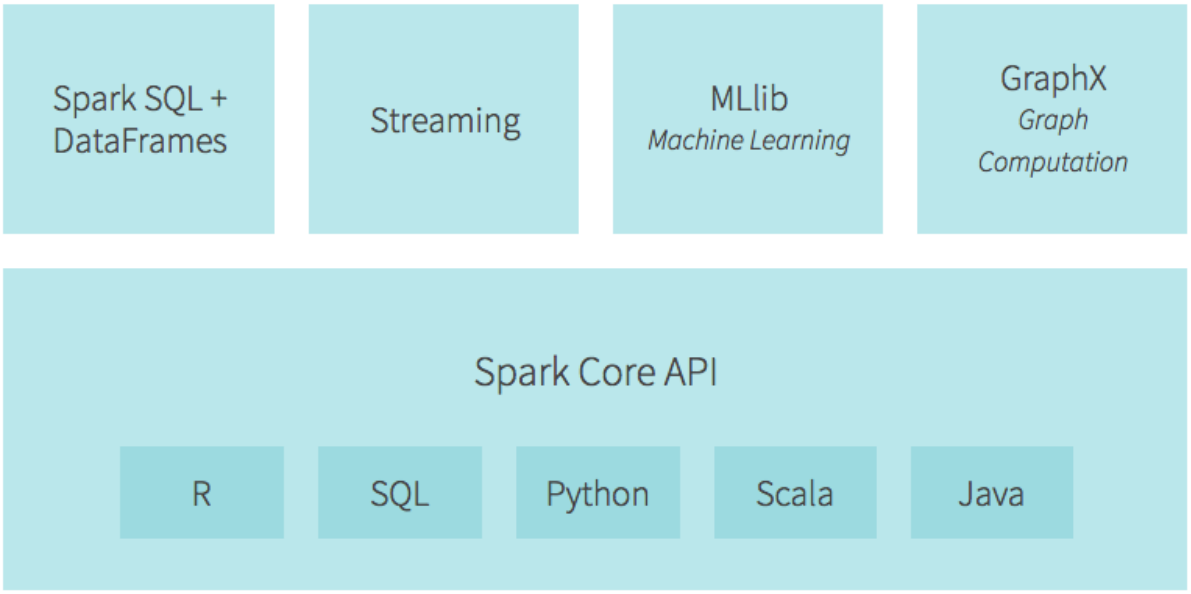
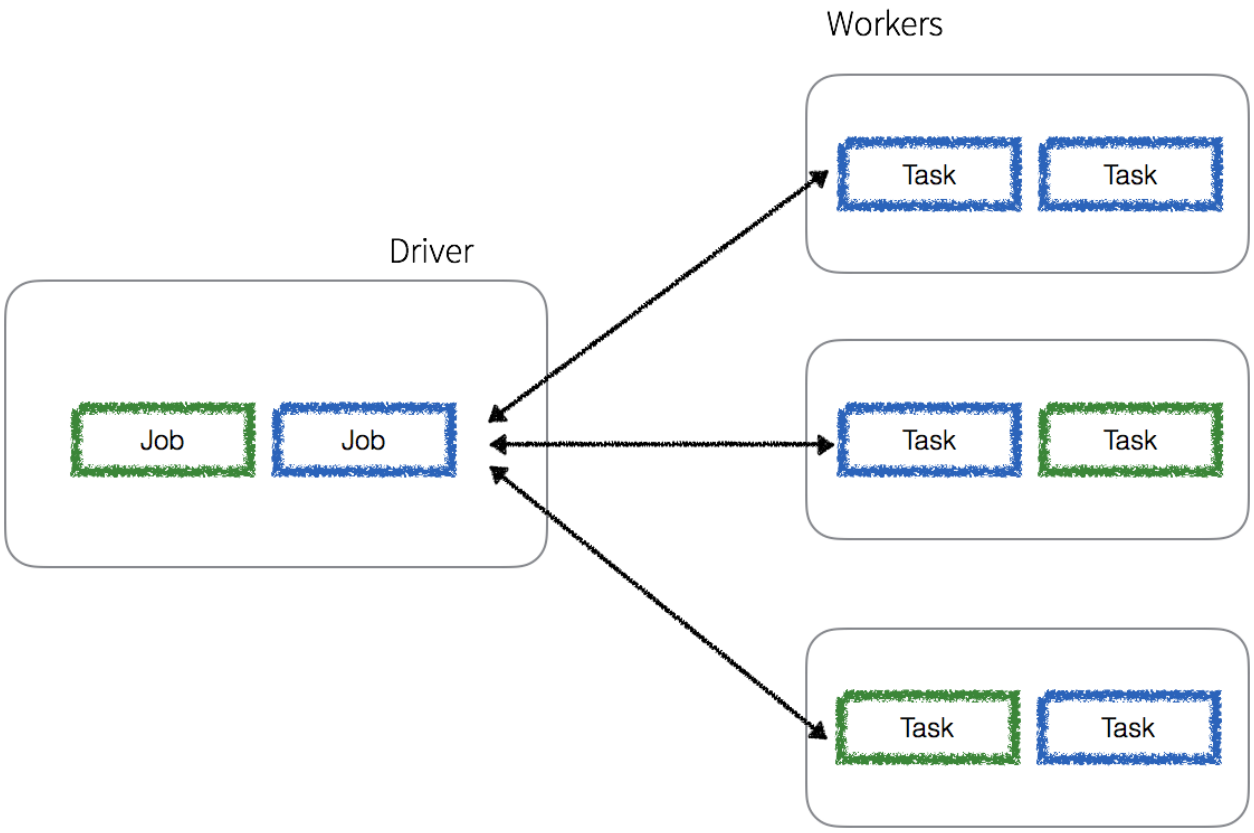
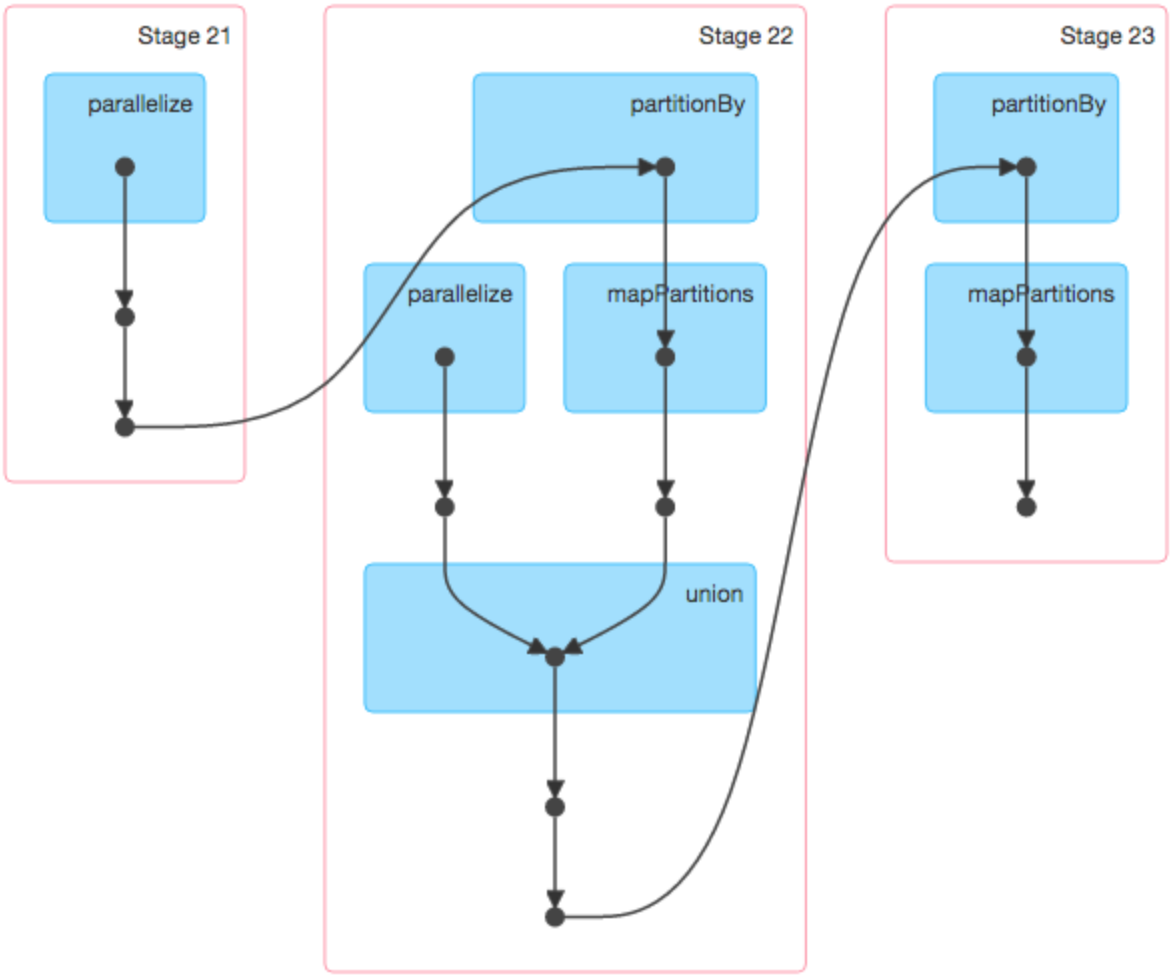
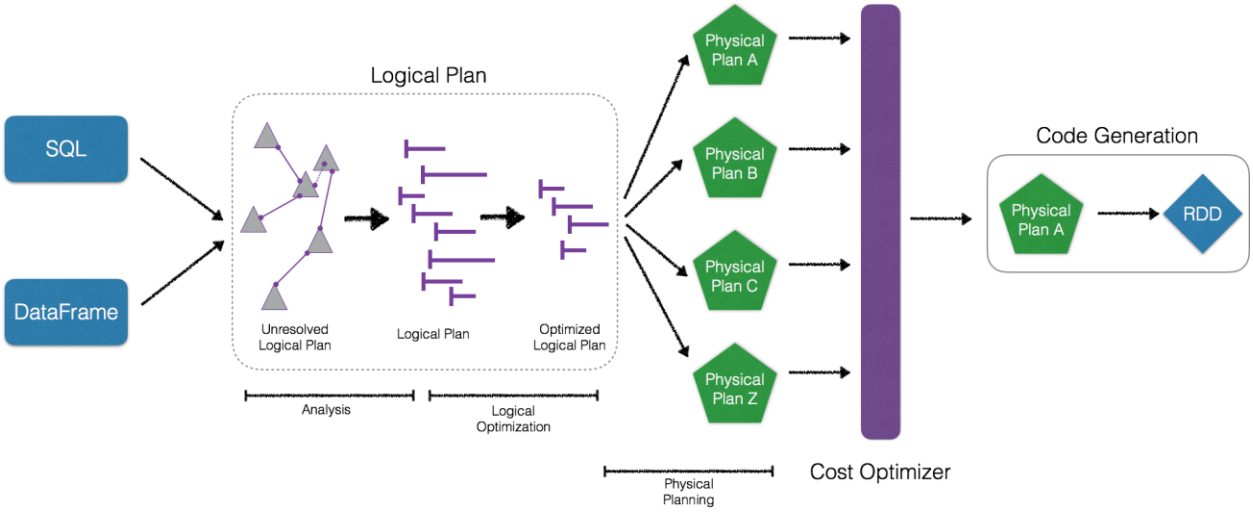
Chapter 1: Understanding Spark




