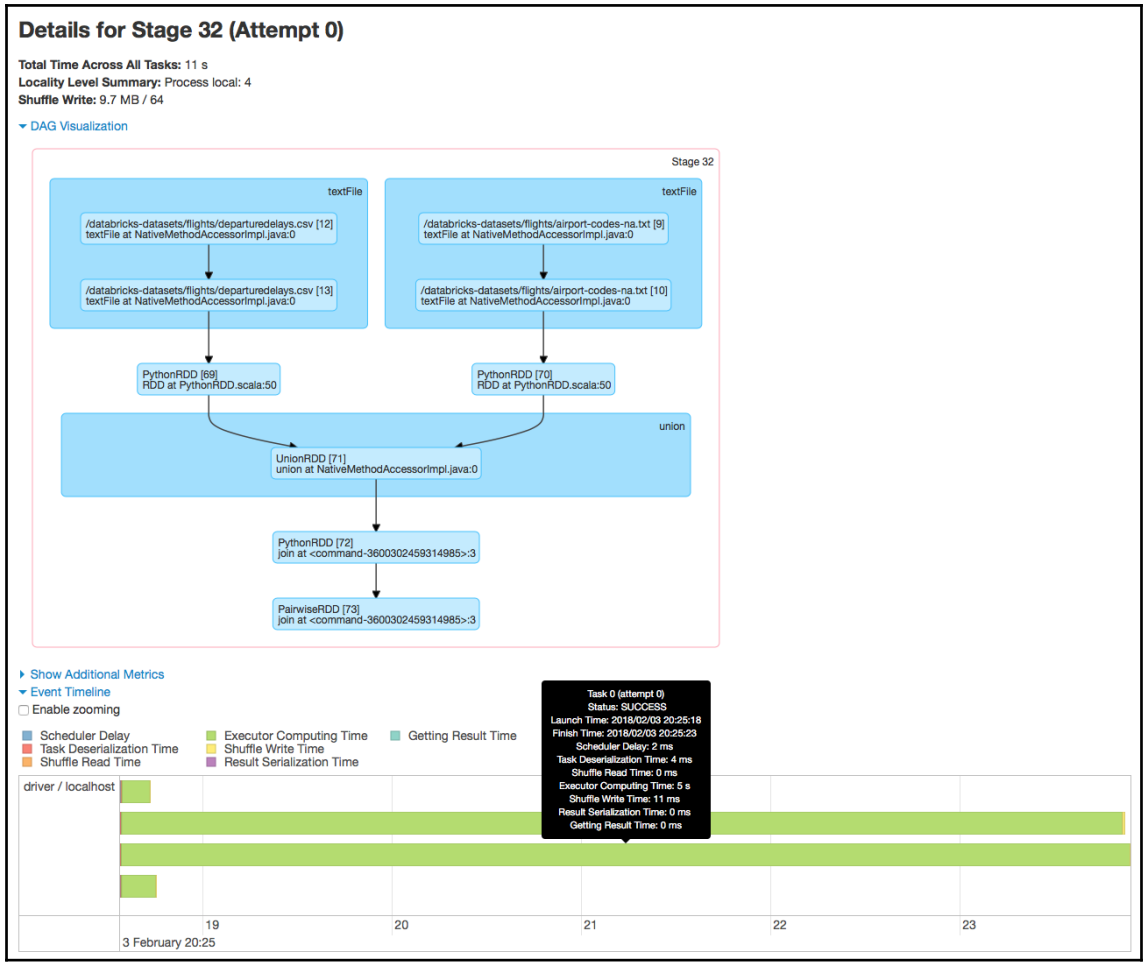
Status: SUCCEEDED
Job Group: 4207587983499043884_5499242308750819941_4c7f1ba459894fcb89ba9a63a10849a8
Completed Stages: 2

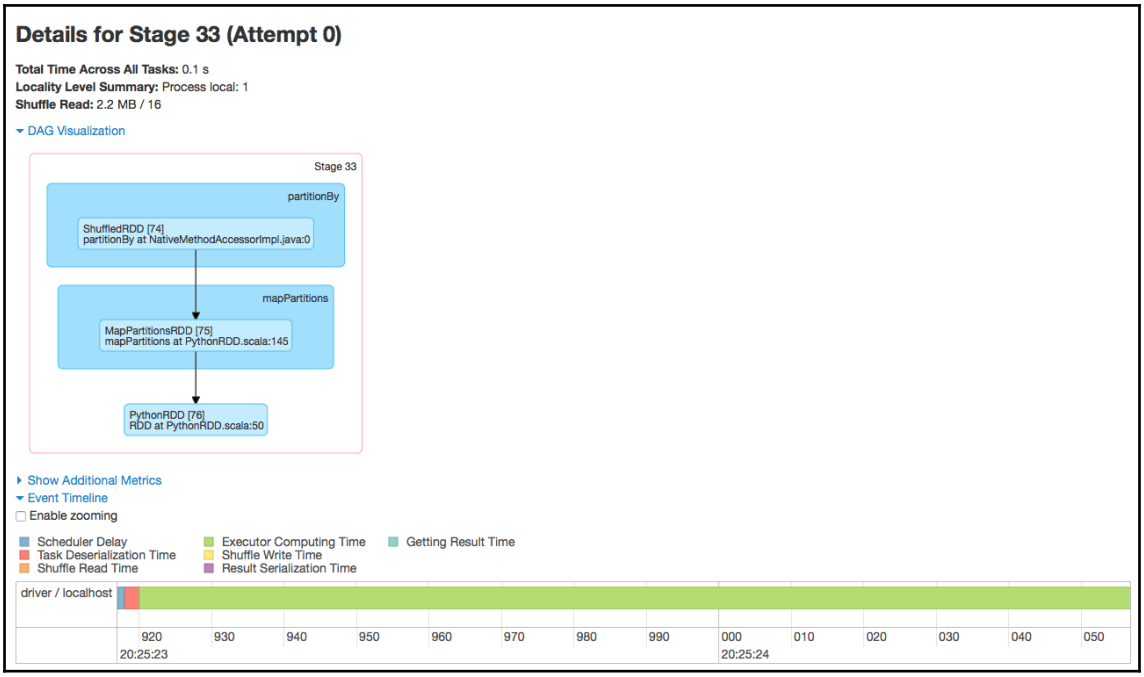
▸ Event Timeline
▾ DAG Visualization



Completed Stages (2)

Stage Id ▾	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
33	4207587983499043884	flt = flights.map(lambda c: (c[3], c[0])) air =... runJob at PythonRDD.scala:165 +details	2018/02/03 20:25:23	0.1 s	1/1			2.2 MB	
32	4207587983499043884	flt = flights.map(lambda c: (c[3], c[0])) air =... join at <command-3600302459314985>:3 +details	2018/02/03 20:25:18	5 s	4/4				9.7 MB

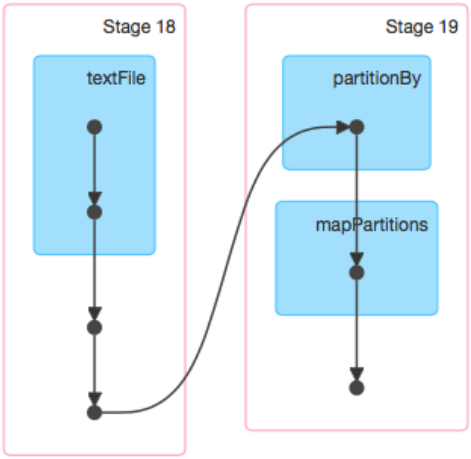

Details for Stage 32 (Attempt 0)
Total Time Across All Tasks: 11 s
Locality Level Summary: Process local: 4
Shuffle Write: 9.7 MB / 64
DAG Visualization
Stage 32
textFile
/databricks-datasets/flights/departuredelays.csv [12]
textFile at NativeMethodAccessorImpl.java:0
/databricks-datasets/flights/departuredelays.csv [13]
textFile at NativeMethodAccessorImpl.java:0
textFile
/databricks-datasets/flights/airport-codes-na.txt [9]
textFile at NativeMethodAccessorImpl.java:0
/databricks-datasets/flights/airport-codes-na.txt [10]
textFile at NativeMethodAccessorImpl.java:0
PythonRDD [69]
RDD at PythonRDD.scala:50
PythonRDD [70]
RDD at PythonRDD.scala:50
union
UnionRDD [71]
union at NativeMethodAccessorImpl.java:0
PythonRDD [72]
join at <command-3600302459314985>:3
PairwiseRDD [73]
join at <command-3600302459314985>:3
Show Additional Metrics
Event Timeline
Enable zooming
Scheduler Delay
Task Deserialization Time
Shuffle Read Time
Executor Computing Time
Shuffle Write Time
Result Serialization Time
Getting Result Time
Task 0 (attempt 0)
Status: SUCCESS
Launch Time: 2018/02/03 20:25:18
Finish Time: 2018/02/03 20:25:23
Scheduler Delay: 2 ms
Task Deserialization Time: 4 ms
Shuffle Read Time: 0 ms
Executor Computing Time: 5 s
Shuffle Write Time: 11 ms
Result Serialization Time: 0 ms
Getting Result Time: 0 ms
driver / localhost
19
20
21
22
23
3 February 20:25


Details for Stage 33 (Attempt 0)
Total Time Across All Tasks: 0.1 s
Locality Level Summary: Process local: 1
Shuffle Read: 2.2 MB / 16
DAG Visualization
Stage 33
partitionBy
ShuffledRDD [74]
partitionBy at NativeMethodAccessorImpl.java:0
mapPartitions
MapPartitionsRDD [75]
mapPartitions at PythonRDD.scala:145
PythonRDD [76]
RDD at PythonRDD.scala:50
Show Additional Metrics
Event Timeline
Enable zooming
Scheduler Delay
Task Deserialization Time
Shuffle Read Time
Executor Computing Time
Shuffle Write Time
Result Serialization Time
Getting Result Time
driver / localhost
920
930
940
950
960
970
980
990
000
010
020
030
040
050
20:25:23
20:25:24