Workers
Task
Task
Driver
Job
Job
Task
Task
Task
Task


Stage 21
parallelize
Stage 22
partitionBy
parallelize
mapPartitions
union
Stage 23
partitionBy
mapPartitions

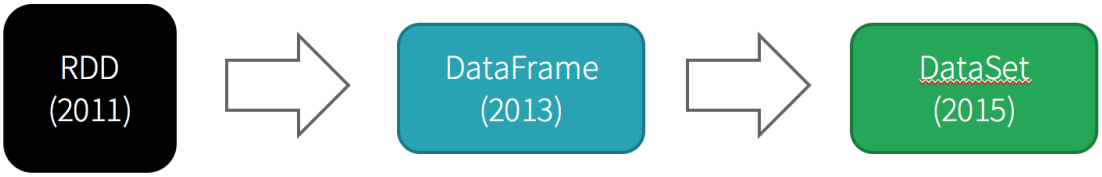

SQL
DataFrame
Logical Plan
Unresolved Logical Plan
Logical Plan
Optimized Logical Plan
Analysis
Logical Optimization
Physical Plan A
Physical Plan B
Physical Plan C
Physical Plan Z
Physical Planning
Cost Optimizer
Code Generation
Physical Plan A
RDD

Tungsten Phase 2
speedups of 5-20x

Structured Streaming

SQL 2003
& Unifying Datasets
and DataFrames

History of Spark APIs



Distribute collection of JVM objects

Functional Operators (map, filter, etc.)

Distribute collection of Row objects

Expression-based operations and UDFs

Logical plans and optimizer

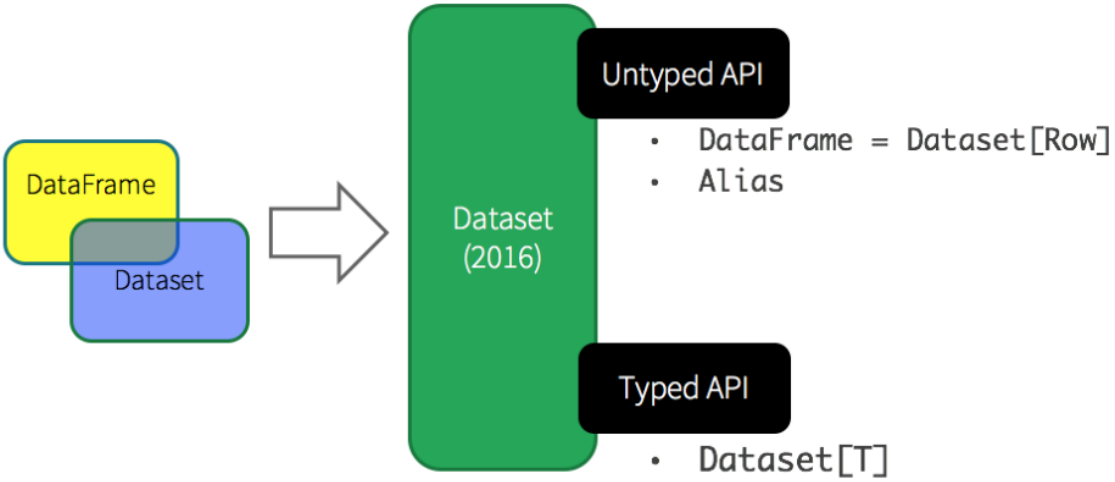
Fast/efficient internal representations

Internally rows, externally JVM objects

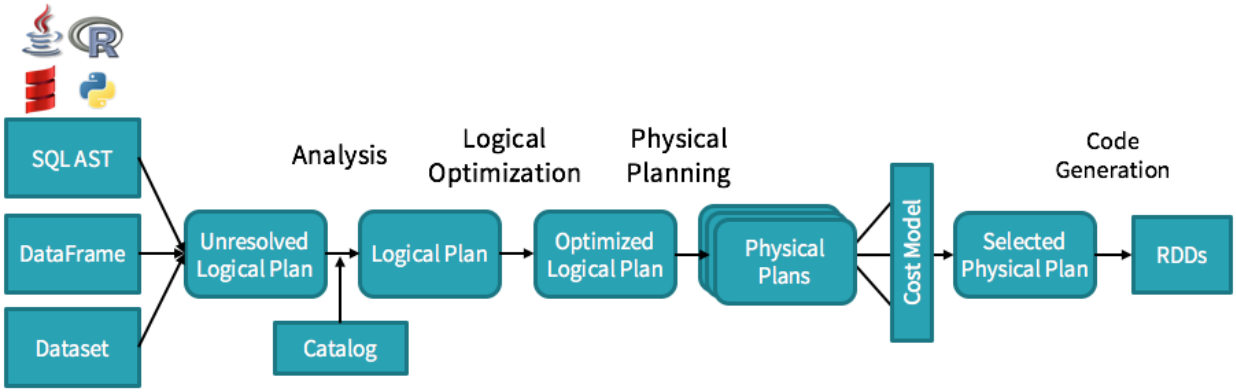
Almost the "Best of both worlds": type safe + fast