Details for Job 14

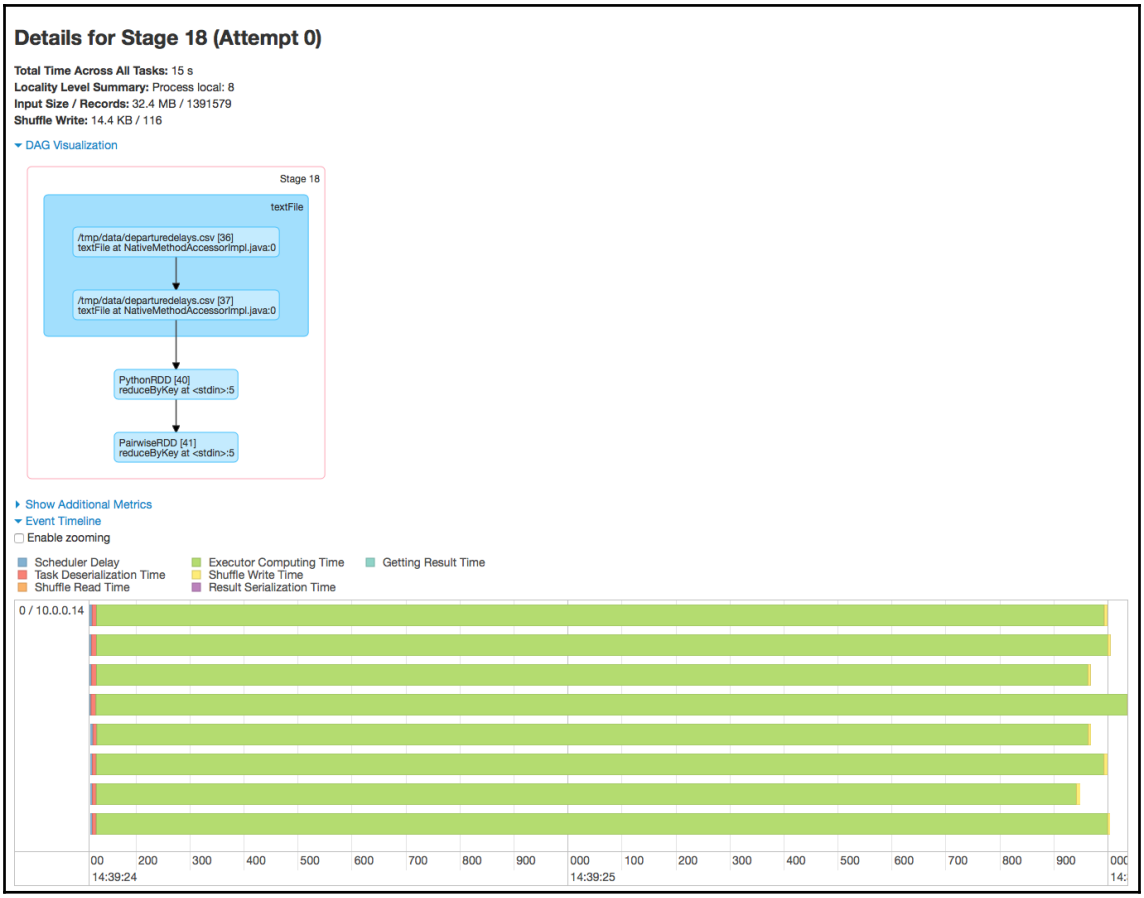
Status: SUCCEEDED
Completed Stages: 2

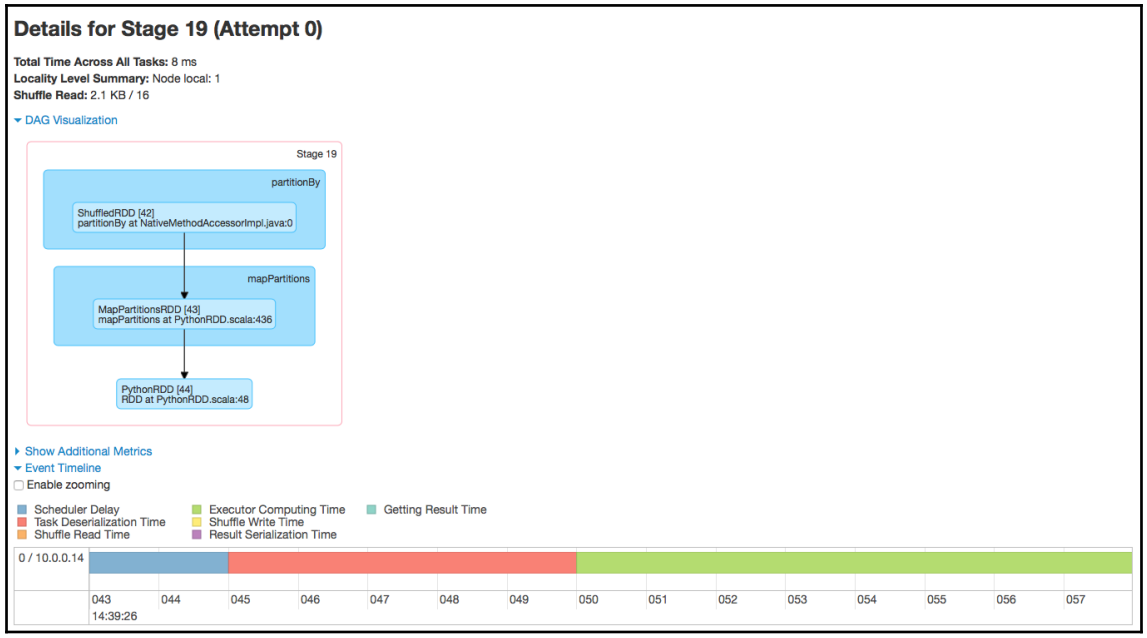
▸ Event Timeline
▾ DAG Visualization

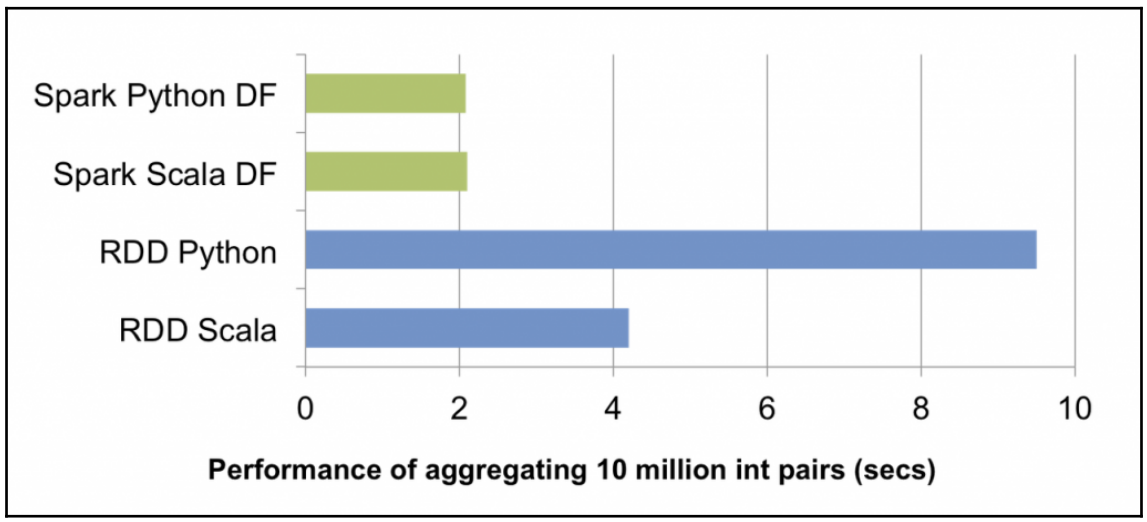


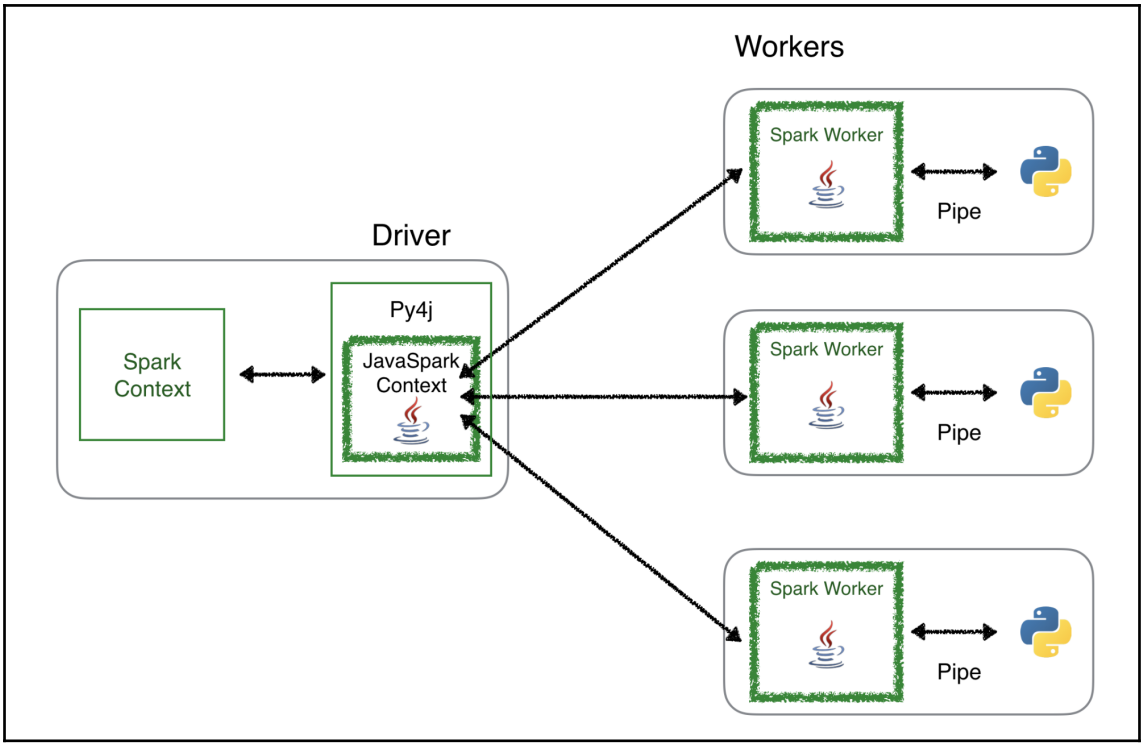
Completed Stages (2)

Stage Id ▾	Description	Submitted	Duration	Tasks: Succeeded/Total
19	runJob at PythonRDD.scala:455 +details	2018/02/05 14:39:26	16 ms	1/1
18	reduceByKey at <stdin>:5 +details	2018/02/05 14:39:24	2 s	8/8


Details for Stage 18 (Attempt 0)
Total Time Across All Tasks: 15 s
Locality Level Summary: Process local: 8
Input Size / Records: 32.4 MB / 1391579
Shuffle Write: 14.4 KB / 116
DAG Visualization
Stage 18
textFile
/tmp/data/departuredelays.csv [36]
textFile at NativeMethodAccessorImpl.java:0
/tmp/data/departuredelays.csv [37]
textFile at NativeMethodAccessorImpl.java:0
PythonRDD [40]
reduceByKey at <stdin>:5
PairwiseRDD [41]
reduceByKey at <stdin>:5
Show Additional Metrics
Event Timeline
Enable zooming
Scheduler Delay
Task Deserialization Time
Shuffle Read Time
Executor Computing Time
Shuffle Write Time
Result Serialization Time
Getting Result Time
0 / 10.0.0.14
14:39:24
14:39:25


Details for Stage 19 (Attempt 0)
Total Time Across All Tasks: 8 ms
Locality Level Summary: Node local: 1
Shuffle Read: 2.1 KB / 16
DAG Visualization
Stage 19
partitionBy
ShuffledRDD [42]
partitionBy at NativeMethodAccessorImpl.java:0
mapPartitions
MapPartitionsRDD [43]
mapPartitions at PythonRDD.scala:436
PythonRDD [44]
RDD at PythonRDD.scala:48
Show Additional Metrics
Event Timeline
Enable zooming
Scheduler Delay
Task Deserialization Time
Shuffle Read Time
Executor Computing Time
Shuffle Write Time
Result Serialization Time
Getting Result Time
0 / 10.0.0.14
043
14:39:26
044
045
046
047
048
049
050
051
052
053
054
055
056
057

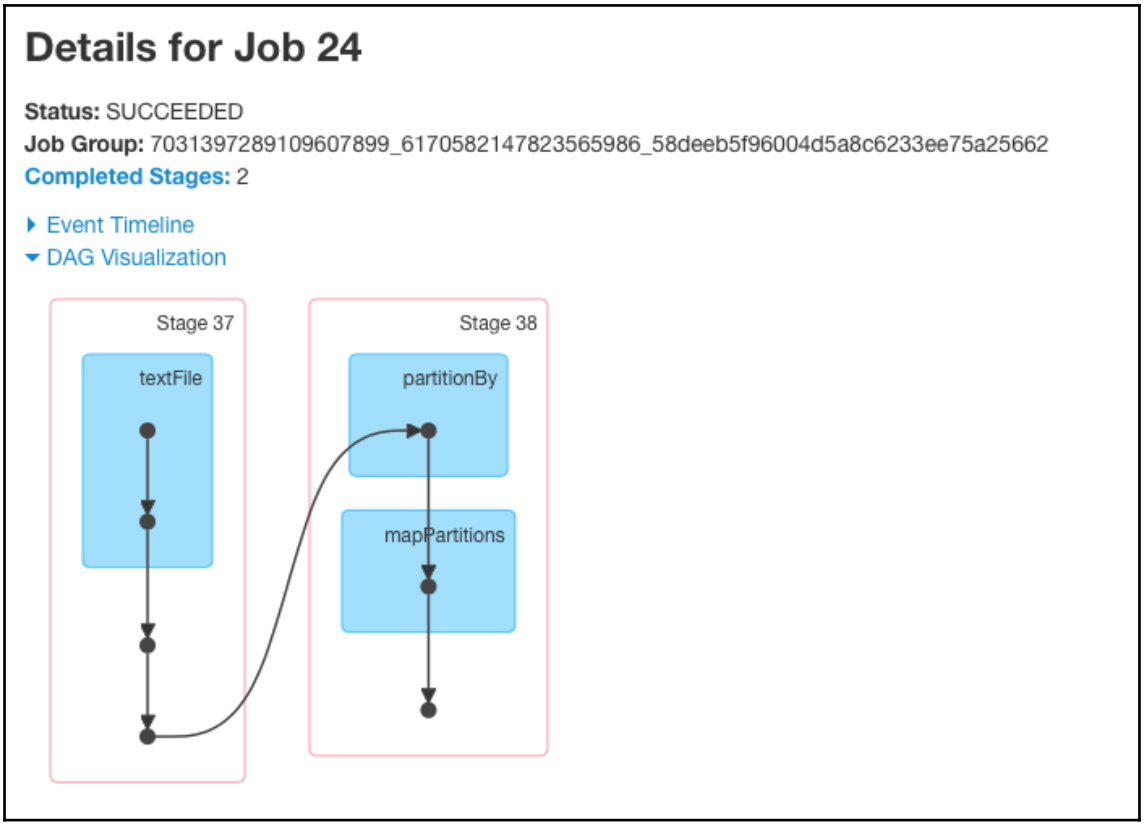

Spark Python DF
Spark Scala DF
RDD Python
RDD Scala
0
2
4
6
8
10
Performance of aggregating 10 million int pairs (secs)

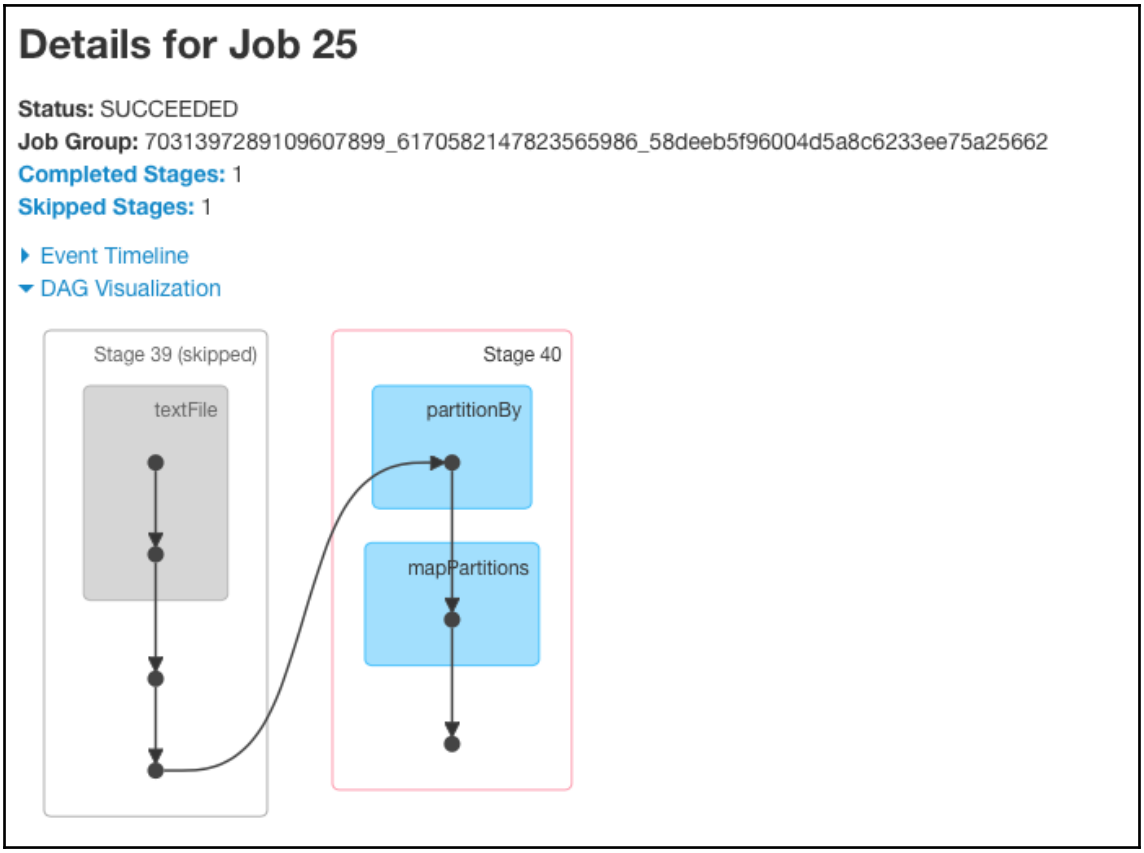



```
# How to do it
flights.map(lambda c: (c[3], int(c[1]))).reduceByKey(lambda x, y: x + y).sortByKey().take(50)
```

(3) Spark Jobs

- Job 24 View (Stages: 2/2)
- Job 25 View (Stages: 1/1, 1 skipped)
- Job 26 View (Stages: 2/2, 1 skipped)

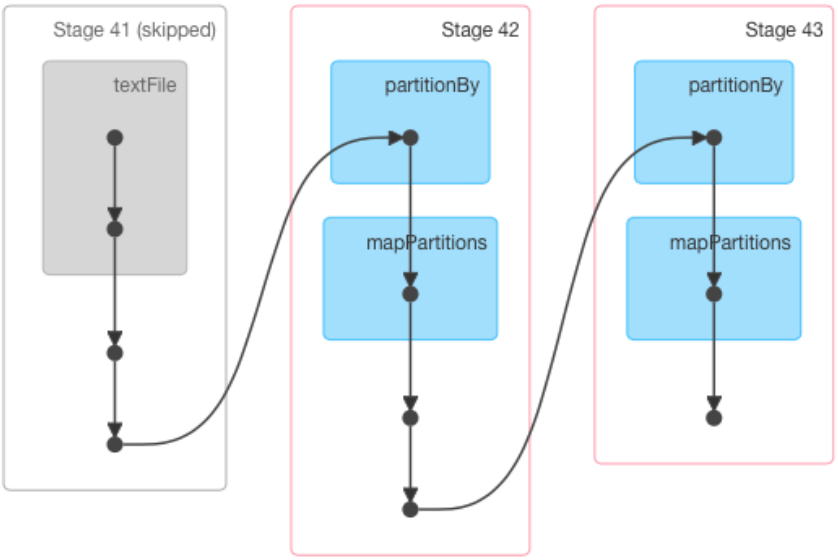

Details for Job 24
Status: SUCCEEDED
Job Group: 7031397289109607899_617058214782356598 6_58deeb5f96004d5a8c6233ee75a25662
Completed Stages: 2
Event Timeline
DAG Visualization
Stage 37
textFile
Stage 38
partitionBy
mapPartitions


Details for Job 25
Status: SUCCEEDED
Job Group: 7031397289109607899_6170582147823565986_58deeb5f96004d5a8c6233ee75a25662
Completed Stages: 1
Skipped Stages: 1
Event Timeline
DAG Visualization
Stage 39 (skipped)
textFile
Stage 40
partitionBy
mapPartitions

Details for Job 26

Status: SUCCEEDED
Job Group: 7031397289109607899_6170582147823565986_58deeb5f96004d5a8c6233ee75a25662
Completed Stages: 2
Skipped Stages: 1

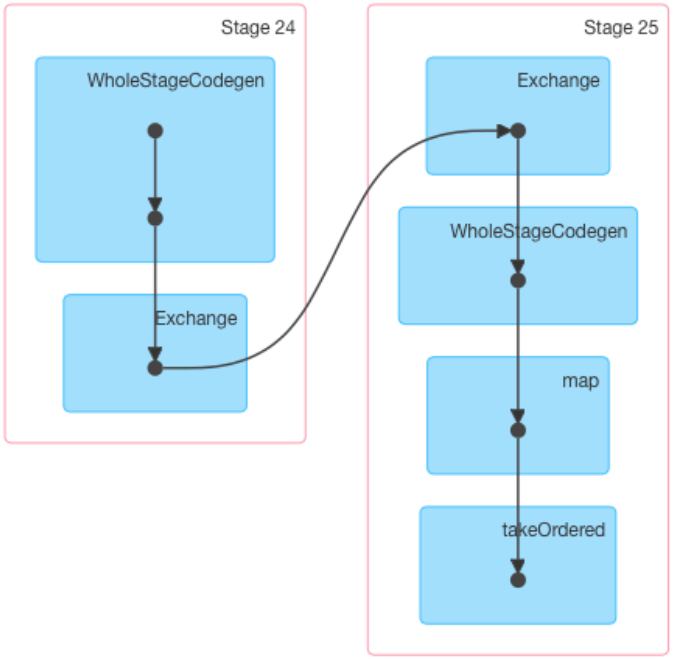
▸ Event Timeline
▾ DAG Visualization



Details for Job 18

Status: SUCCEEDED
Job Group: 7031397289109607899_8363433783550482188_001a3bc4a5164b2a9dbedee2c4947cfa
Completed Stages: 2

▸ Event Timeline
▾ DAG Visualization



Chapter 3: Abstracting Data with DataFrames

```
[(1, 'MacBook Pro', 2015, '15"', '16GB', '512GB SSD', 13.75, 9.48, 0.61, 4.02)]
```

```
[Row(Id=1, Model='MacBook Pro', Year=2015, ScreenSize='15"', RAM='16GB', HDD='512GB SSD', W=13.75, D=9.48, H=0.61, Wei
ght=4.02)]
```

```
+---+-----------+----+----------+----+---------+-----+----+----+------+------------+
| Id|      Model|Year|ScreenSize| RAM|      HDD|    W|   D|   H|Weight|    HDDSplit|
+---+-----------+----+----------+----+---------+-----+----+----+------+------------+
|  1|MacBook Pro|2015|     "15\""|16GB|512GB SSD|13.75|9.48|0.61|  4.02|[512GB, SSD]|
|  2|    MacBook|2016|     "12\""| 8GB|256GB SSD|11.04|7.74|0.52|  2.03|[256GB, SSD]|
|  3|MacBook Air|2016|   "13.3\""| 8GB|128GB SSD| 12.8|8.94|0.68|  2.96|[128GB, SSD]|
|  4|        iMac|2017|     "27\""|64GB|  1TB SSD| 25.6| 8.0|20.3|  20.8|  [1TB, SSD]|
+---+-----------+----+----------+----+---------+-----+----+----+------+------------+
```

```
root
 |--  D: double (nullable = true)
 |--  H: double (nullable = true)
 |--  HDD: string (nullable = true)
 |--  Model: string (nullable = true)
 |--  RAM: string (nullable = true)
 |--  ScreenSize: string (nullable = true)
 |--  W: double (nullable = true)
 |--  Weight: double (nullable = true)
 |--  Year: long (nullable = true)
 |-- Id: long (nullable = true)
```

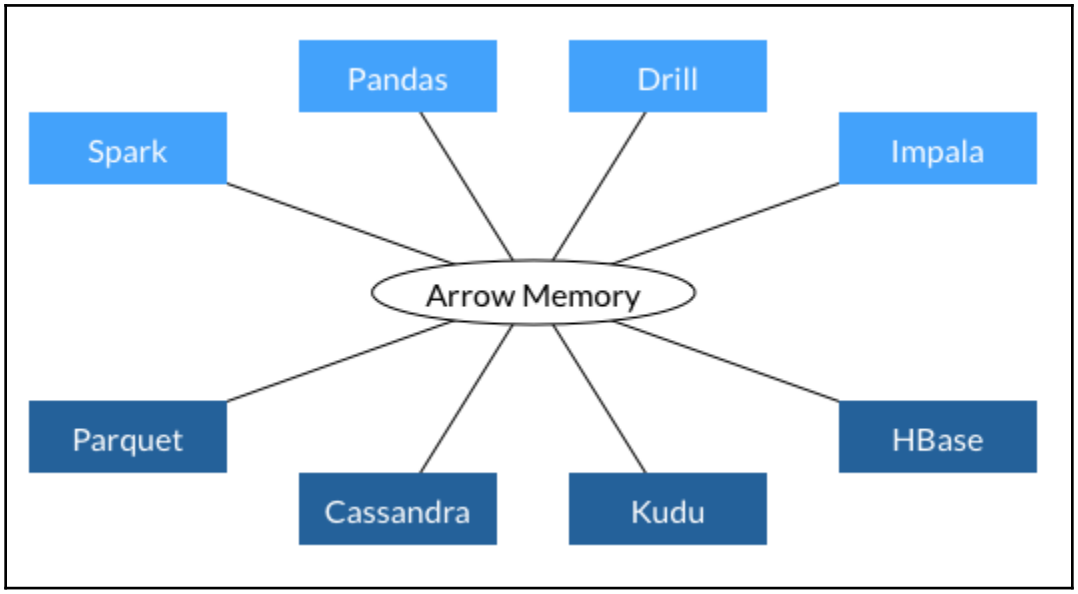
```
+----+----+---------+-----------+----+-----------+-----+-------+-----+---+
|   D|   H|      HDD|      Model| RAM| ScreenSize|    W| Weight| Year| Id|
+----+----+---------+-----------+----+-----------+-----+-------+-----+---+
|9.48|0.61|512GB SSD|MacBook Pro|16GB|        15"|13.75|   4.02| 2015|  1|
|7.74|0.52|256GB SSD|    MacBook| 8GB|        12"|11.04|   2.03| 2016|  2|
|8.94|0.68|128GB SSD|MacBook Air| 8GB|      13.3"| 12.8|   2.96| 2016|  3|
| 8.0|20.3|  1TB SSD|       iMac|64GB|        27"| 25.6|   20.8| 2017|  4|
+----+----+---------+-----------+----+-----------+-----+-------+-----+---+
```

```
root
 |--  D: double (nullable = true)
 |--  H: double (nullable = true)
 |--  HDD: string (nullable = true)
 |--  Model: string (nullable = true)
 |--  RAM: string (nullable = true)
 |--  ScreenSize: string (nullable = true)
 |--  W: double (nullable = true)
 |--  Weight: double (nullable = true)
 |--  Year: long (nullable = true)
 |-- Id: long (nullable = true)
```

```
[Row(Id=1, Model='MacBook Pro', Year=2015, ScreenSize='15"', RAM='16GB', HDD='512GB SSD', W=13.75, D=9.48, H=0.61, Weight=4.02)]
```

```
[(1, 'MacBook Pro', 2015, '15"', '16GB', '512GB SSD', 13.75, 9.48, 0.61, 4.02)]
```

```
+---+-----------+----+----------+----+---------+-----+----+----+------+--------+--------+-----------+
| Id|      Model|Year|ScreenSize| RAM|      HDD|    W|   D|   H|Weight|HDD_size|HDD_type|Volume_cuIn|
+---+-----------+----+----------+----+---------+-----+----+----+------+--------+--------+-----------+
|  1|MacBook Pro|2015|       15"|16GB|512GB SSD|13.75|9.48|0.61|  4.02|   512GB|     SSD|       80.0|
|  2|    MacBook|2016|       12"| 8GB|256GB SSD|11.04|7.74|0.52|  2.03|   256GB|     SSD|       44.0|
|  3|MacBook Air|2016|     13.3"| 8GB|128GB SSD| 12.8|8.94|0.68|  2.96|   128GB|     SSD|       78.0|
|  4|       iMac|2017|       27"|64GB|  1TB SSD| 25.6| 8.0|20.3|  20.8|     1TB|     SSD|     4157.0|
+---+-----------+----+----------+----+---------+-----+----+----+------+--------+--------+-----------+
```




```
+---+--------------------+--------------------+
| id|                 val|         probability|
+---+--------------------+--------------------+
|  0|  0.9453946488613437| 0.2551703151423011|
|  1| 0.39388041568859766| 0.3691657960230967|
|  2|  0.1356767412456391| 0.3952872266471388|
|  3|0.05098508750393837|0.39842409615712904|
|  4|  0.5167556509690651|  0.349079100100191|
+---+--------------------+--------------------+
only showing top 5 rows
```

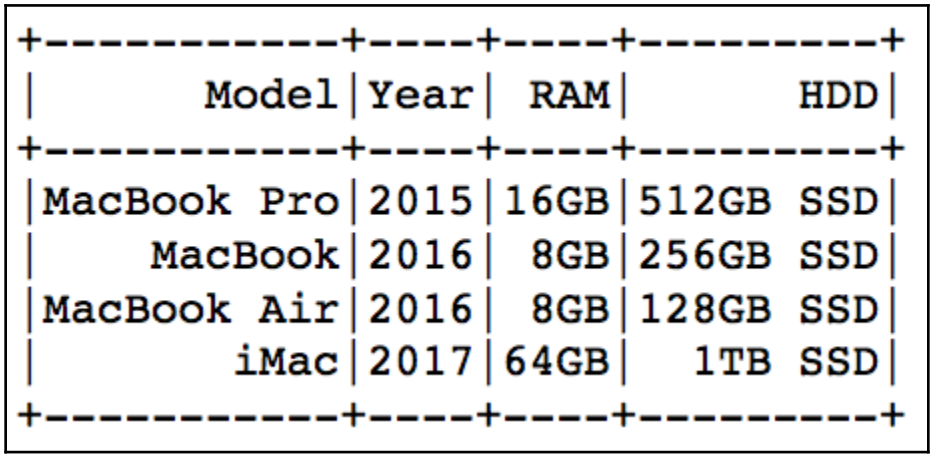
```
+------------------+
|count(probability)|
+------------------+
|           1000000|
+------------------+

23.1 s ± 937 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)
```

```
+------------------+
|count(probability)|
+------------------+
|           1000000|
+------------------+

23.1 s ± 937 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)
```

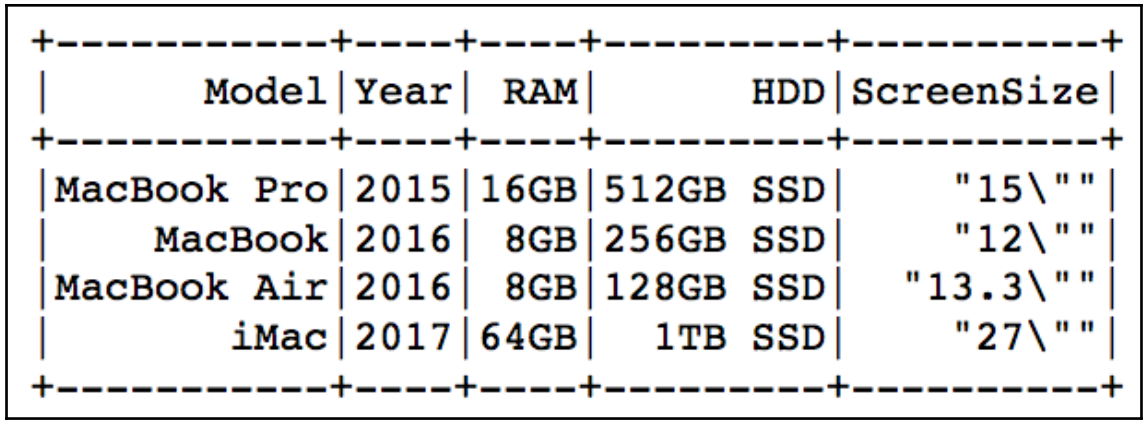
```
['Id, Model, Year, ScreenSize, RAM, HDD, W, D, H, Weight',
 '1,MacBook Pro,2015,"15\\"",16GB,512GB SSD,13.75,9.48,0.61,4.02']
```

```
+----+----+---------+---+----------+----+----------+-----+------+----+
|   D|   H|      HDD| Id|     Model| RAM|ScreenSize|    W|Weight|Year|
+----+----+---------+---+----------+----+----------+-----+------+----+
|9.48|0.61|512GB SSD|  1|MacBook Pro|16GB|     "15\""|13.75|  4.02|2015|
|7.74|0.52|256GB SSD|  2|    MacBook| 8GB|     "12\""|11.04|  2.03|2016|
|8.94|0.68|128GB SSD|  3|MacBook Air| 8GB|   "13.3\""| 12.8|  2.96|2016|
+----+----+---------+---+----------+----+----------+-----+------+----+
only showing top 3 rows
```

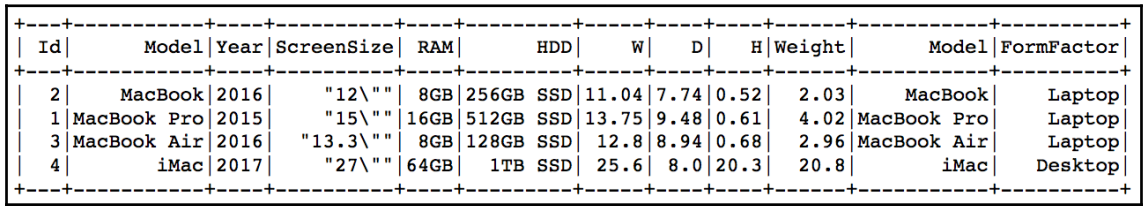
```
+---+----------+----+----------+----+---------+-----+----+----+------+
| Id|     Model|Year|ScreenSize| RAM|      HDD|    W|   D|   H|Weight|
+---+----------+----+----------+----+---------+-----+----+----+------+
|  1|MacBook Pro|2015|       15"|16GB|512GB SSD|13.75|9.48|0.61|  4.02|
|  2|    MacBook|2016|       12"| 8GB|256GB SSD|11.04|7.74|0.52|  2.03|
|  3|MacBook Air|2016|     13.3"| 8GB|128GB SSD| 12.8|8.94|0.68|  2.96|
|  4|       iMac|2017|       27"|64GB|  1TB SSD| 25.6| 8.0|20.3|  20.8|
+---+----------+----+----------+----+---------+-----+----+----+------+
```



Model	Year	RAM	HDD
MacBook Pro	2015	16GB	512GB SSD
MacBook	2016	8GB	256GB SSD
MacBook Air	2016	8GB	128GB SSD
iMac	2017	64GB	1TB SSD



Model	Year	RAM	HDD	ScreenSize
MacBook Pro	2015	16GB	512GB SSD	"15\""
MacBook	2016	8GB	256GB SSD	"12\""
MacBook Air	2016	8GB	128GB SSD	"13.3\""
iMac	2017	64GB	1TB SSD	"27\""

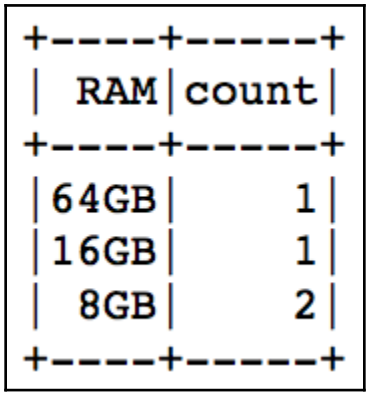


Id	Model	Year	ScreenSize	RAM	HDD	W	D	H	Weight	Model	FormFactor
2	MacBook	2016	"12\""	8GB	256GB SSD	11.04	7.74	0.52	2.03	MacBook	Laptop
1	MacBook Pro	2015	"15\""	16GB	512GB SSD	13.75	9.48	0.61	4.02	MacBook Pro	Laptop
3	MacBook Air	2016	"13.3\""	8GB	128GB SSD	12.8	8.94	0.68	2.96	MacBook Air	Laptop
4	iMac	2017	"27\""	64GB	1TB SSD	25.6	8.0	20.3	20.8	iMac	Desktop

FormFactor	ComputerCnt
Laptop	3
Desktop	1

Model	ScreenSize
MacBook Pro	"15\""
MacBook	"12\""
MacBook Air	"13.3\""
iMac	"27\""

Id	Model	Year	ScreenSize	RAM	HDD	W	D	H	Weight
2	MacBook	2016	"12\""	8GB	256GB SSD	11.04	7.74	0.52	2.03
3	MacBook Air	2016	"13.3\""	8GB	128GB SSD	12.8	8.94	0.68	2.96
4	iMac	2017	"27\""	64GB	1TB SSD	25.6	8.0	20.3	20.8



RAM	count
64GB	1
16GB	1
8GB	2

Id	Model	Year	ScreenSize	RAM	HDD	W	D	H	Weight
2	MacBook	2016	"12\""	8GB	256GB SSD	11.04	7.74	0.52	2.03
3	MacBook Air	2016	"13.3\""	8GB	128GB SSD	12.8	8.94	0.68	2.96
1	MacBook Pro	2015	"15\""	16GB	512GB SSD	13.75	9.48	0.61	4.02
4	iMac	2017	"27\""	64GB	1TB SSD	25.6	8.0	20.3	20.8

Id	Model	Year	ScreenSize	RAM	HDD	W	D	H	Weight
4	iMac	2017	"27\""	64GB	1TB SSD	25.6	8.0	20.3	20.8
3	MacBook Air	2016	"13.3\""	8GB	128GB SSD	12.8	8.94	0.68	2.96
1	MacBook Pro	2015	"15\""	16GB	512GB SSD	13.75	9.48	0.61	4.02
2	MacBook	2016	"12\""	8GB	256GB SSD	11.04	7.74	0.52	2.03