But slower than DF
Not as good for interactive analysis, especially Python

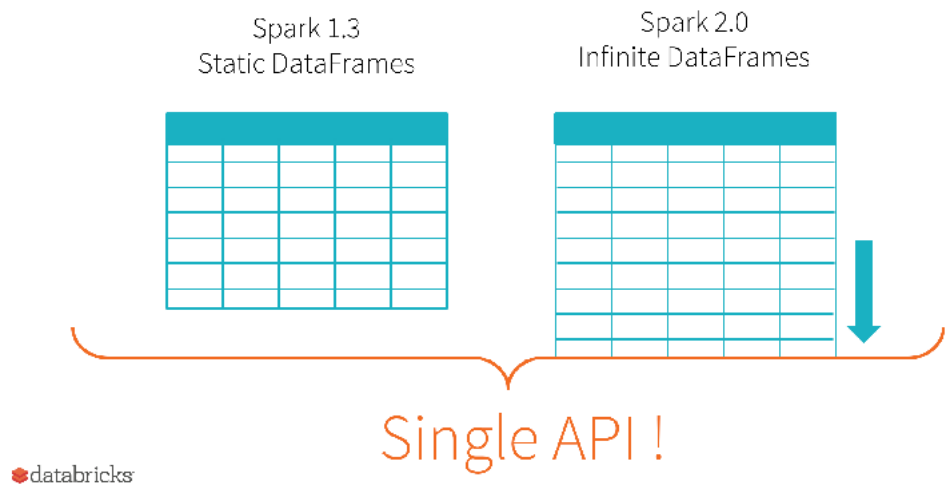
Unified Apache Spark 2.0 API

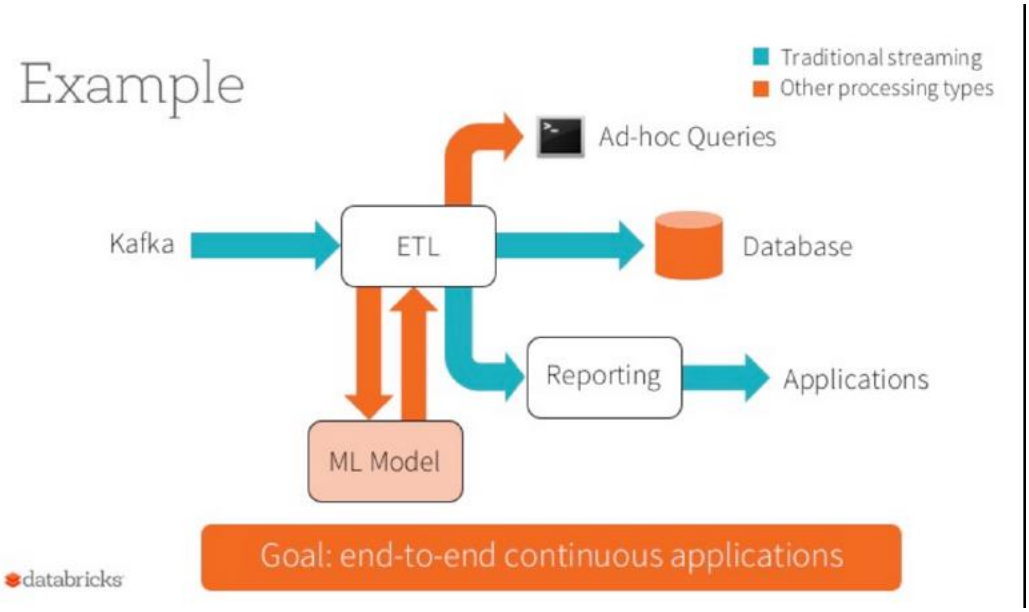


databricks



DataFrames, Datasets and SQL
share the same optimization/execution pipeline


Spark 1.3
Static DataFrames
Spark 2.0
Infinite DataFrames
Single API !
databricks


Example
Traditional streaming
Other processing types
Ad-hoc Queries
Kafka
ETL
Database
Reporting
Applications
ML Model
Goal: end-to-end continuous applications
databricks

Chapter 2: Resilient Distributed Datasets

```
Out[7]: ['                    1
         2101  M1087 432311  4M4                2014U7CN
         I64 238 070   24 0111I64
         01 I64
         01  11                                 100 601']
```

```
Out[4]: [array(['1', '  ', '2', '1', '01', 'M', '1', '087', ' ', '43', '23', '1
        1',
                '  ', '4', 'M', '4', '2014', 'U', '7', 'C', 'N', ' ', ' ', 'I64
        ',
                '238', '070', '   ', '24', '01', '11I64  ', '       ', '
        ',
                '       ', '       ', '       ', '       ', '       ', '
        ',
                '       ', '       ', '       ', '       ', '       ', '
        ',
                '       ', '       ', '       ', '       ', '       ', '01',
                'I64  ', '     ', '     ', '     ', '     ', '     ', '     ',
                '     ', '     ', '     ', '     ', '     ', '     ', '     ',
                '     ', '     ', '     ', '     ', '     ', '01', ' '
        ,
                ' ', '1', '1', '100', '6'],
               dtype='<U40')]
```

```
Out[11]: [2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, -99]
```

```
Out[12]: [('2014', 2014),
          ('2014', 2014),
          ('2014', 2014),
          ('2014', 2014),
          ('2014', 2014),
          ('2014', 2014),
          ('2014', 2014),
          ('2014', 2014),
          ('2014', 2014),
          ('-99', -99)]
```

Out[14]: ['2014', 2015, '2014', 2015, '2014', 2015, '2014', 2015, '2014', 2015]

Out[22]: ['-99', 'M', 'F']

Original dataset: 2631171, sample: 263247

Out[52]: [('c', (10, None)), ('b', (4, '6')), ('a', (1, 4)), ('a', (1, 1))]

Out[48]: [('b', (4, '6')), ('a', (1, 4)), ('a', (1, 1))]

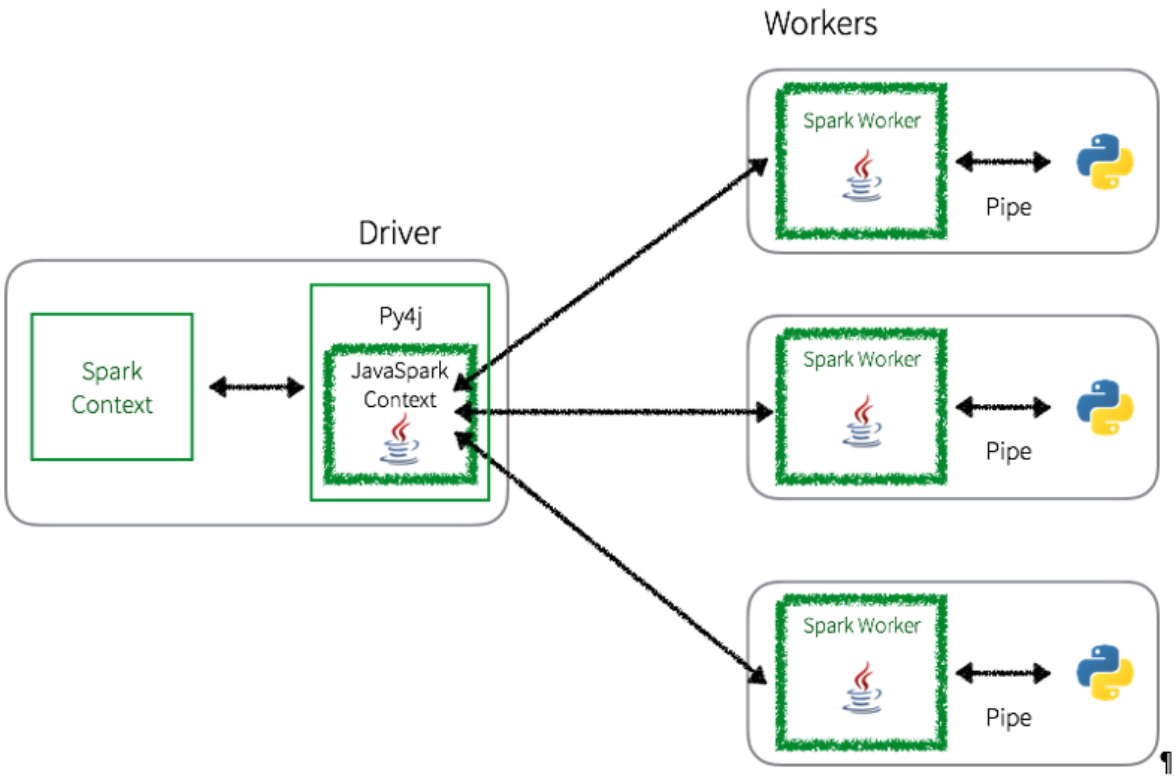
Out[88]: [('a', 1)]

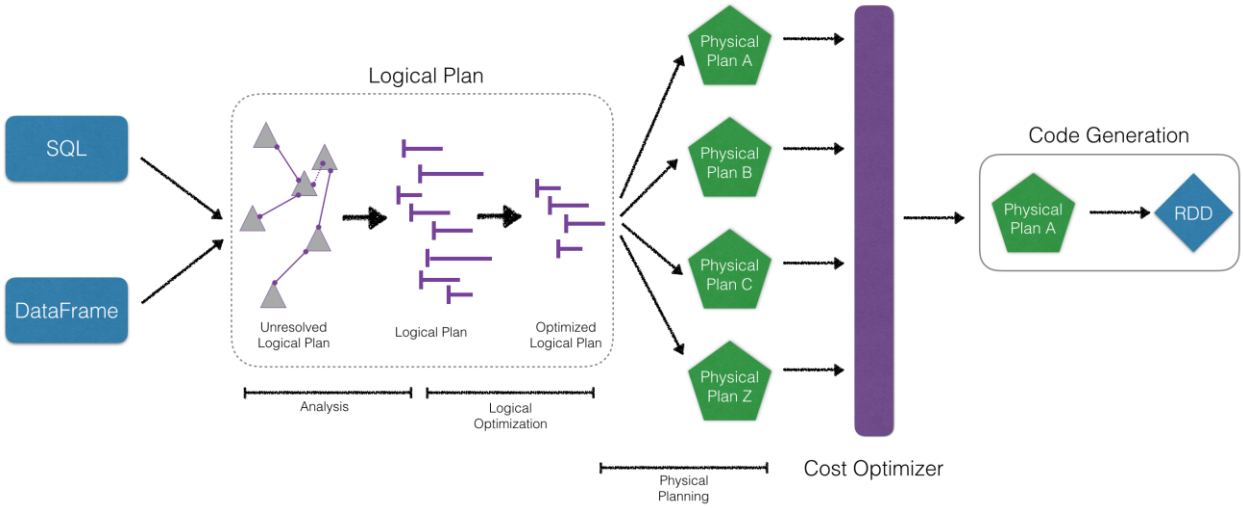
Out[122]: [('b', 4), ('c', 2), ('a', 12), ('d', 5)]

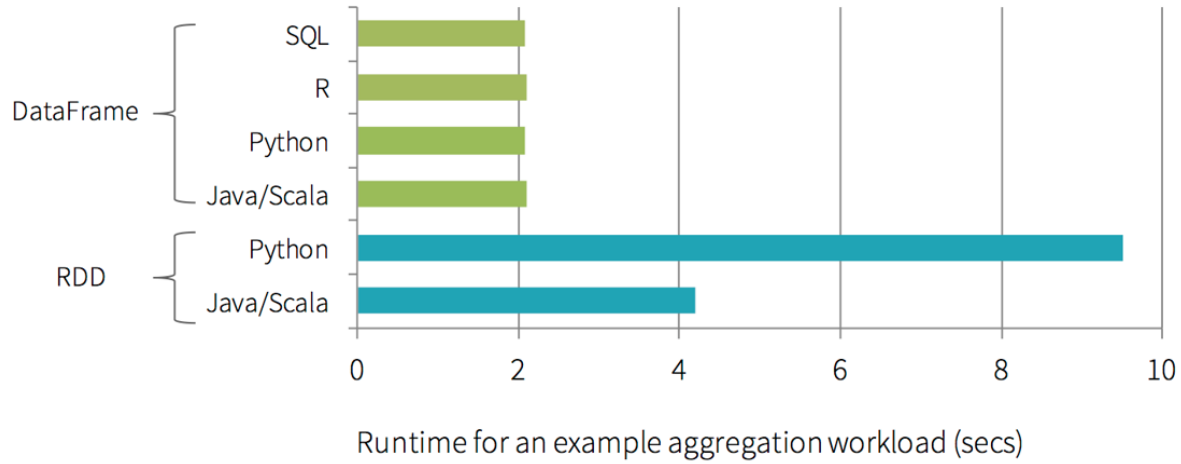
Out[132]: dict_items([('a', 2), ('b', 2), ('d', 2), ('c', 1)])

Out[159]: [('a', 4), ('b', 3), ('c', 2), ('a', 8), ('d', 2), ('b', 1), ('d', 3)]

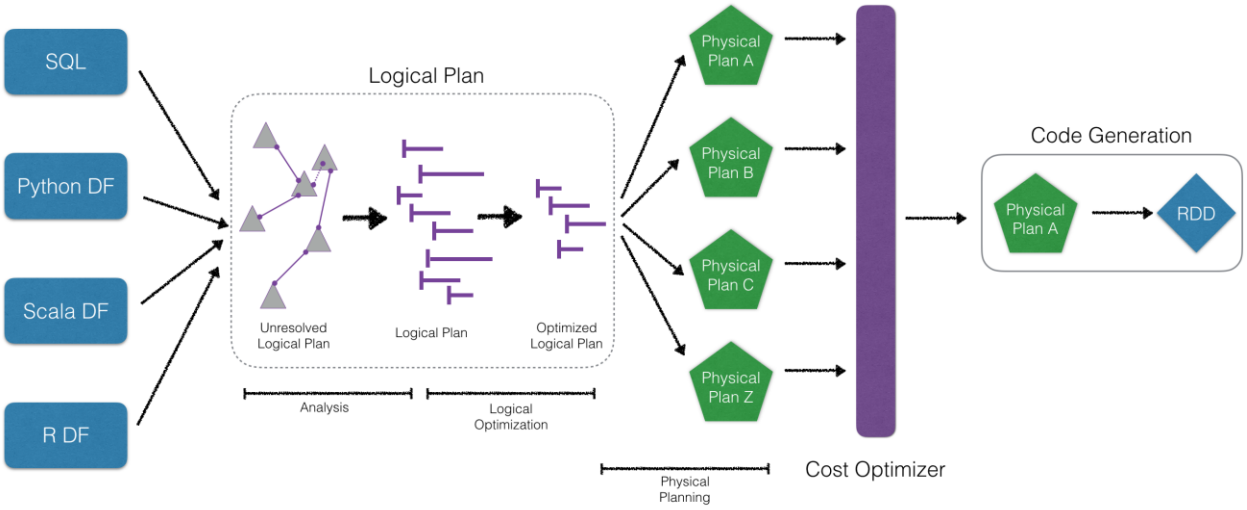
Chapter 3: DataFrames

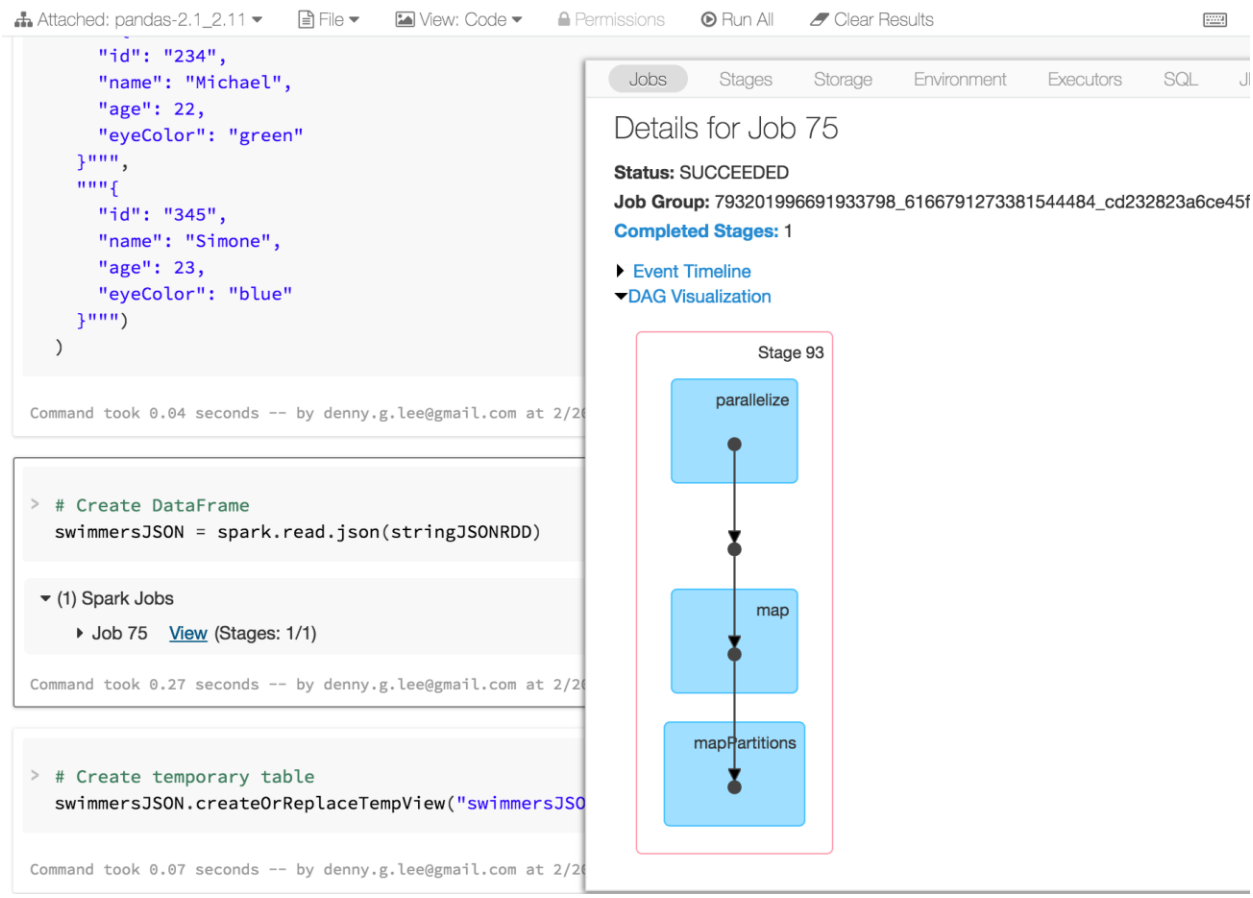






Runtime for an example aggregation workload (secs)