Id	Model	Year	ScreenSize	RAM	HDD	W	D	H	Weight	HDDSplit
1	MacBook Pro	2015	"15\""	16GB	512GB SSD	13.75	9.48	0.61	4.02	[512GB, SSD]
2	MacBook	2016	"12\""	8GB	256GB SSD	11.04	7.74	0.52	2.03	[256GB, SSD]
3	MacBook Air	2016	"13.3\""	8GB	128GB SSD	12.8	8.94	0.68	2.96	[128GB, SSD]
4	iMac	2017	"27\""	64GB	1TB SSD	25.6	8.0	20.3	20.8	[1TB, SSD]

```
+---+-----------+----+----------+----+---------+-----+----+----+------+-----------+----------+
| Id|      Model|Year|ScreenSize| RAM|      HDD|    W|   D|   H|Weight|      Model|FormFactor|
+---+-----------+----+----------+----+---------+-----+----+----+------+-----------+----------+
|  2|    MacBook|2016|     "12\""| 8GB|256GB SSD|11.04|7.74|0.52|  2.03|    MacBook|    Laptop|
|  1|MacBook Pro|2015|     "15\""|16GB|512GB SSD|13.75|9.48|0.61|  4.02|MacBook Pro|    Laptop|
|  3|MacBook Air|2016|   "13.3\""| 8GB|128GB SSD| 12.8|8.94|0.68|  2.96|MacBook Air|    Laptop|
|  4|       iMac|2017|     "27\""|64GB|  1TB SSD| 25.6| 8.0|20.3|  20.8|       iMac|   Desktop|
+---+-----------+----+----------+----+---------+-----+----+----+------+-----------+----------+
```

```
+---+-----------+----+----------+----+---------+-----+----+----+------+-----------+----------+
| Id|      Model|Year|ScreenSize| RAM|      HDD|    W|   D|   H|Weight|      Model|FormFactor|
+---+-----------+----+----------+----+---------+-----+----+----+------+-----------+----------+
|  2|    MacBook|2016|     "12\""| 8GB|256GB SSD|11.04|7.74|0.52|  2.03|       null|      null|
|  1|MacBook Pro|2015|     "15\""|16GB|512GB SSD|13.75|9.48|0.61|  4.02|MacBook Pro|    Laptop|
|  3|MacBook Air|2016|   "13.3\""| 8GB|128GB SSD| 12.8|8.94|0.68|  2.96|MacBook Air|    Laptop|
|  4|       iMac|2017|     "27\""|64GB|  1TB SSD| 25.6| 8.0|20.3|  20.8|       iMac|   Desktop|
+---+-----------+----+----------+----+---------+-----+----+----+------+-----------+----------+
```

```
+---+-----------+----+----------+----+---------+-----+----+----+------+-----------+----------+
| Id|      Model|Year|ScreenSize| RAM|      HDD|    W|   D|   H|Weight|      Model|FormFactor|
+---+-----------+----+----------+----+---------+-----+----+----+------+-----------+----------+
|  1|MacBook Pro|2015|     "15\""|16GB|512GB SSD|13.75|9.48|0.61|  4.02|MacBook Pro|    Laptop|
|  3|MacBook Air|2016|   "13.3\""| 8GB|128GB SSD| 12.8|8.94|0.68|  2.96|MacBook Air|    Laptop|
|  4|       iMac|2017|     "27\""|64GB|  1TB SSD| 25.6| 8.0|20.3|  20.8|       iMac|   Desktop|
+---+-----------+----+----------+----+---------+-----+----+----+------+-----------+----------+
```

```
+---+-----------+----+----------+----+---------+-----+----+----+------+
| Id|      Model|Year|ScreenSize| RAM|      HDD|    W|   D|   H|Weight|
+---+-----------+----+----------+----+---------+-----+----+----+------+
|  1|MacBook Pro|2015|     "15\""|16GB|512GB SSD|13.75|9.48|0.61|  4.02|
|  3|MacBook Air|2016|   "13.3\""| 8GB|128GB SSD| 12.8|8.94|0.68|  2.96|
|  4|       iMac|2017|     "27\""|64GB|  1TB SSD| 25.6| 8.0|20.3|  20.8|
+---+-----------+----+----------+----+---------+-----+----+----+------+
```

```
+---+-------+----+----------+---+---------+-----+----+----+------+
| Id|  Model|Year|ScreenSize|RAM|      HDD|    W|   D|   H|Weight|
+---+-------+----+----------+---+---------+-----+----+----+------+
|  2|MacBook|2016|     "12\""|8GB|256GB SSD|11.04|7.74|0.52|  2.03|
+---+-------+----+----------+---+---------+-----+----+----+------+
```

Id	Model	Year	ScreenSize	RAM	HDD	W	D	H	Weight
1	MacBook Pro	2015	"15\""	16GB	512GB SSD	13.75	9.48	0.61	4.02
2	MacBook	2016	"12\""	8GB	256GB SSD	11.04	7.74	0.52	2.03
3	MacBook Air	2016	"13.3\""	8GB	128GB SSD	12.8	8.94	0.68	2.96
4	iMac	2017	"27\""	64GB	1TB SSD	25.6	8.0	20.3	20.8
5	MacBook Pro	2018	15"	16GB	256GB SSD	13.75	9.48	0.61	4.02

RAM
64GB
16GB
8GB

A	B	C
21.4	36.3	24.2
1.6	32.1	27.9
3.2	38.7	24.7
2.8	21.4	23.9
3.9	34.1	27.9
9.2	21.4	21.4

```
+----+------------------+-----+
|   A|                 B|    C|
+----+------------------+-----+
|4.14|              36.3| 24.2|
| 1.6|              32.1| 27.9|
| 3.2|              38.7| 24.7|
| 2.8|35.300000000000004| 23.9|
| 3.9|              34.1| 27.9|
| 9.2|35.300000000000004|25.72|
+----+------------------+-----+
```

```
+---+----+----+
|  A|   B|   C|
+---+----+----+
|1.6|32.1|27.9|
|3.9|34.1|27.9|
|3.2|38.7|24.7|
+---+----+----+
```

```
+----+----+----+
|   A|   B|   C|
+----+----+----+
|null|36.3|24.2|
| 1.6|32.1|27.9|
| 3.2|38.7|24.7|
| 2.8|null|23.9|
| 3.9|34.1|27.9|
+----+----+----+
```

A	B	C
1.6	32.1	27.9
3.2	38.7	24.7
3.9	34.1	27.9

```
+-------+------------------+
|summary|                 W|
+-------+------------------+
|  count|                 4|
|   mean|15.797500000000001|
| stddev| 6.630738395281983|
|    min|             11.04|
|    25%|             11.04|
|    50%|              12.8|
|    75%|             13.75|
|    max|              25.6|
+-------+------------------+


+-------+------------------+
|summary|                 W|
+-------+------------------+
|  count|                 4|
|   mean|15.797500000000001|
| stddev| 6.630738395281983|
|    min|             11.04|
|    max|              25.6|
+-------+------------------+
```

```
+-------+------------------+
|summary|                 W|
+-------+------------------+
|  count|                 4|
|   mean|15.797500000000001|
| stddev| 6.630738395281983|
|    min|             11.04|
|    25%|             11.04|
|    50%|              12.8|
|    75%|             13.75|
|    max|              25.6|
+-------+------------------+

+-------+------------------+
|summary|                 W|
+-------+------------------+
|  count|                 4|
|   mean|15.797500000000001|
| stddev| 6.630738395281983|
|    min|             11.04|
|    max|              25.6|
+-------+------------------+
```

```
+-------+------------------+
|summary|                 W|
+-------+------------------+
|  count|                 4|
|   mean|15.797500000000001|
| stddev| 6.630738395281983|
|    min|             11.04|
|    max|              25.6|
+-------+------------------+
```

```
[Row(Year=2015, count=1), Row(Year=2016, count=2), Row(Year=2017, count=1)]
```

```
[Row(Id=1, Model='MacBook Pro', Year=2015, ScreenSize='"15\\""', RAM='16GB', HDD='512GB SSD', W=13.75, D=9.48, H=0.61,
Weight=4.02),
 Row(Id=2, Model='MacBook', Year=2016, ScreenSize='"12\\""', RAM='8GB', HDD='256GB SSD', W=11.04, D=7.74, H=0.52, Weig
ht=2.03)]
```

	Id	Model	Year	ScreenSize	RAM	HDD	W	D	H	Weight
0	1	MacBook Pro	2015	"15\""	16GB	512GB SSD	13.75	9.48	0.61	4.02
1	2	MacBook	2016	"12\""	8GB	256GB SSD	11.04	7.74	0.52	2.03
2	3	MacBook Air	2016	"13.3\""	8GB	128GB SSD	12.80	8.94	0.68	2.96
3	4	iMac	2017	"27\""	64GB	1TB SSD	25.60	8.00	20.30	20.80

Chapter 4: Preparing Data for Modeling

Id	Manufacturer	Model	EngineType	Displacement	Cylinders	FuelEconomy	MSRP	count
16	Toyota	CAMRY HYBRID LE	Aspirated	2.5	4	46	null	2

Manufacturer	Model	EngineType	Displacement	Cylinders	FuelEconomy	MSRP	count
BMW	440i Coupe	Turbo	3.0	6	23	null	2
Hyundai	G80 AWD	Turbo	3.3	6	20	null	2

```
In [9]: # count
        id_removed.count()

        19
```

CountOfIDs	CountOfDistinctIDs
19	18

Id	count
3	2

Id	Manufacturer	Model	EngineType	Displacement	Cylinders	FuelEconomy	MSRP
3	General Motors	SPARK ACTIV	Aspirated	1.4	null	32	null
3	Porsche	911 Carrera 4S Ca...	Turbo	3.0	6	24	null

Id	Manufacturer	Model	EngineType	Displacement	Cylinders	FuelEconomy	MSRP
8589934592	General Motors	SPARK ACTIV	Aspirated	1.4	null	32	null
188978561024	Mercedes-Benz	CLS 550	Turbo	4.7	8	21	79231
197568495616	Mercedes-Benz	null	null	null	null	27	null
206158430208	Ford Motor Company	FUSION AWD	Turbo	2.7	6	20	null
438086664192	BMW	COOPER S HARDTOP ...	Turbo	2.0	4	26	null
523986010112	Aston Martin	Vanquish	Aspirated	6.0	12	16	null
721554505728	Volkswagen	GTI	Turbo	2.0	4	null	null
764504178688	Kia	Stinger RWD	Turbo	2.0	4	25	null
919123001344	BMW	330i	Turbo	2.0	null	27	null
944892805120	Porsche	Boxster S	Turbo	2.5	4	22	null
970662608896	FCA US LLC	300	Aspirated	3.6	6	23	null
1030792151040	Hyundai	G80 AWD	Turbo	3.3	6	20	null
1039382085632	BMW	440i Coupe	Turbo	3.0	6	23	null
1116691496960	Nissan	Q50 AWD RED SPORT	Turbo	3.0	6	22	null
1211180777472	BMW	X5 M	Turbo	4.4	8	18	121231
1331439861760	Nissan	Q70 AWD	Aspirated	5.6	8	18	null
1606317768704	Porsche	911 Carrera 4S Ca...	Turbo	3.0	6	24	null
1614907703296	Toyota	CAMRY HYBRID LE	Aspirated	2.5	4	46	null
1700807049216	GE	K1500 SUBURBAN 4WD	Aspirated	5.3	8	18	null

Id	CountMissing
197568495616	5
8589934592	2
919123001344	2
721554505728	2

Id	Manufacturer	Model	EngineType	Displacement	Cylinders	FuelEconomy	MSRP
197568495616	Mercedes-Benz	null	null	null	null	27	null

```
+---+------------+-----+----------+------------+---------+-----------+----+
| Id|Manufacturer|Model|EngineType|Displacement|Cylinders|FuelEconomy|MSRP|
+---+------------+-----+----------+------------+---------+-----------+----+
+---+------------+-----+----------+------------+---------+-----------+----+
```

```
MSRP_miss 0.8888888888888888
Cylinders_miss 0.1111111111111116
FuelEconomy_miss 0.05555555555555558
EngineType_miss 0.0
Manufacturer_miss 0.0
Id_miss 0.0
Model_miss 0.0
Displacement_miss 0.0
```

```
{'FuelEconomy': 1.4957485048359973, 'Cylinders': 1.8353365984789105}
```

```
+------------+-----------------+--------------------+----------+------------+---------+------------------+
|          Id|     Manufacturer|               Model|EngineType|Displacement|Cylinders|       FuelEconomy|
+------------+-----------------+--------------------+----------+------------+---------+------------------+
|  8589934592|    General Motors|         SPARK ACTIV| Aspirated|         1.4|        2|    4.188095813552|
| 188978561024|     Mercedes-Benz|             CLS 550|    Turbo|         4.7|        8|              21.0|
| 206158430208|Ford Motor Company|          FUSION AWD|    Turbo|         2.7|        5|16.666666666666668|
| 438086664192|              BMW|COOPER S HARDTOP ...|    Turbo|         2.0|        4|              26.0|
| 523986010112|      Aston Martin|            Vanquish| Aspirated|         6.0|       12|              16.0|
| 721554505728|       Volkswagen|                 GTI|    Turbo|         2.0|        4|    11.96598803872|
| 764504178688|              Kia|          Stinger RWD|    Turbo|         2.0|        4|              25.0|
| 919123001344|              BMW|                330i|    Turbo|         2.0|        3| 8.974491029040001|
| 944892805120|          Porsche|            Boxster S|    Turbo|         2.5|        4|              22.0|
| 970662608896|       FCA US LLC|                 300| Aspirated|         3.6|        6|              23.0|
|1030792151040|          Hyundai|             G80 AWD|    Turbo|         3.3|        6|              20.0|
|1039382085632|              BMW|          440i Coupe|    Turbo|         3.0|        6|23.000000000000004|
|1116691496960|           Nissan|   Q50 AWD RED SPORT|    Turbo|         3.0|        6|21.999999999999996|
|1211180777472|              BMW|                X5 M|    Turbo|         4.4|        8|              18.0|
|1331439861760|           Nissan|             Q70 AWD| Aspirated|         5.6|        8|              18.0|
|1606317768704|          Porsche|911 Carrera 4S Ca...|    Turbo|         3.0|        6|              24.0|
|1614907703296|           Toyota|     CAMRY HYBRID LE| Aspirated|         2.5|        4|              46.0|
|1700807049216|               GE|  K1500 SUBURBAN 4WD| Aspirated|         5.3|        8|              18.0|
+------------+-----------------+--------------------+----------+------------+---------+------------------+
```

```
{   'Cylinders': [-2.0, 14.0],
    'Displacement': [-1.6000000000000005, 8.0],
    'FuelEconomy': [7.166666666666664, 32.50000000000001]}
```

id	Displacement_o	Cylinders_o	FuelEconomy_o
8589934592	false	false	true
188978561024	false	false	false
206158430208	false	false	false
438086664192	false	false	false
523986010112	false	false	false
721554505728	false	false	false
764504178688	false	false	false
919123001344	false	false	false
944892805120	false	false	false
970662608896	false	false	false
1030792151040	false	false	false
1039382085632	false	false	false
1116691496960	false	false	false
1211180777472	false	false	false
1331439861760	false	false	false
1606317768704	false	false	false
1614907703296	false	false	true
1700807049216	false	false	false

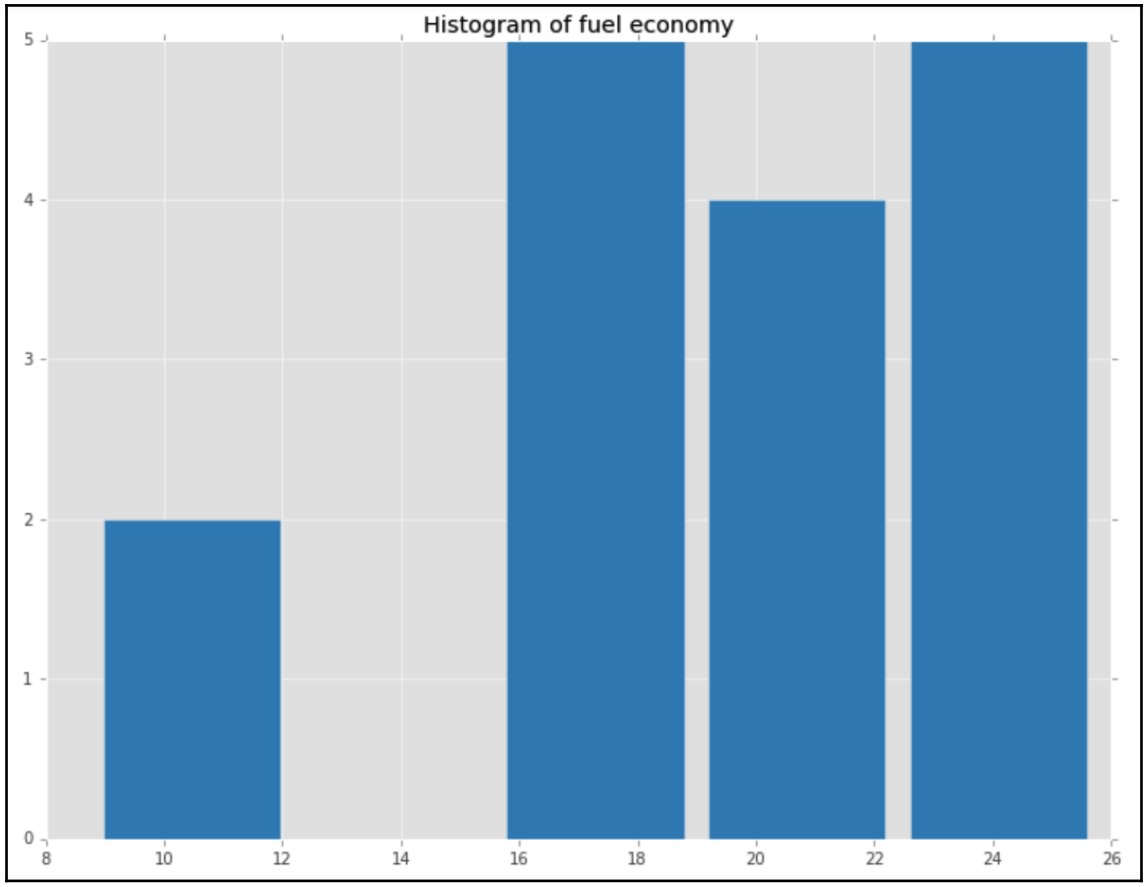
Id	Manufacturer	Model	FuelEconomy
8589934592	General Motors	SPARK ACTIV	4.188095813552
1614907703296	Toyota	CAMRY HYBRID LE	46.0

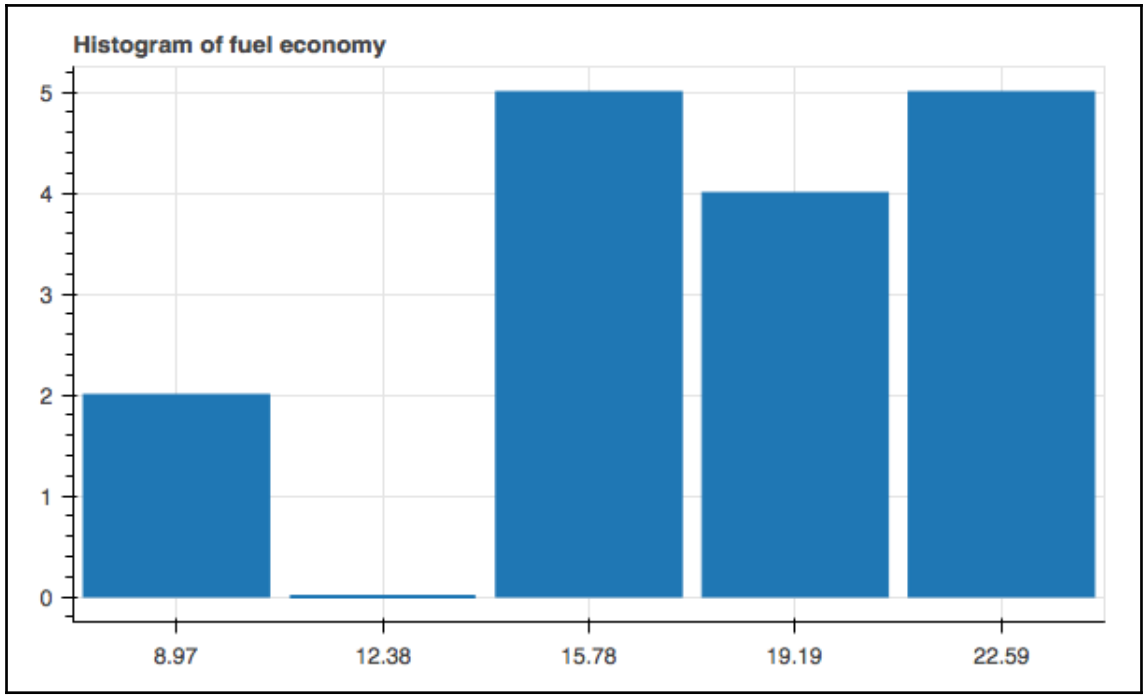
summary	Displacement	Cylinders	FuelEconomy
count	16	16	16
mean	3.44375	6.125	19.60044660840167
stddev	1.354975399530683	2.276693508870558	4.666647767366612
min	2.0	3	8.974491029040001
max	6.0	12	26.0

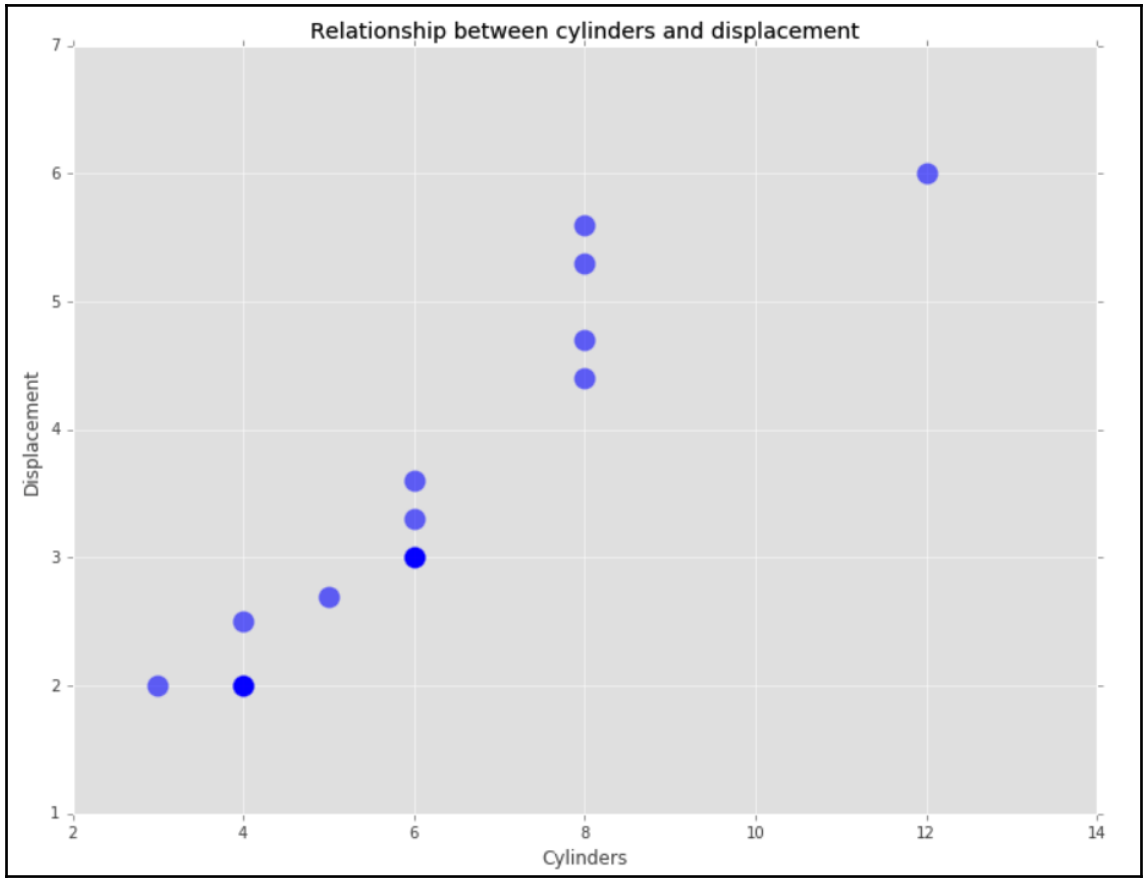
summary	Id	Manufacturer	Model	EngineType	Displacement	Cylinders	FuelEconomy
count	16	16	16	16	16	16	16
mean	9.19659872256E11	null	300.0	null	3.44375	6.125	19.60044660840167
stddev	4.396778949583304E11	null	NaN	null	1.354975399530683	2.276693508870558	4.666647767366612
min	188978561024	Aston Martin	300	Aspirated	2.0	3	8.974491029040001
max	1700807049216	Volkswagen	X5 M	Turbo	6.0	12	26.0

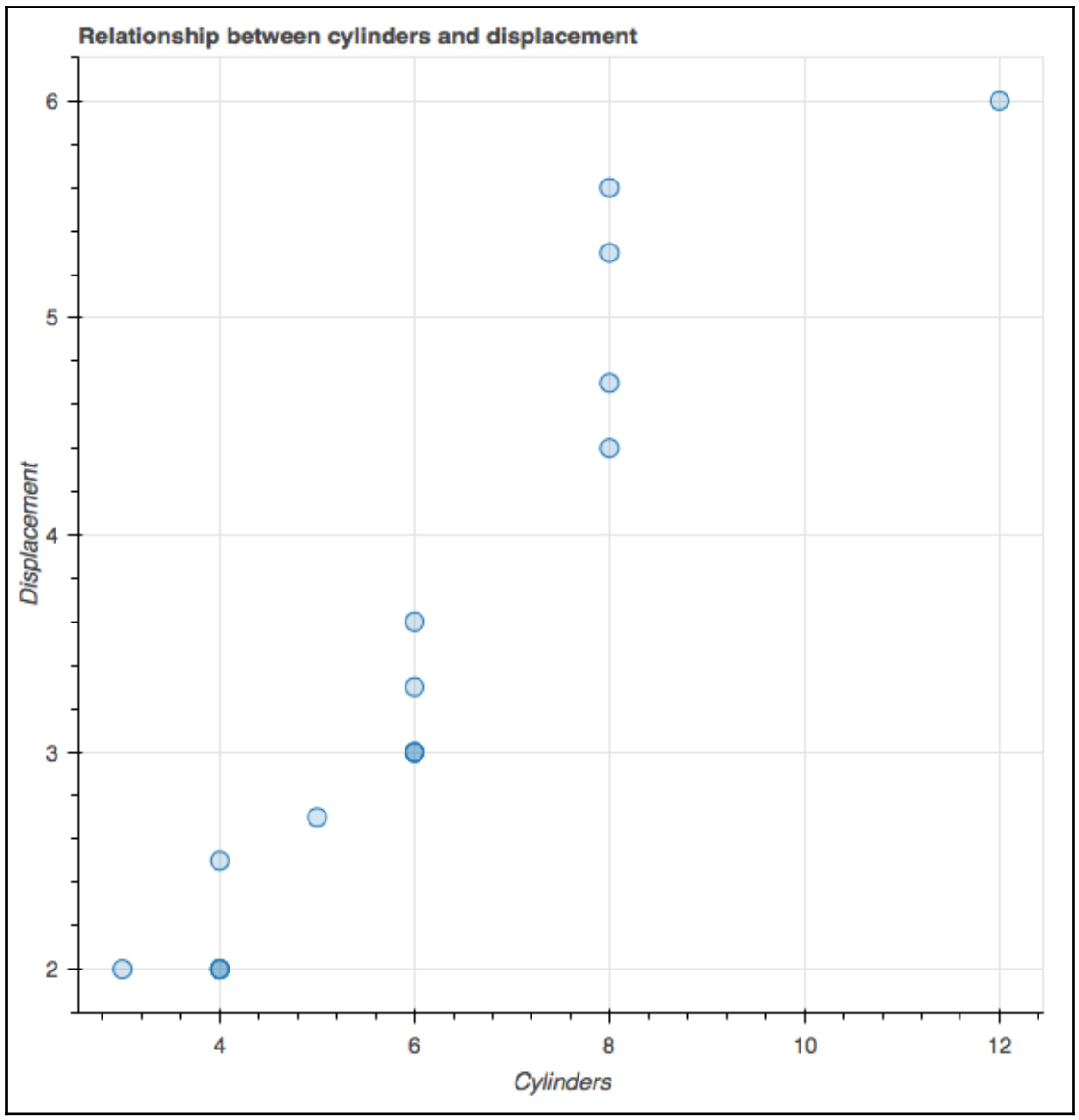
Cylinders	Count	MPG_avg	Disp_avg	MPG_stdev	Disp_stdev
3	1	8.974491029040001	2.0	NaN	NaN
4	4	21.24149700968	2.125	6.413009924983552	0.24999999999999994
5	1	16.666666666666668	2.7	NaN	NaN
6	5	22.4	3.1799999999999997	1.5165750888103104	0.26832815729997467
8	4	18.75	5.0	1.5	0.5477225575051655
12	1	16.0	6.0	NaN	NaN

Column	Displacement	Cylinders	FuelEconomy
Displacement	1.0	0.9381829964408113	-0.10757908872489412
Cylinders	null	1.0	-0.04218546545131555
FuelEconomy	null	null	1.0


Histogram of fuel economy
5
4
3
2
1
0
8
10
12
14
16
18
20
22
24
26


Histogram of fuel economy
5
4
3
2
1
0
8.97
12.38
15.78
19.19
22.59


Relationship between cylinders and displacement
Displacement
Cylinders
7
6
5
4
3
2
1
2
4
6
8
10
12
14

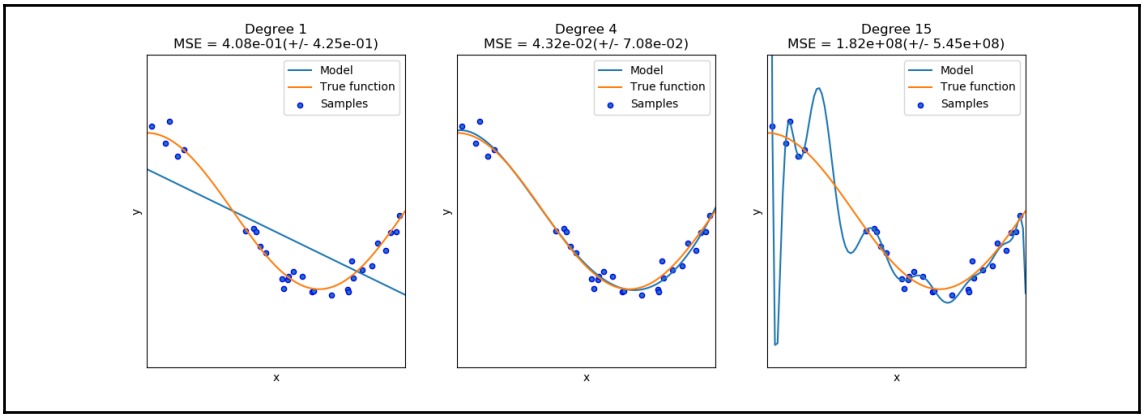

Relationship between cylinders and displacement
Displacement
Cylinders
2
3
4
5
6
4
6
8
10
12

```
root
 |-- age: integer (nullable = true)
 |-- workclass: string (nullable = true)
 |-- fnlwgt: integer (nullable = true)
 |-- education: string (nullable = true)
 |-- education-num: integer (nullable = true)
 |-- marital-status: string (nullable = true)
 |-- occupation: string (nullable = true)
 |-- relationship: string (nullable = true)
 |-- race: string (nullable = true)
 |-- sex: string (nullable = true)
 |-- capital-gain: integer (nullable = true)
 |-- capital-loss: integer (nullable = true)
 |-- hours-per-week: integer (nullable = true)
 |-- native-country: string (nullable = true)
 |-- label: string (nullable = true)
```

```
age: min->17.0, mean->38.6, max->90.0, stdev->13.6
capital-gain: min->0.0, mean->1077.6, max->99999.0, stdev->7385.3
capital-loss: min->0.0, mean->87.3, max->4356.0, stdev->403.0
hours-per-week: min->1.0, mean->40.4, max->99.0, stdev->12.3
```

```
sex [('Male', 21790), ('Female', 10771)]

race [('White', 27816), ('Black', 3124), ('Asian-Pac-Islander', 1039), ('Amer-Indian-Eskimo', 311)
, ('Other', 271)]

label [('<=50K', 24720), ('>50K', 7841)]

native-country [('United-States', 29170), ('Mexico', 643), ('?', 583), ('Philippines', 198), ('Ger
many', 137), ('Canada', 121), ('Puerto-Rico', 114), ('El-Salvador', 106), ('India', 100), ('Cuba',
95), ('England', 90), ('Jamaica', 81), ('South', 80), ('China', 75), ('Italy', 73), ('Dominican-Re
public', 70), ('Vietnam', 67), ('Guatemala', 64), ('Japan', 62), ('Poland', 60), ('Columbia', 59),
('Taiwan', 51), ('Haiti', 44), ('Iran', 43), ('Portugal', 37), ('Nicaragua', 34), ('Peru', 31), ('
France', 29), ('Greece', 29), ('Ecuador', 28), ('Ireland', 24), ('Hong', 20), ('Trinadad&Tobago',
19), ('Cambodia', 19), ('Laos', 18), ('Thailand', 18), ('Yugoslavia', 16), ('Outlying-US(Guam-USVI
-etc)', 14), ('Hungary', 13), ('Honduras', 13), ('Scotland', 12), ('Holand-Netherlands', 1)]
```

```
age
     capital-gain 0.077674498166
     capital-loss 0.057774539479
     hours-per-week 0.0687557075095

capital-gain
     age 0.077674498166
     hours-per-week 0.0784086153901

capital-loss
     age 0.057774539479
     hours-per-week 0.0542563622727

hours-per-week
     age 0.0687557075095
     capital-gain 0.0784086153901
     capital-loss 0.0542563622727
```