Attached: pandas-2.1_2.11
File
View: Code
Permissions
Run All
Clear Results
"id": "234",
"name": "Michael",
"age": 22,
"eyeColor": "green"
}""",
"""{
"id": "345",
"name": "Simone",
"age": 23,
"eyeColor": "blue"
}""")
)
Command took 0.04 seconds -- by denny.g.lee@gmail.com at 2/2
> # Create DataFrame
swimmersJSON = spark.read.json(stringJSONRDD)
(1) Spark Jobs
Job 75 View (Stages: 1/1)
Command took 0.27 seconds -- by denny.g.lee@gmail.com at 2/2
> # Create temporary table
swimmersJSON.createOrReplaceTempView("swimmersJSO
Command took 0.07 seconds -- by denny.g.lee@gmail.com at 2/2
Jobs
Stages
Storage
Environment
Executors
SQL
Details for Job 75
Status: SUCCEEDED
Job Group: 793201996691933798_6166791273381544484_cd232823a6ce45f
Completed Stages: 1
Event Timeline
DAG Visualization
Stage 93
parallelize
map
mapPartitions

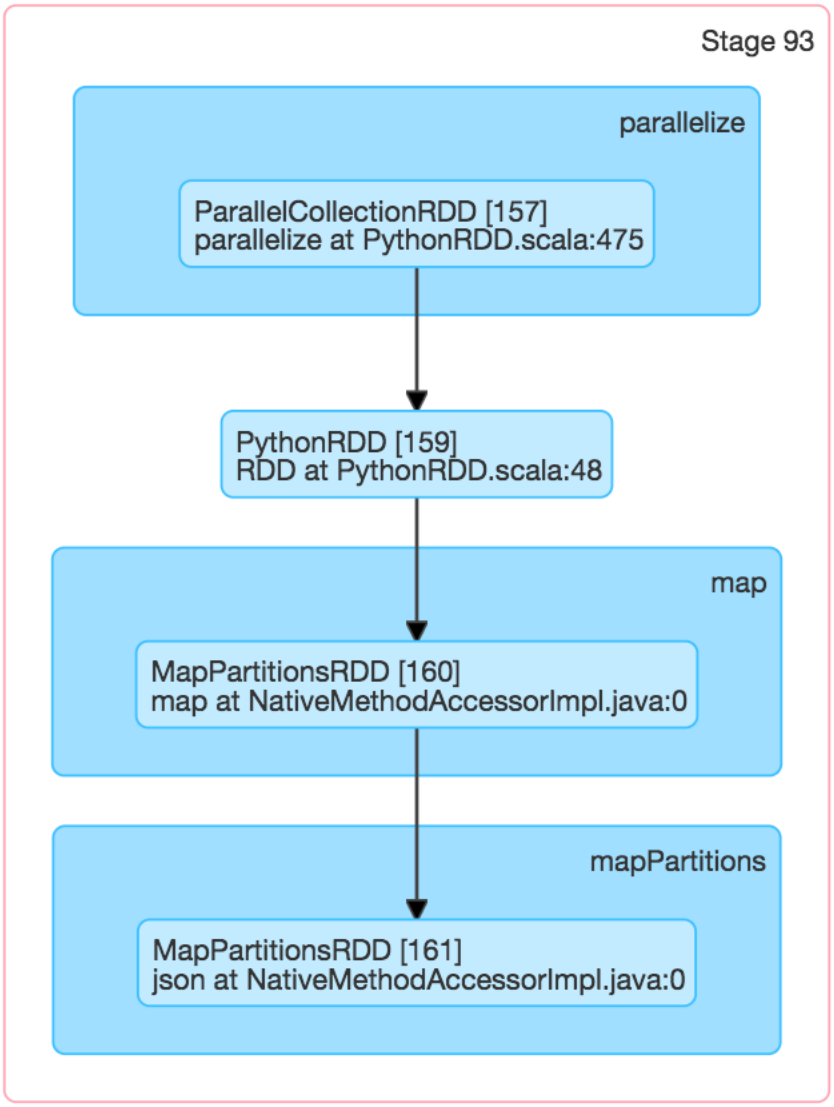
Jobs | Stages | Storage | Environment | Executors | SQL

Details for Stage 93 (Attempt 0)

Total Time Across All Tasks: 1 s

Locality Level Summary: Process local: 8

▾DAG Visualization



▸ (2) Spark Jobs

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

Command took 0.22s

▸ (1) Spark Jobs

Out[6]:

```
[Row(age=19, eyeColor=u'brown', id=u'123', name=u'Katie'),
 Row(age=22, eyeColor=u'green', id=u'234', name=u'Michael'),
 Row(age=23, eyeColor=u'blue', id=u'345', name=u'Simone')]
```

Command took 0.17s

```
> %sql
  -- Query Data
  select * from swimmersJSON
```

▸ (3) Spark Jobs

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

Command took 0.23 seconds -- by denny.g.lee@gmail.com at 2/20/2017, 10:18:30 AM on pandas-2.1_2.11

```
root
 |-- age: long (nullable = true)
 |-- eyeColor: string (nullable = true)
 |-- id: string (nullable = true)
 |-- name: string (nullable = true)
```

Command took 0.07s

```
root
 |-- id: long (nullable = true)
 |-- name: string (nullable = true)
 |-- age: long (nullable = true)
 |-- eyeColor: string (nullable = true)
```

Command took 0.04s

▸ (2) Spark Jobs

```
+---+---+
| id|age|
+---+---+
|234| 22|
+---+---+
```

Command took 0.22s

▸ (2) Spark Jobs

```
+------+--------+
|  name|eyeColor|
+------+--------+
| Katie|   brown|
|Simone|    blue|
+------+--------+
```

Command took 0.22s

▸ (1) Spark Jobs

```
+--------+
|count(1)|
+--------+
|       3|
+--------+
```

Command took 0.42s

▸ (2) Spark Jobs

```
+---+---+
| id|age|
+---+---+
|234| 22|
+---+---+
```

Command took 0.27s

▸ (2) Spark Jobs

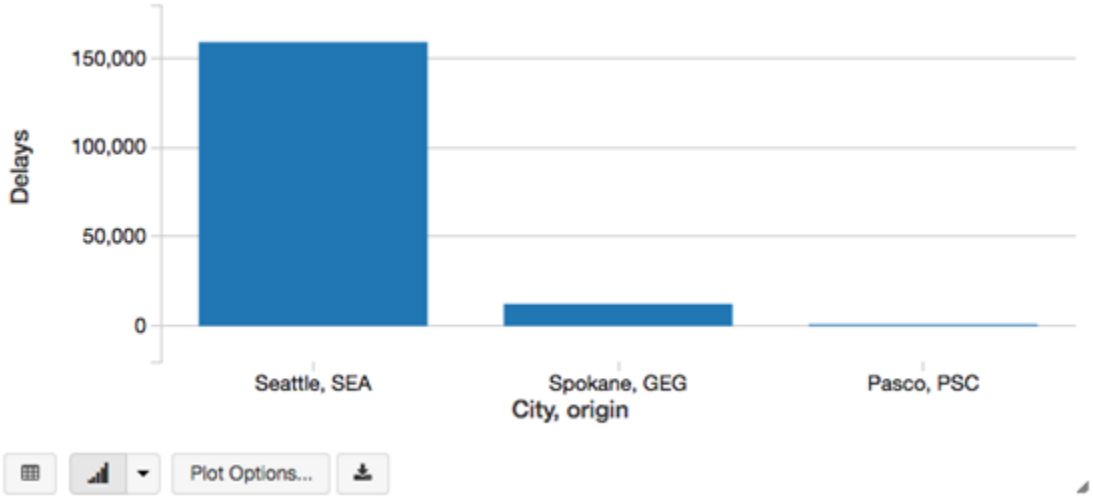
```
+------+--------+
|  name|eyeColor|
+------+--------+
| Katie|   brown|
|Simone|    blue|
+------+--------+
```

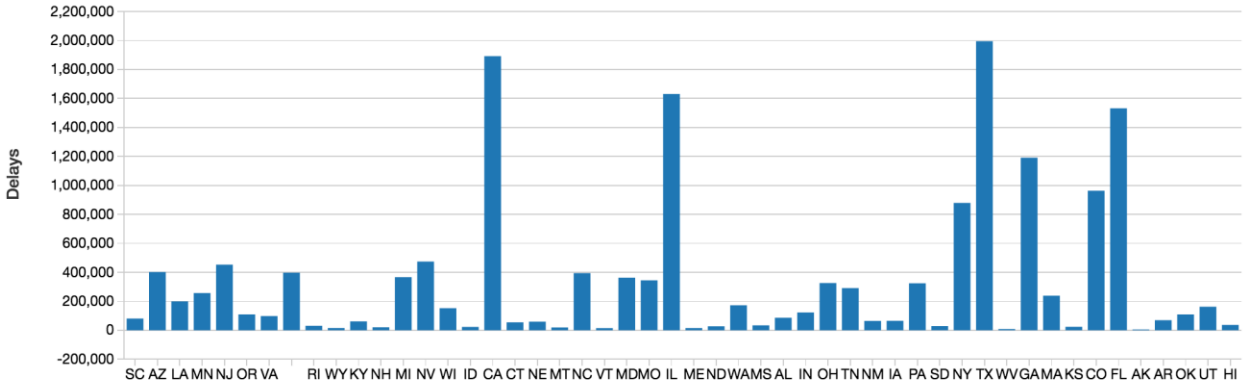
Command took 0.27s

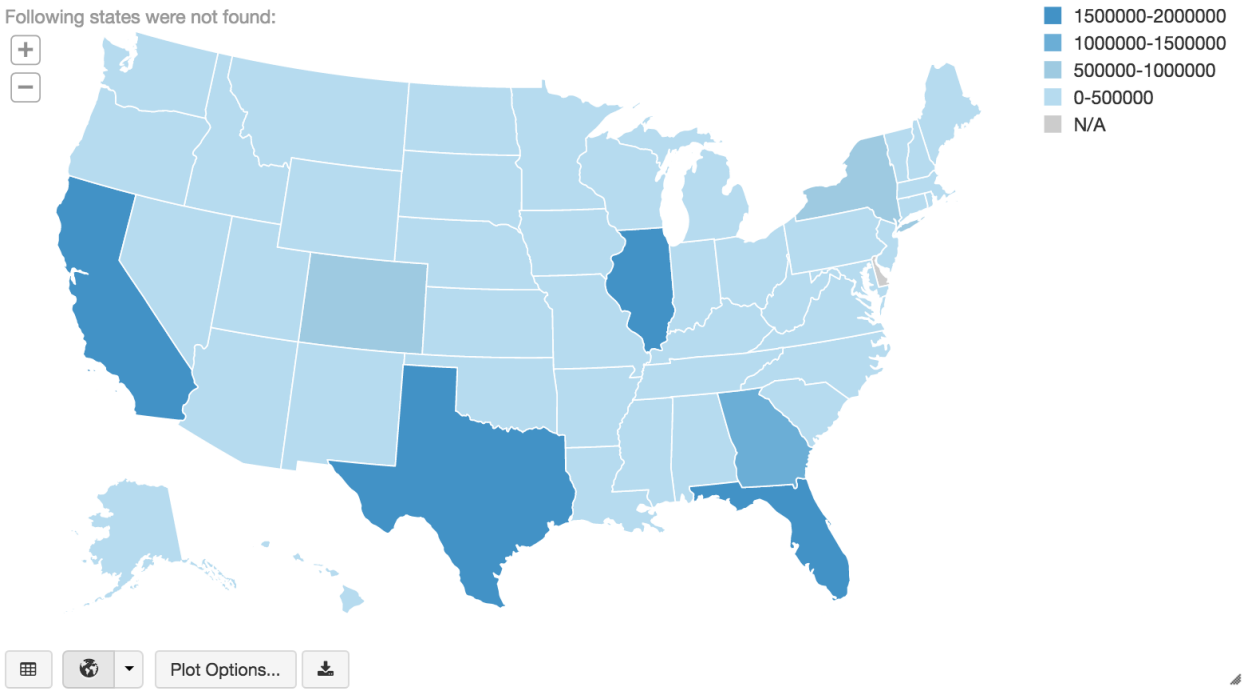
▸ (2) Spark Jobs

```
+-------+------+--------+
|   City|origin|  Delays|
+-------+------+--------+
|Seattle|   SEA|159086.0|
|Spokane|   GEG| 12404.0|
|  Pasco|   PSC|   949.0|
+-------+------+--------+
```

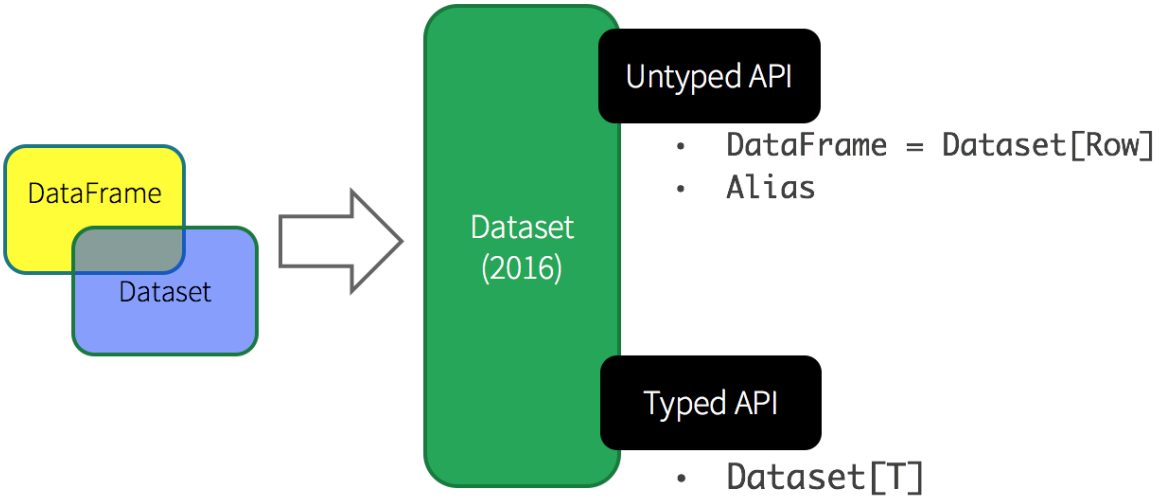
Command took 3.93s







Unified Apache Spark 2.0 API



Chapter 4: Prepared Data for Modeling

```
Count of rows: 7
Count of distinct rows: 6
```

id	weight	height	age	gender
4	144.5	5.9	33	M
1	144.5	5.9	33	M
5	129.2	5.3	42	M
5	133.2	5.7	54	F
2	167.2	5.4	45	M
3	124.1	5.2	23	F

```
Count of ids: 6
Count of distinct ids: 5
```

id	weight	height	age	gender
5	133.2	5.7	54	F
4	144.5	5.9	33	M
2	167.2	5.4	45	M
3	124.1	5.2	23	F
5	129.2	5.3	42	M

```
+-----+--------+
|count|distinct|
+-----+--------+
|    5|       4|
+-----+--------+
```

```
+---+------+------+---+------+-------------+
| id|weight|height|age|gender|       new_id|
+---+------+------+---+------+-------------+
|  5| 133.2|   5.7| 54|     F|  25769803776|
|  4| 144.5|   5.9| 33|     M| 171798691840|
|  2| 167.2|   5.4| 45|     M| 592705486848|
|  3| 124.1|   5.2| 23|     F|1236950581248|
|  5| 129.2|   5.3| 42|     M|1365799600128|
+---+------+------+---+------+-------------+
```