occupation	<=50K	>50K
Sales	2667	983
Exec-managerial	2098	1968
Prof-specialty	2281	1859
Handlers-cleaners	1284	86
Farming-fishing	879	115
Craft-repair	3170	929
Transport-moving	1277	320
Priv-house-serv	148	1
Protective-serv	438	211
Other-service	3158	137
Tech-support	645	283
Machine-op-inspct	1752	250
Armed-Forces	8	1
?	1652	191
Adm-clerical	3263	507

```
DenseMatrix(15, 2, [2667.0, 983.0, 2098.0, 1968.0, 2281.0, 1859.0, 1284.0, 86.0, ..., 1752.0, 250.0, 8.0, 1.0, 1652.0
, 191.0, 3263.0, 507.0], True)
```

```
array([[   2.66700000e+03,   9.83000000e+02],
       [   2.09800000e+03,   1.96800000e+03],
       [   2.28100000e+03,   1.85900000e+03],
       [   1.28400000e+03,   8.60000000e+01],
       [   8.79000000e+02,   1.15000000e+02],
       [   3.17000000e+03,   9.29000000e+02],
       [   1.27700000e+03,   3.20000000e+02],
       [   1.48000000e+02,   1.00000000e+00],
       [   4.38000000e+02,   2.11000000e+02],
       [   3.15800000e+03,   1.37000000e+02],
       [   6.45000000e+02,   2.83000000e+02],
       [   1.75200000e+03,   2.50000000e+02],
       [   8.00000000e+00,   1.00000000e+00],
       [   1.65200000e+03,   1.91000000e+02],
       [   3.26300000e+03,   5.07000000e+02]])
```

```
0.0
the occurrence of the outcomes is statistically independent.
```

```
{'sex': 2, 'race': 5, 'native-country': 42, 'marital-status': 7, 'workclass': 9, 'education': 16, 'occupation': 15, '
relationship': 6}
```

```
['label', 'age', 'capital-gain', 'capital-loss', 'hours-per-week', 'workclass', 'education', 'marital-status', 'occup
ation', 'relationship', 'race', 'sex', 'native-country']
```

```
[[['<=50K'], [39], [2174], [0], [40], [1.0, 2.0, 1.0, 5.0], [3.0, 3.0, 0.0, 0.0, 1.0, 0.0, 1.0, 1.0], [2.0, 3.0, 8.0]
, [0.0, 3.0, 3.0, 1.0, 4.0, 1.0, 0.0], [5.0, 5.0, 3.0], [3.0, 2.0], [4.0], [1.0, 0.0, 0.0, 3.0, 0.0, 1.0, 0.0, 0.0, 0
.0, 0.0, 0.0, 2.0, 0.0, 0.0, 0.0, 1.0, 1.0, 2.0, 1.0, 1.0, 0.0]], [['<=50K'], [50], [0], [0], [13], [4.0, 3.0, 1.0, 8
.0], [3.0, 3.0, 0.0, 0.0, 1.0, 0.0, 1.0, 1.0], [5.0, 5.0, 8.0], [0.0, 1.0, 2.0, 2.0, 8.0, 1.0, 1.0], [4.0, 2.0, 1.0],
[3.0, 2.0], [4.0], [1.0, 0.0, 0.0, 3.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2.0, 0.0, 0.0, 0.0, 1.0, 1.0, 2.0, 1.0, 1.
0, 0.0]], [['<=50K'], [38], [0], [0], [40], [2.0, 2.0, 0.0, 3.0], [2.0, 2.0, 0.0, 0.0, 0.0, 0.0, 2.0, 1.0], [3.0, 2.0
, 3.0], [2.0, 3.0, 1.0, 3.0, 7.0, 0.0, 1.0], [5.0, 5.0, 3.0], [3.0, 2.0], [4.0], [1.0, 0.0, 0.0, 3.0, 0.0, 1.0, 0.0,
0.0, 0.0, 0.0, 0.0, 2.0, 0.0, 0.0, 0.0, 1.0, 1.0, 2.0, 1.0, 1.0, 0.0]]]
```

```
[[0, 39, 2174, 0, 40, 1.0, 2.0, 1.0, 5.0, 3.0, 3.0, 0.0, 0.0, 1.0, 0.0, 1.0, 1.0, 2.0, 3.0, 8.0, 0.0, 3.0, 3.0, 1.0,
4.0, 1.0, 0.0, 5.0, 5.0, 3.0, 3.0, 2.0, 4.0, 1.0, 0.0, 0.0, 3.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2.0, 0.0, 0.0, 0.
0, 1.0, 1.0, 2.0, 1.0, 1.0, 0.0], [0, 50, 0, 0, 13, 4.0, 3.0, 1.0, 8.0, 3.0, 3.0, 0.0, 0.0, 1.0, 0.0, 1.0, 1.0, 5.0,
5.0, 8.0, 0.0, 1.0, 2.0, 2.0, 8.0, 1.0, 1.0, 4.0, 2.0, 1.0, 3.0, 2.0, 4.0, 1.0, 0.0, 0.0, 3.0, 0.0, 1.0, 0.0, 0.0, 0.
0, 0.0, 0.0, 2.0, 0.0, 0.0, 0.0, 1.0, 1.0, 2.0, 1.0, 1.0, 0.0], [0, 38, 0, 0, 40, 2.0, 2.0, 0.0, 3.0, 2.0, 2.0, 0.0,
0.0, 0.0, 0.0, 2.0, 1.0, 3.0, 2.0, 3.0, 2.0, 3.0, 1.0, 3.0, 7.0, 0.0, 1.0, 5.0, 5.0, 3.0, 3.0, 2.0, 4.0, 1.0, 0.0, 0.
0, 3.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2.0, 0.0, 0.0, 0.0, 1.0, 1.0, 2.0, 1.0, 1.0, 0.0]]
```

```
[(0, DenseVector([0.0307, 0.1485, -0.2167, -0.0354, -1.2635, 0.008, 1.7796, 1.0001, 0.83, 0.5743, -0.3473, -0.443, 0.
6826, -0.4007, -0.3862, -0.4685, -1.1369, -0.4555, 0.4551, -1.1329, 2.0776, 1.8713, -1.0381, -0.3381, -0.2381, -0.775
, 0.9805, 1.5207, 0.3083, -0.0634, -0.2574, -0.7031, 0.3208, -0.0901, -0.1263, 0.3355, -0.1223, 0.3378, -0.0853, -0.0
937, -0.0887, -0.2104, 0.0, 0.1286, -0.1976, -0.1433, -0.1419, 0.1895, 0.298, 0.2896, 0.1638, 0.1221, 0.0]))]
```

```
(DenseVector([38.5816, 1077.6488, 87.3038, 40.4375, 2.2307, 1.9942, 0.2391, 3.4004, 2.2001, 2.2522, 0.1836, 0.2528, 0
.5707, 0.1384, 1.2445, 1.5914, 3.6564, 3.6415, 7.1161, 1.0957, 0.838, 1.4236, 2.461, 4.7319, 1.223, 0.4287, 3.8206, 2
.8271, 2.472, 3.0508, 2.4882, 4.6616, 0.9066, 0.008, 0.0157, 2.7153, 0.0157, 0.8967, 0.0072, 0.0108, 0.0078, 0.0432,
0.0, 1.9416, 0.0445, 0.0238, 0.0197, 0.9444, 0.9136, 1.8534, 0.9618, 0.964, 0.0]), DenseVector([13.6404, 7385.2921, 4
02.9602, 12.3474, 0.974, 0.722, 0.4276, 1.5994, 0.9637, 1.3021, 0.5285, 0.5706, 0.6288, 0.3453, 0.6331, 1.2623, 1.457
, 1.4084, 1.9423, 0.9671, 1.0407, 0.8424, 1.4074, 2.1649, 0.9365, 0.5532, 1.2029, 1.4289, 1.7124, 0.8013, 1.8969, 0.9
41, 0.2911, 0.0893, 0.1243, 0.8486, 0.1281, 0.3057, 0.0846, 0.1151, 0.088, 0.2055, 0.0, 0.4543, 0.2251, 0.1658, 0.139
1, 0.2935, 0.29, 0.5062, 0.2331, 0.2945, 0.0]))
```

```
16.0 3.28395400384
40.0 45.7930482573
35.0 19.46708583
60.0 41.8838718837
52.0 11.0061417631
43.0 1.28423467636
35.0 9.96743282811
40.0 -9.36306561637
60.0 10.8259138175
40.0 1.09195397136
```

```
1.0 1
1.0 1
1.0 1
0.0 0
0.0 0
0.0 1
0.0 0
0.0 0
0.0 0
0.0 0
```

```
0.153472872815
0.122339061021
```

```
R^2:  -6.754451242767173
Explained Variance:  1145.7421086452416
meanAbsoluteError:  29.866629018908615
```

```
areaUnderPR:  0.7050195379236808
areaUnderROC:  0.7951114398791014
```

```
areaUnderPR:  0.691156113863938
areaUnderROC:  0.779415478708815 2
```

```
[((0.0, 0.0), 4967), ((1.0, 1.0), 2163), ((0.0, 1.0), 196), ((1.0, 0.0), 2410)]
```

```
[((0.0, 0.0), 4848), ((1.0, 1.0), 2127), ((0.0, 1.0), 232), ((1.0, 0.0), 2529)]
```

```
Training Error = 0.2676663927691043 7
```

```
Training Error = 0.2835866885784716 4
```

Chapter 6: Machine Learning with the ML Module

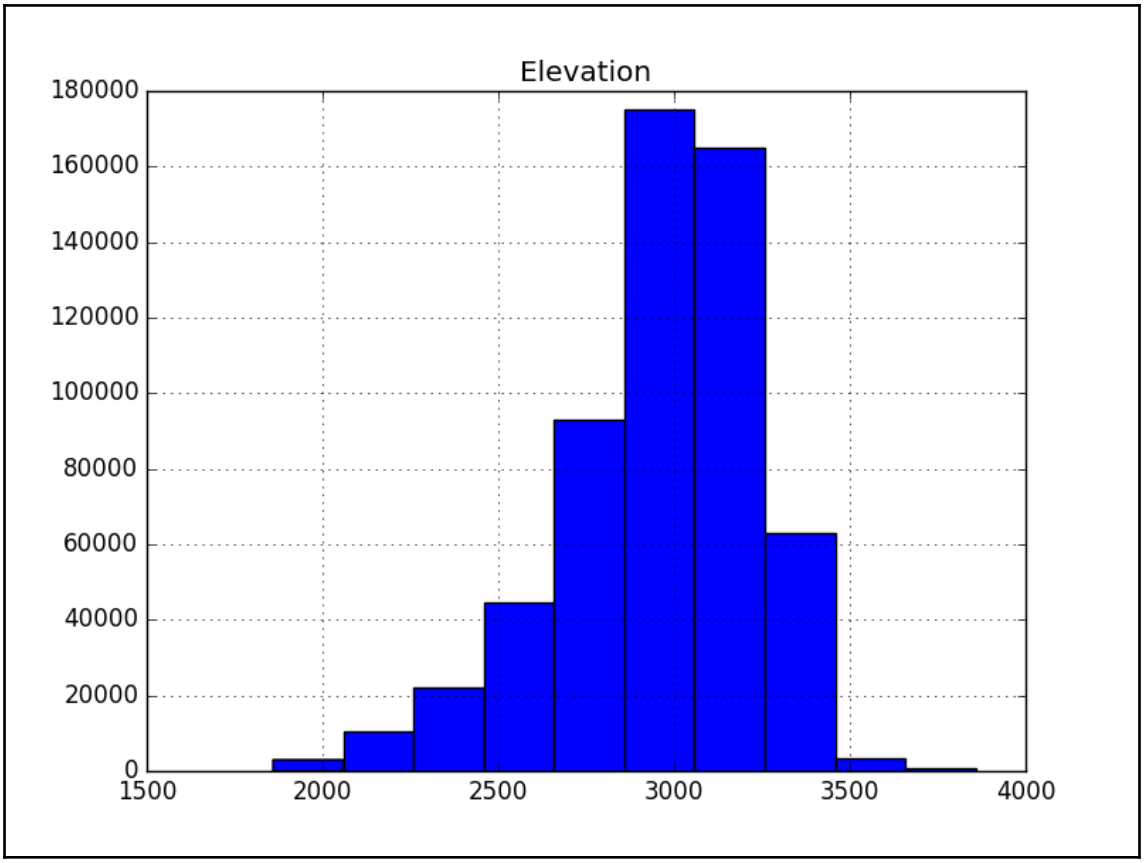
```
+-----------------------------------+-----------------------------------+
|Horizontal_Distance_To_Hydrology|Horizontal_Distance_To_Hydrology_Bkt|
+-----------------------------------+-----------------------------------+
|                              258|                               2.0|
|                              212|                               1.0|
|                              268|                               2.0|
|                              242|                               1.0|
|                              153|                               1.0|
+-----------------------------------+-----------------------------------+
```

```
[Row(feat=SparseVector(55, {0: 2596.0, 1: 51.0, 2: 3.0, 3: 258.0, 5: 510.0, 6: 221.0, 7: 232.0, 8: 148.0, 9: 6279.0, 1
0: 1.0, 42: 1.0, 54: 5.0}), pca_feat=DenseVector([-3887.7711, 4996.8103, 2323.0932, 1014.5873, -135.1702]))]
```

```
DenseVector([-0.0001, -0.0, -0.0023, -0.0, -0.0001, 0.0, -0.001, -0.0017, -0.0003, -0.0, 0.0, 0.0401, -0.0071, -0.0958
, -0.0901, -0.0653, -0.0655, -0.0437, -0.0928, -0.0848, -0.0211, -0.0045, -0.0498, -0.0829, -0.0522, -0.0325, -0.0263,
-0.0923, -0.0889, -0.0275, -0.0606, -0.0595, 0.0341, -0.003, 0.0822, 0.0607, 0.0351, 0.0093, 0.0048, -0.0154, 0.0422,
-0.0673, -0.0039, -0.0142, 0.0036, 0.0078, 0.0, -0.0117, 0.0283, -0.0002, -0.0463, 0.0394, 0.0292, 0.0358])
```

```
DenseVector([0.0309, 0.6522, 0.1911, 0.1424, 0.0342, 0.7402, 1.053, -0.0017, -0.0041, 2.7163, 189.0362, 27.8238, -265.
8505, -407.4379, -346.0612, -364.3841, -302.6788, -400.5852, -212.9918, -126.1329, -117.7423, -312.0478, -248.7118, -2
21.4788, -155.1459, -84.5129, -398.0433, -387.8102, -179.4485, -261.3875, -337.7875, 48.0629, -94.7813, 149.8043, 135.
144, 80.0901, 64.3659, 124.0233, -115.0126, 119.1285, -181.7498, 10.8056, -42.7849, 65.5441, 102.2562, 36.9865, -48.11
63, 379.2091, 256.0169, 497.1714, 313.0607, 337.172, 397.0758, -14.4551])
```

```
0.7860412464754236 129.50871925702438 103.34079732698483
```


Elevation
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
1500
2000
2500
3000
3500
4000

Elevation	prediction
2596	2840.7801831411316
2590	2828.7464246669683
2804	2842.761272955131
2785	2966.057500325109
2595	2817.1687155114637

selected
(10,[0,1,2,3,5,6,...
(10,[0,1,2,3,4,5,...
(10,[0,1,2,3,4,5,...
(10,[0,1,2,3,4,5,...
(10,[0,1,2,3,4,5,...

```
DenseMatrix([[ 1.        ,  0.01573494, -0.24269664, ...,  0.19359464,
               0.21261232, -0.26955378],
             [ 0.01573494,  1.        ,  0.07872841, ...,  0.00829428,
              -0.00586558,  0.0170798 ],
             [-0.24269664,  0.07872841,  1.        , ...,  0.09360193,
               0.02563691,  0.14828541],
             ...,
             [ 0.19359464,  0.00829428,  0.09360193, ...,  1.        ,
              -0.01929168,  0.15566826],
             [ 0.21261232, -0.00586558,  0.02563691, ..., -0.01929168,
               1.        ,  0.1283513 ],
             [-0.26955378,  0.0170798 ,  0.14828541, ...,  0.15566826,
               0.1283513 ,  1.        ]])
```

```
array(['Wilderness_Area_CacheLaPoudre', 'Soil_type_4703',
       'Horizontal_Distance_To_Roadways',
       'Horizontal_Distance_To_Hydrology', 'CoverType', 'Slope',
       'Wilderness_Area_Neota', 'Soil_type_8771', 'Soil_type_2717',
       'Soil_type_8776'], dtype='<U34')
```

```
(0.6638467009427569, 0.6632784396900246, 0.691296432850954)
```

```
(0.6638467009427569, 0.6632784396900246, 0.691296432850954)
```

```
0.826423672209303 4
```

```
0.833598109692272
```

```
+--------------------+---------+----------+
|            features|CoverType|prediction|
+--------------------+---------+----------+
|(54,[0,1,2,3,5,6,...|        5|         1|
|(54,[0,1,2,3,4,5,...|        5|         1|
|(54,[0,1,2,3,4,5,...|        2|         1|
|(54,[0,1,2,3,4,5,...|        2|         1|
|(54,[0,1,2,3,4,5,...|        5|         1|
+--------------------+---------+----------+
```

```
0.499982613164406 1
```

```
0.602428186128145 3
0.660204857590561 2
0.660204857590561 4
```

```
0.602428186128145 3
0.660204857590561 2
0.660204857590561 4
```

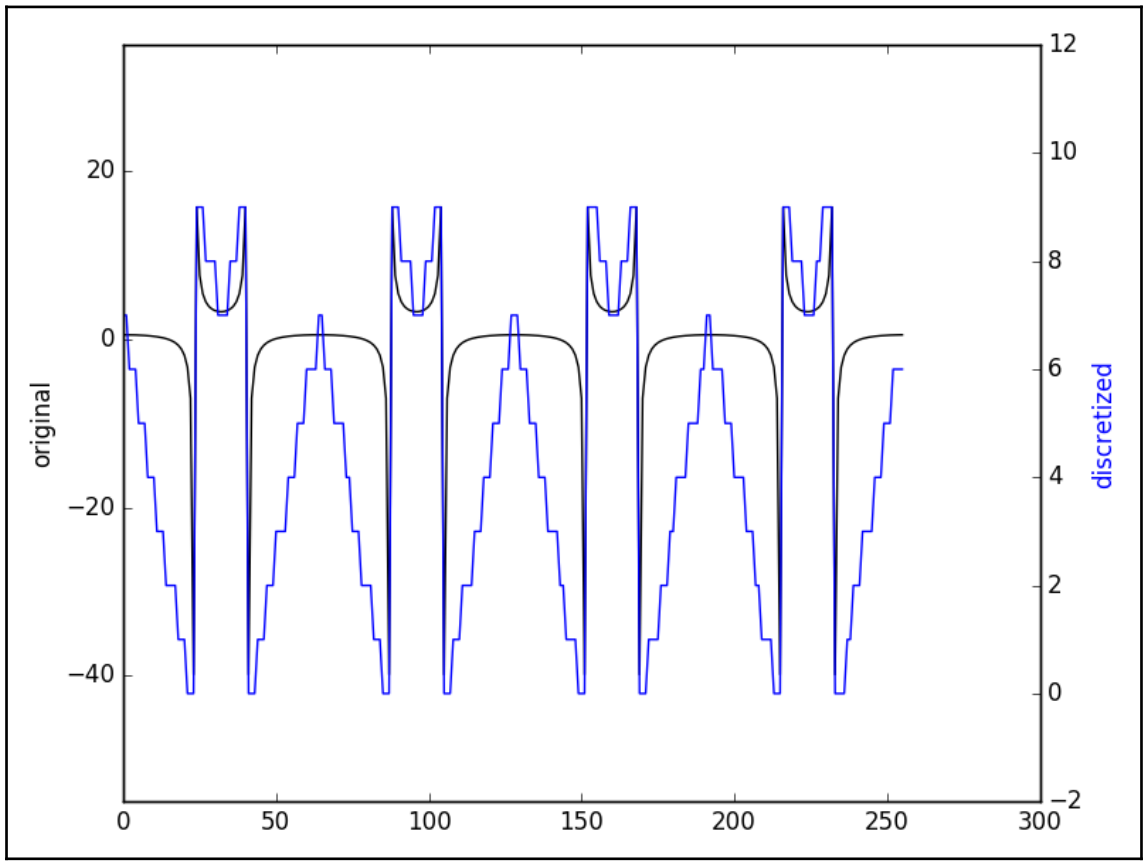
```
[Row(text_split=['apache', 'spark', 'achieves', 'high', 'performance', 'for', 'both', 'batch', 'and', 'streaming', 'da
ta', 'using', 'a', 'state-of-the-art', 'dag', 'scheduler', 'a', 'query', 'optimizer', 'and', 'a', 'physical', 'executi
on', 'engine'])]
```

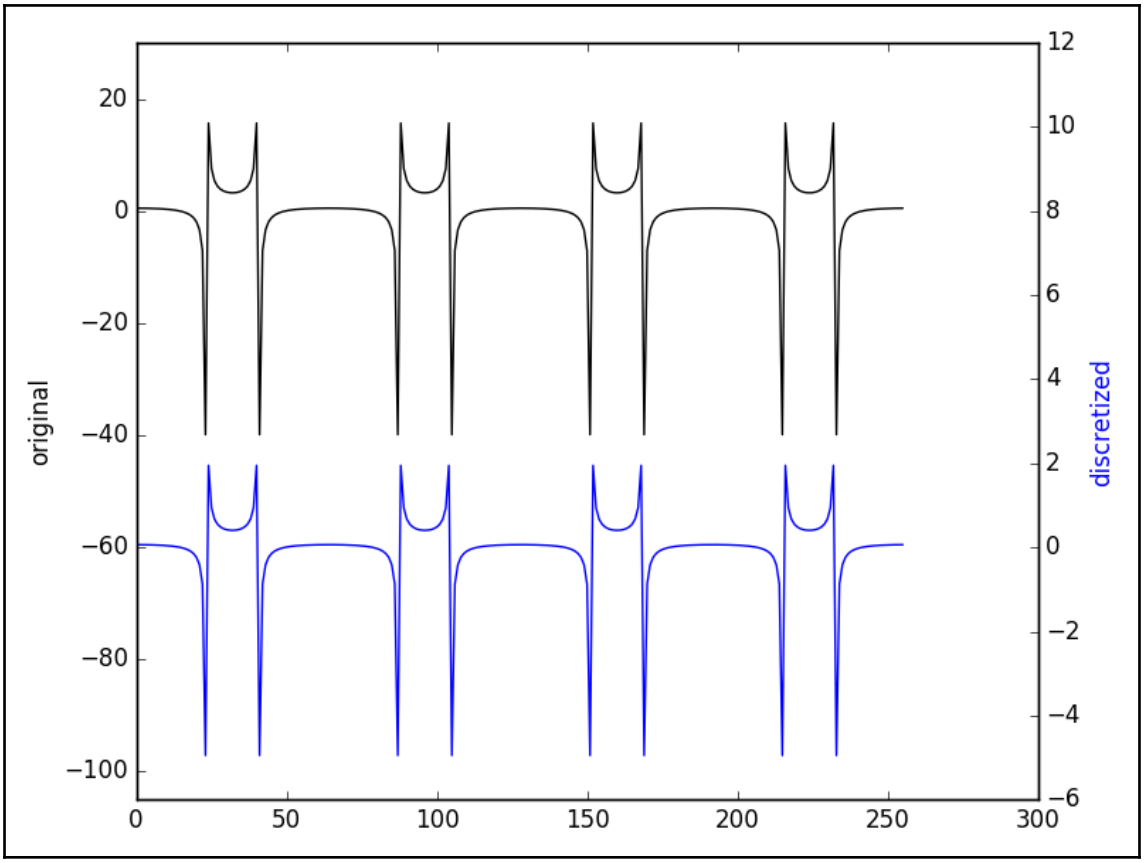
```
[Row(no_stopWords=['apache', 'spark', 'achieves', 'high', 'performance', 'batch', 'streaming', 'data', 'using', 'state
-of-the-art', 'dag', 'scheduler', 'query', 'optimizer', 'physical', 'execution', 'engine'])]
```

```
[Row(features=SparseVector(20, {2: 0.0, 3: 0.0, 4: 0.0, 5: 0.863, 8: 0.2877, 9: 0.0, 15: 0.0, 16: 0.6931, 18: 0.2877,
19: 0.0}))]
```

```
[Row(text='\n    Apache Spark achieves high performance for both batch\n    and streaming data, using a state-of-the-a
rt DAG scheduler, \n    a query optimizer, and a physical execution engine.\n    ', features=SparseVector(20, {2: 0.0,
3: 0.0, 4: 0.0, 5: 0.863, 8: 0.2877, 9: 0.0, 15: 0.0, 16: 0.6931, 18: 0.2877, 19: 0.0}))]
```

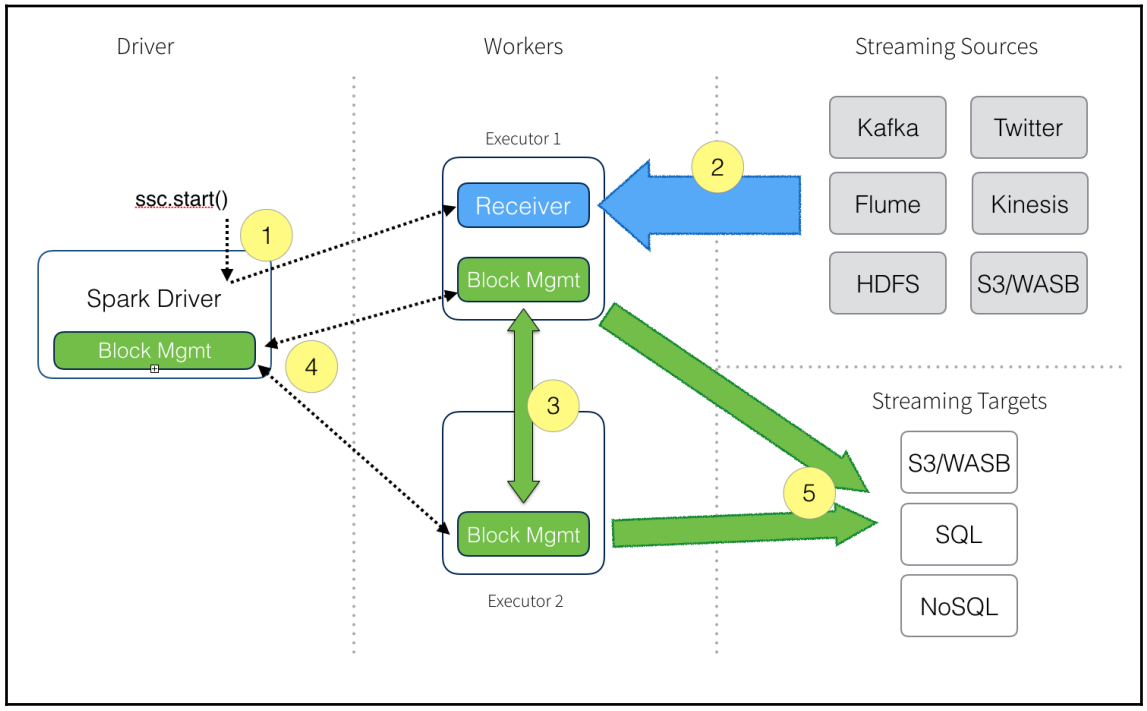
```
[Row(vector=DenseVector([0.0187, -0.0121, -0.0208, -0.0028, 0.002]))]
```

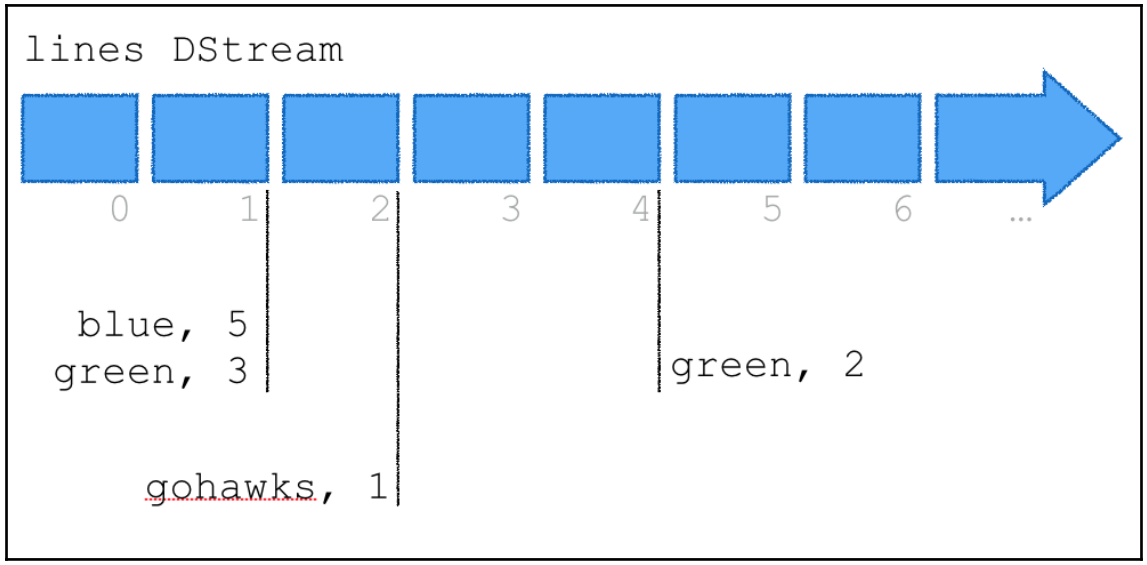


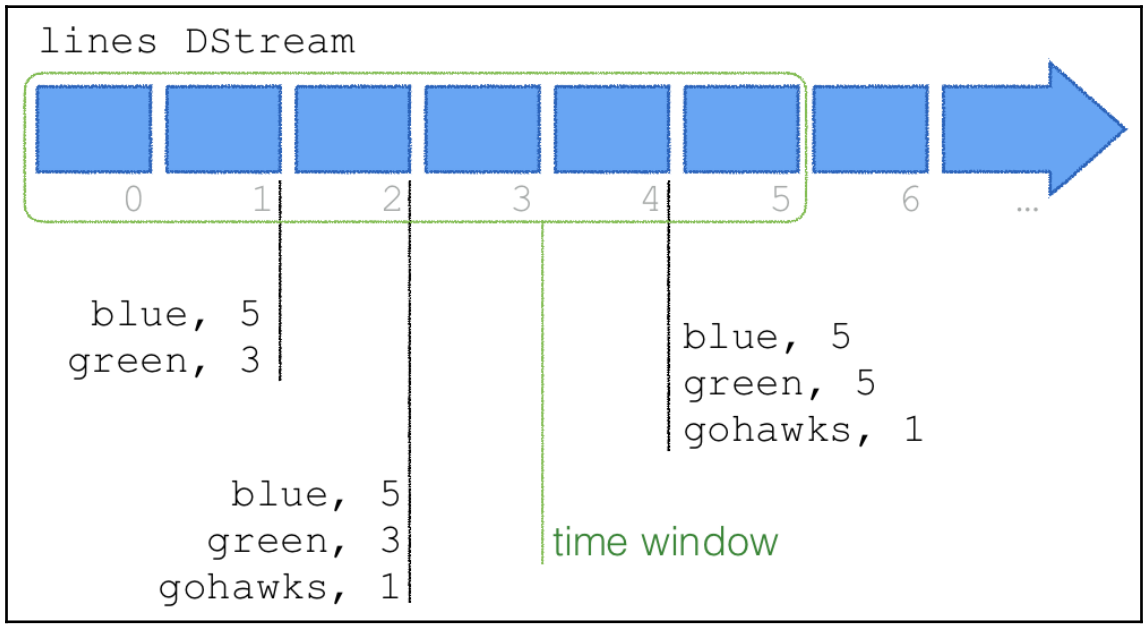


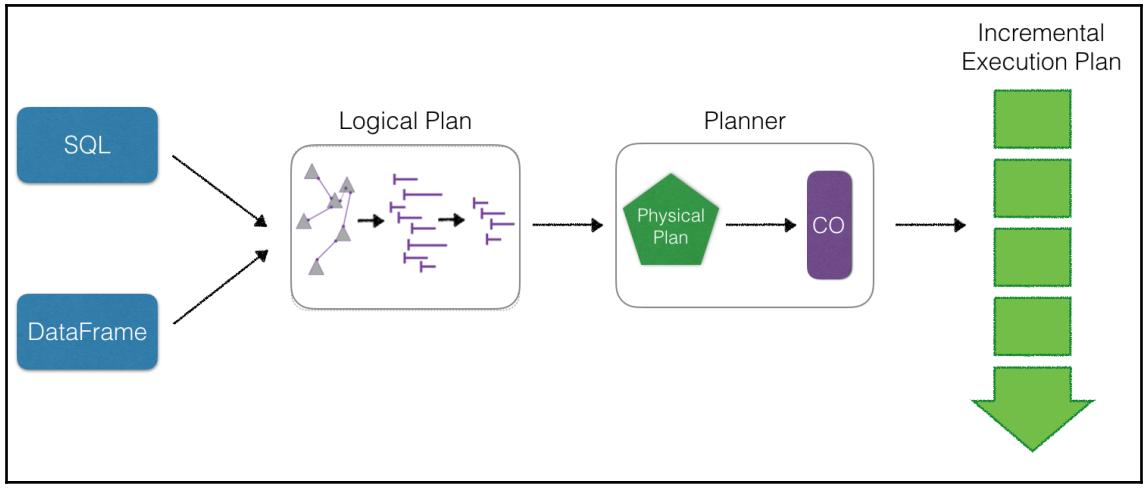

```
Galaxy Andromeda 2 [0.003053456550444906,0.0033317477861422363,0.9936147956634129]
Galaxy Milky Way 2 [0.004752646858051239,0.0050467276024757125,0.9902006255394731]
Geography Australia 1 [0.00632938201257351,0.9877519489900843,0.005918668997342191]
Geography USA 1 [0.002525770470526258,0.9951088020926291,0.002365427436844653]
Geography China 1 [0.0051541381704948135,0.6008937537867546,0.3939521080427506]
Geography Poland 1 [0.006814345676648856,0.986849415140345,0.006336239183006135]
Animal Dog 0 [0.9901640623662747,0.005226762717124236,0.004609174916600995]
Animal Dog 0 [0.9926300349445092,0.003938103061207765,0.0034318619942831073]
Geography Washington State 1 [0.005261811808175384,0.9898606664191076,0.004877521772717041]
```