Out[9]: [(1, 0), (2, 1), (3, 4), (4, 1), (5, 1), (6, 2), (7, 0)]

```
+---+------+------+----+------+------+
| id|weight|height| age|gender|income|
+---+------+------+----+------+------+
|  3|  null|   5.2|null|  null|  null|
+---+------+------+----+------+------+
```

```
+----------+-----------------+-------------+-----------------+-----------------+-----------------+
|id_missing|   weight_missing|height_missing|      age_missing|   gender_missing|   income_missing|
+----------+-----------------+-------------+-----------------+-----------------+-----------------+
|       0.0|0.1428571428571429|          0.0|0.2857142857142857|0.1428571428571429|0.7142857142857143|
+----------+-----------------+-------------+-----------------+-----------------+-----------------+
```

```
+---+------+------+----+------+
| id|weight|height| age|gender|
+---+------+------+----+------+
|  1| 143.5|   5.6|  28|     M|
|  2| 167.2|   5.4|  45|     M|
|  4| 144.5|   5.9|  33|     M|
|  5| 133.2|   5.7|  54|     F|
|  6| 124.1|   5.2|null|     F|
|  7| 129.2|   5.3|  42|     M|
+---+------+------+----+------+
```

```
+---+-------------+------+---+-------+
| id|       weight|height|age| gender|
+---+-------------+------+---+-------+
|  1|        143.5|   5.6| 28|      M|
|  2|        167.2|   5.4| 45|      M|
|  3|140.283333333|   5.2| 40|missing|
|  4|        144.5|   5.9| 33|      M|
|  5|        133.2|   5.7| 54|      F|
|  6|        124.1|   5.2| 40|      F|
|  7|        129.2|   5.3| 42|      M|
+---+-------------+------+---+-------+
```

```
{'age': 40.399999999999999,
 'height': 5.4714285714285706,
 'id': 4.0,
 'weight': 140.28333333333333}
```

Out[17]:
```
{'age': [9.0, 51.0],
 'height': [4.8999999999999995, 5.6],
 'weight': [115.0, 146.84999999999997]}
```

id	weight_o	height_o	age_o
1	false	false	false
2	true	false	false
3	true	false	true
4	false	false	false
5	false	false	true
6	false	false	false
7	false	false	false

id	weight
3	342.3
2	154.2

id	age
5	54
3	99

```
"custID","gender","state","cardholder","balance","numTrans","numIntlTrans","creditLine","fraudRisk"
1,1,35,1,3000,4,14,2,0
2,2,2,1,0,9,0,18,0
3,2,2,1,0,27,9,16,0
4,1,15,1,0,12,0,5,0
5,1,46,1,0,11,16,7,0
```

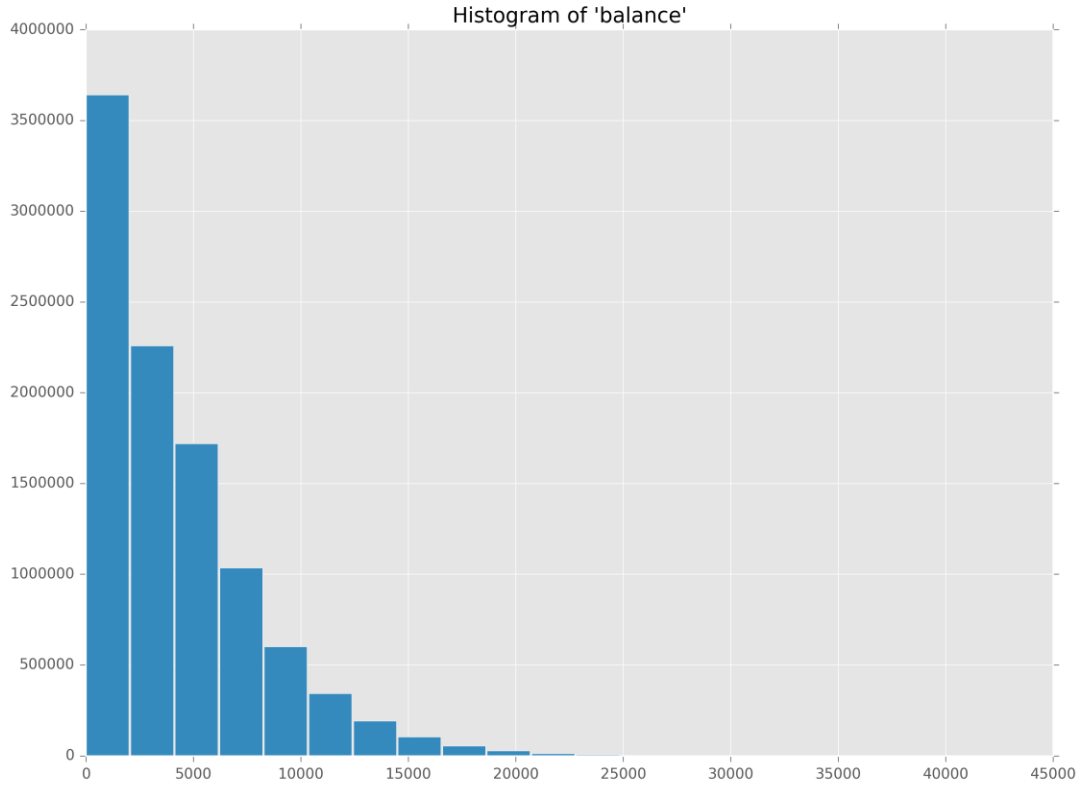
```
root
 |-- custID: integer (nullable = true)
 |-- gender: integer (nullable = true)
 |-- state: integer (nullable = true)
 |-- cardholder: integer (nullable = true)
 |-- balance: integer (nullable = true)
 |-- numTrans: integer (nullable = true)
 |-- numIntlTrans: integer (nullable = true)
 |-- creditLine: integer (nullable = true)
 |-- fraudRisk: integer (nullable = true)
```