Chapter 7: Structured Streaming with PySpark

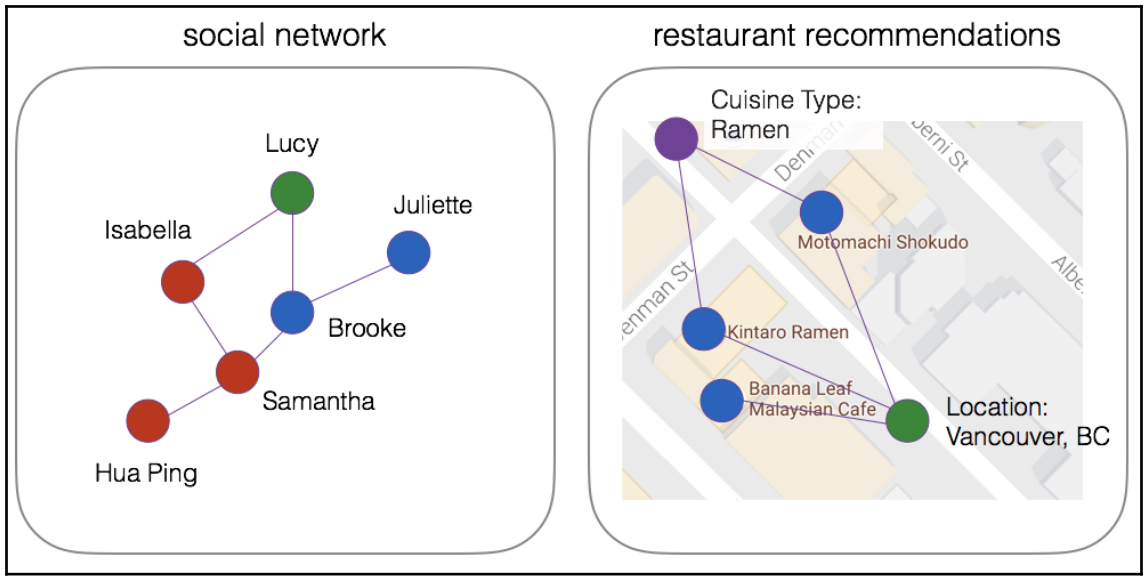


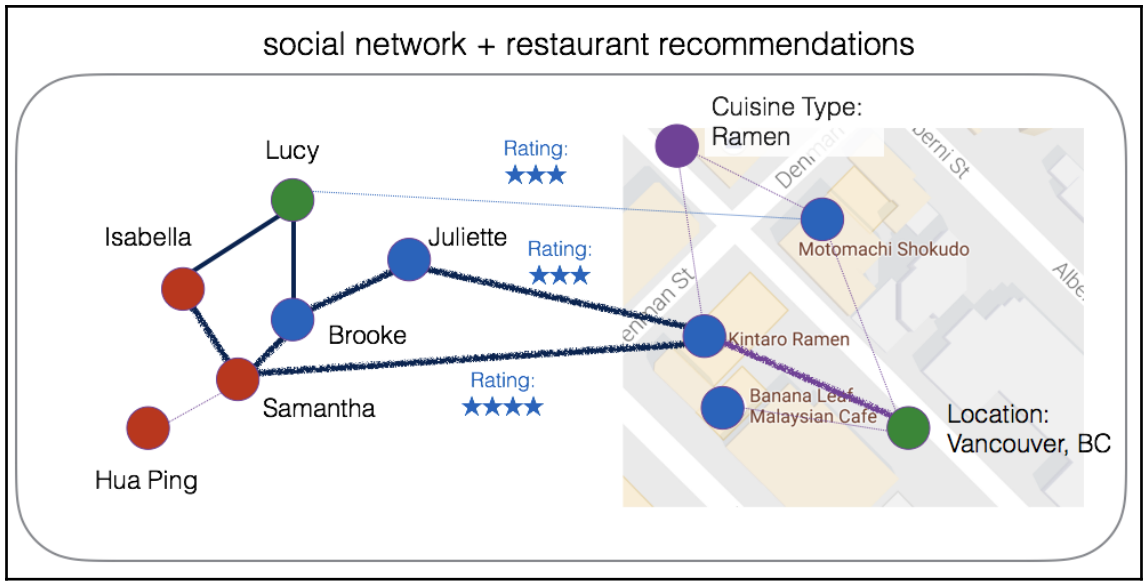

lines DStream
0
1
2
3
4
5
6
...
blue, 5
green, 3
gohawks, 1
green, 2


lines DStream
0
1
2
3
4
5
6
...
blue, 5
green, 3
blue, 5
green, 3
gohawks, 1
blue, 5
green, 5
gohawks, 1
time window


Incremental
Execution Plan
SQL
DataFrame
Logical Plan
Planner
Physical
Plan
CO

Chapter 8: GraphFrames – Graph Theory with PySpark

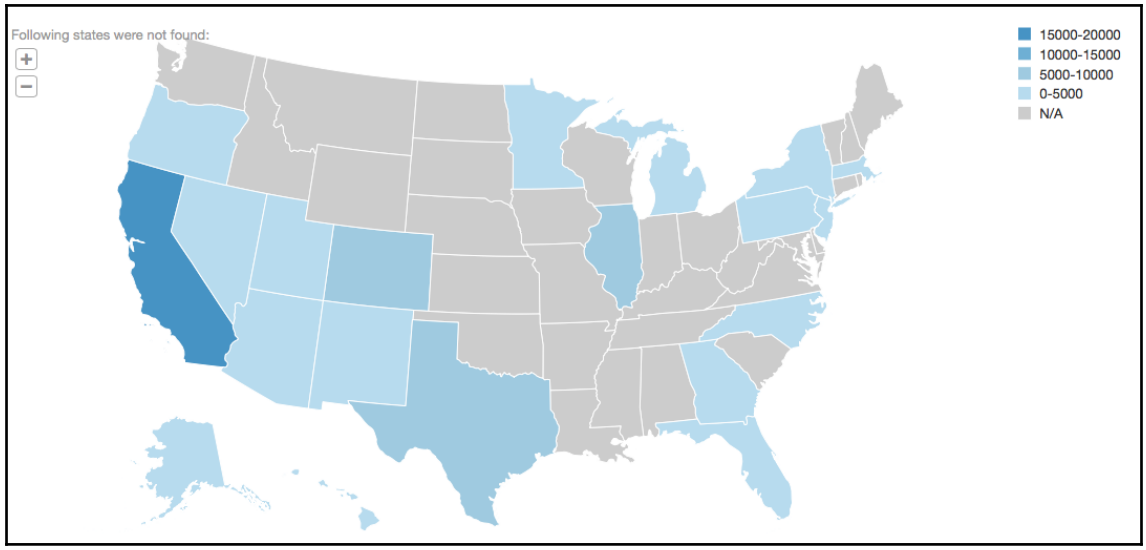




id	City	State	Country
STL	St. Louis	MO	USA
EKO	Elko	NV	USA
RAP	Rapid City	SD	USA
GRK	Killeen	TX	USA
ABQ	Albuquerque	NM	USA
LBB	Lubbock	TX	USA
SAT	San Antonio	TX	USA
BDL	Hartford	CT	USA

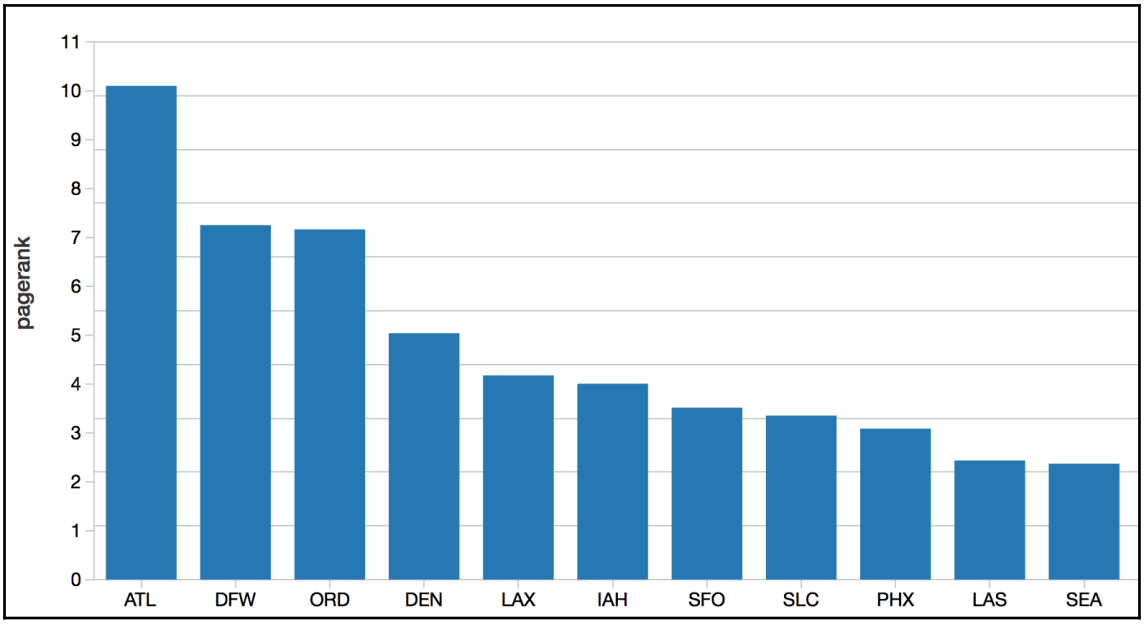
tripid	delay	src	dst	city_dst	state_dst
1011158	-5	SAN	IAH	Houston	TX
1011516	-8	SAN	IAH	Houston	TX
1010937	16	SAN	IAH	Houston	TX
1010702	1	SAN	IAH	Houston	TX
1010620	3	SAN	IAH	Houston	TX
1020620	1	SAN	IAH	Houston	TX
1021507	10	SAN	IAH	Houston	TX
1020814	-3	SAN	IAH	Houston	TX
1021158	0	SAN	IAH	Houston	TX

src	dst	avg(delay)
SFO	OKC	59.073170731707314
SFO	JAC	57.13333333333333
SFO	COS	53.976190476190474
SFO	OTH	48.09090909090909
SFO	SAT	47.625
SFO	MOD	46.80952380952381
SFO	SUN	46.723404255319146



from	e0	v1
{"id":"SFO","City":"San Francisco","State":"CA","Country":"USA"}	{"tripid":1010630,"delay":-3,"src":"SFO","dst":"MSP","city_dst":"Minneapolis","state_dst":"MN"}	{"id":"MSP","City":"Minneapolis","State":"MN","Country":"US/
{"id":"SFO","City":"San Francisco","State":"CA","Country":"USA"}	{"tripid":1010630,"delay":-3,"src":"SFO","dst":"MSP","city_dst":"Minneapolis","state_dst":"MN"}	{"id":"MSP","City":"Minneapolis","State":"MN","Country":"US/
{"id":"SFO","City":"San Francisco","State":"CA","Country":"USA"}	{"tripid":1010830,"delay":52,"src":"SFO","dst":"MSP","city_dst":"Minneapolis","state_dst":"MN"}	{"id":"MSP","City":"Minneapolis","State":"MN","Country":"US/
{"id":"SFO","City":"San Francisco","State":"CA","Country":"USA"}	{"tripid":1010830,"delay":52,"src":"SFO","dst":"MSP","city_dst":"Minneapolis","state_dst":"MN"}	{"id":"MSP","City":"Minneapolis","State":"MN","Country":"US/
{"id":"SFO","City":"San Francisco","State":"CA","Country":"USA"}	{"tripid":1011300,"delay":-6,"src":"SFO","dst":"MSP","city_dst":"Minneapolis","state_dst":"MN"}	{"id":"MSP","City":"Minneapolis","State":"MN","Country":"US/

Showing the first 1000 rows.

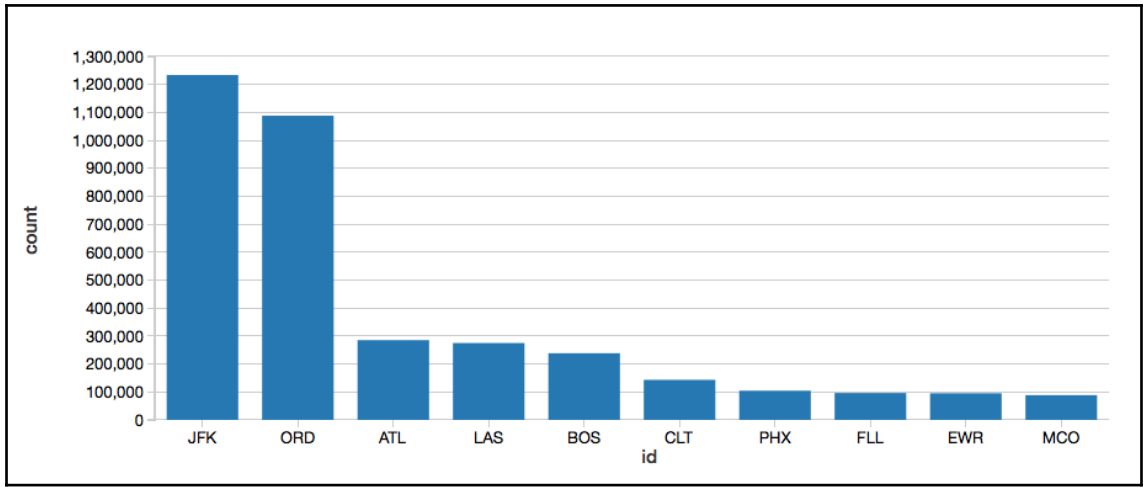


from	e0	to
{"id":"SEA","City":"Seattle","State":"WA","Country":"USA"}	{"tripid":1010710,"delay":31,"src":"SEA","dst":"SFO","city_dst":"San Francisco","state_dst":"CA"}	{"id":"SFO","City":"San Francisco","State":"CA","Country":"USA"}
{"id":"SEA","City":"Seattle","State":"WA","Country":"USA"}	{"tripid":1012125,"delay":-4,"src":"SEA","dst":"SFO","city_dst":"San Francisco","state_dst":"CA"}	{"id":"SFO","City":"San Francisco","State":"CA","Country":"USA"}
{"id":"SEA","City":"Seattle","State":"WA","Country":"USA"}	{"tripid":1011840,"delay":-5,"src":"SEA","dst":"SFO","city_dst":"San Francisco","state_dst":"CA"}	{"id":"SFO","City":"San Francisco","State":"CA","Country":"USA"}
{"id":"SEA","City":"Seattle","State":"WA","Country":"USA"}	{"tripid":1010610,"delay":-4,"src":"SEA","dst":"SFO","city_dst":"San Francisco","state_dst":"CA"}	{"id":"SFO","City":"San Francisco","State":"CA","Country":"USA"}
{"id":"SEA","City":"Seattle","State":"WA","Country":"USA"}	{"tripid":1011230,"delay":-2,"src":"SEA","dst":"SFO","city_dst":"San Francisco","state_dst":"CA"}	{"id":"SFO","City":"San Francisco","State":"CA","Country":"USA"}

Showing the first 1000 rows.

from	e0	v1
{"id":"SFO","City":"San Francisco","State":"CA","Country":"USA"}	{"tripid":1010630,"delay":-3,"src":"SFO","dst":"MSP","city_dst":"Minneapolis","state_dst":"MN"}	{"id":"MSP","City":"Minneapolis","State":"MN","Country":"US/
{"id":"SFO","City":"San Francisco","State":"CA","Country":"USA"}	{"tripid":1010630,"delay":-3,"src":"SFO","dst":"MSP","city_dst":"Minneapolis","state_dst":"MN"}	{"id":"MSP","City":"Minneapolis","State":"MN","Country":"US/
{"id":"SFO","City":"San Francisco","State":"CA","Country":"USA"}	{"tripid":1010830,"delay":52,"src":"SFO","dst":"MSP","city_dst":"Minneapolis","state_dst":"MN"}	{"id":"MSP","City":"Minneapolis","State":"MN","Country":"US/
{"id":"SFO","City":"San Francisco","State":"CA","Country":"USA"}	{"tripid":1010830,"delay":52,"src":"SFO","dst":"MSP","city_dst":"Minneapolis","state_dst":"MN"}	{"id":"MSP","City":"Minneapolis","State":"MN","Country":"US/
{"id":"SFO","City":"San Francisco","State":"CA","Country":"USA"}	{"tripid":1011300,"delay":-6,"src":"SFO","dst":"MSP","city_dst":"Minneapolis","state_dst":"MN"}	{"id":"MSP","City":"Minneapolis","State":"MN","Country":"US/

Showing the first 1000 rows.


1,300,000
1,200,000
1,100,000
1,000,000
900,000
800,000
700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
count
JFK
ORD
ATL
LAS
BOS
CLT
PHX
FLL
EWR
MCO
id