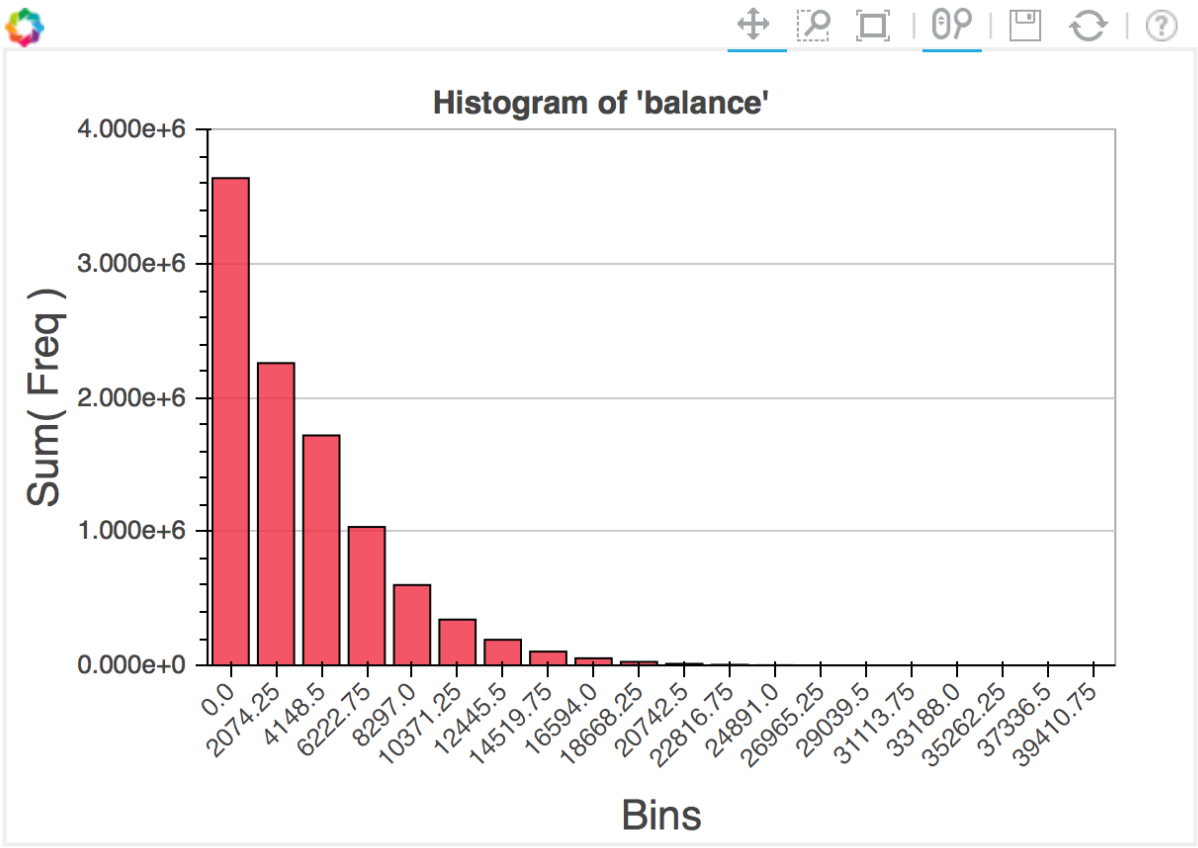
gender	count
1	6178231
2	3821769

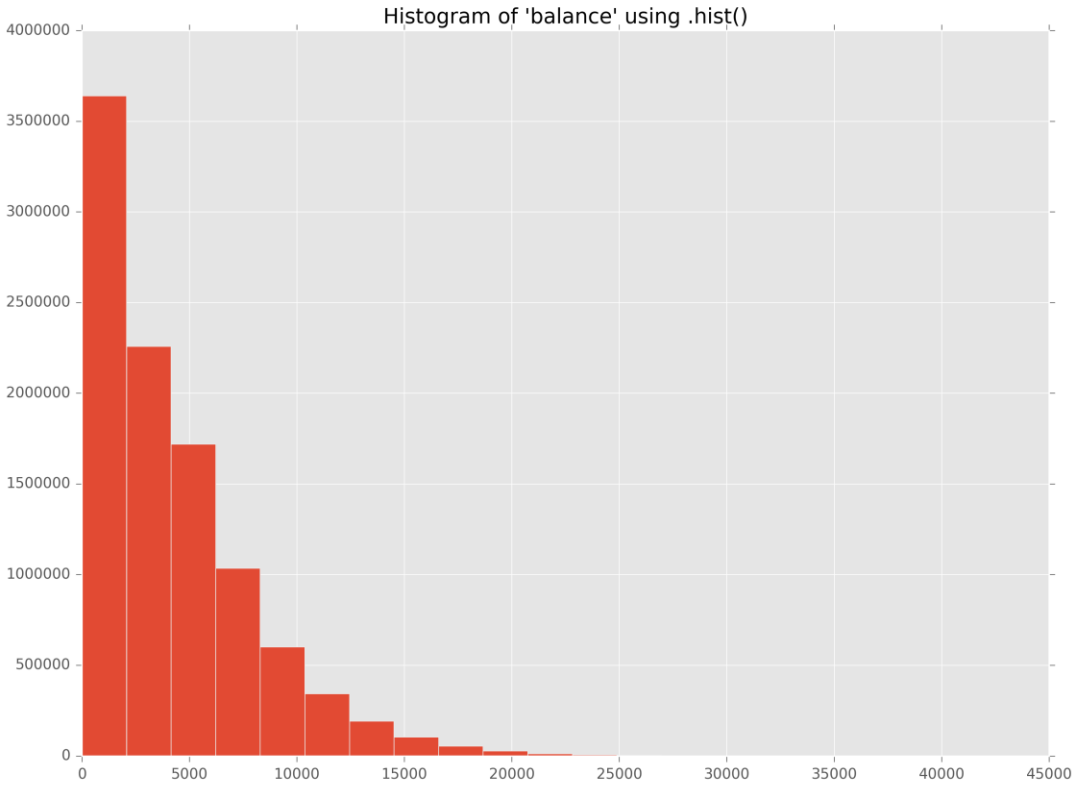
summary	balance	numTrans	numIntlTrans
count	10000000	10000000	10000000
mean	4109.9199193	28.9351871	4.0471899
stddev	3996.847309737077	26.553781024522852	8.602970115863767
min	0	0	0
max	41485	100	60

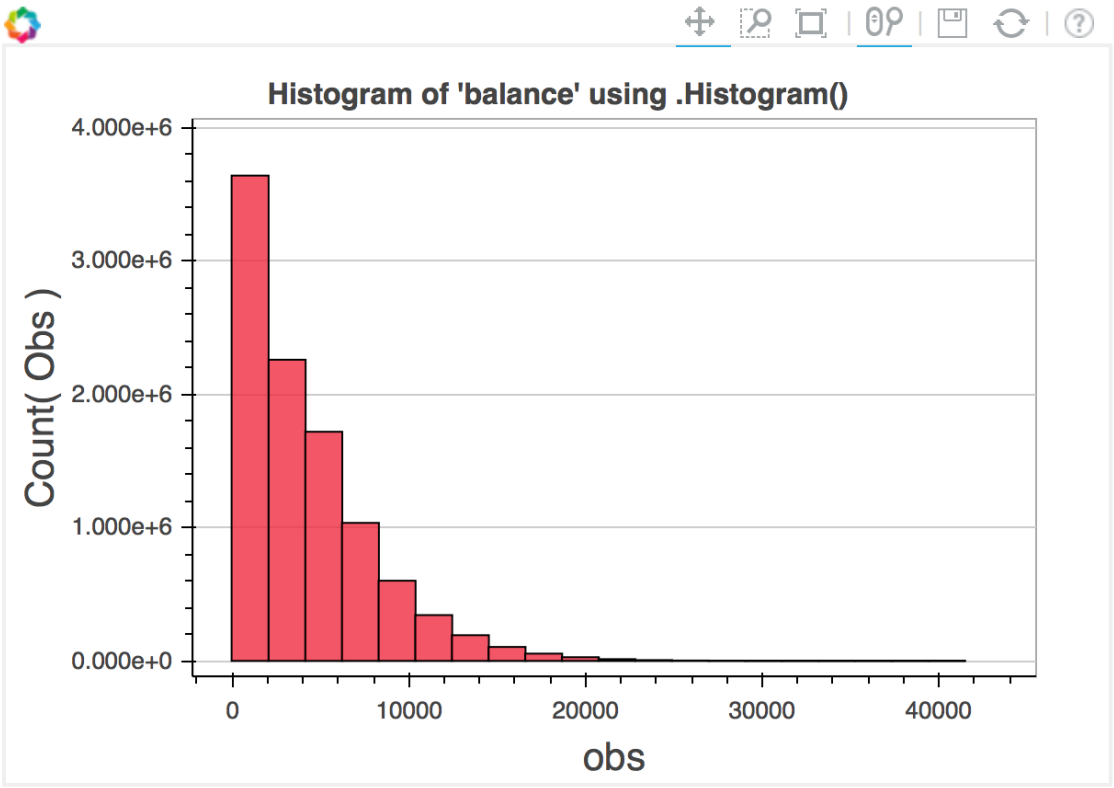
skewness(balance)
1.1818315552995033

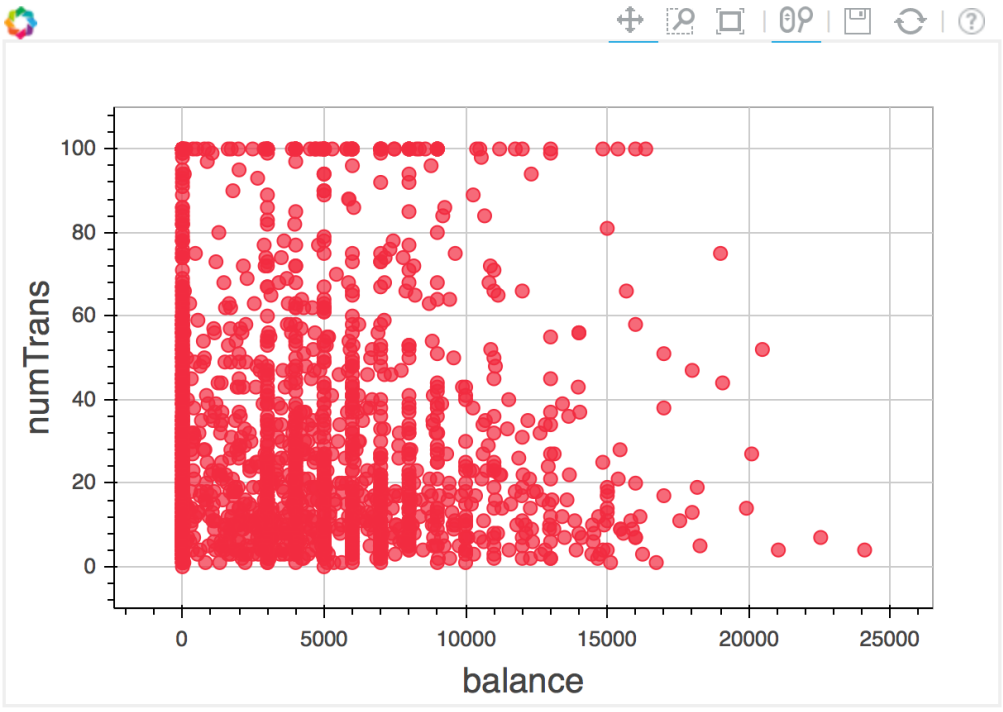
```
Out[30]: [[1.0, 0.00044523140172659576, 0.00027139913398184604],
          [None, 1.0, -0.00028057128198161 79],
          [None, None, 1.0]]
```




Histogram of 'balance'
4.000e+6
3.000e+6
2.000e+6
1.000e+6
0.000e+0
Sum(Freq)
0.0
2074.25
4148.5
6222.75
8297.0
10371.25
12445.5
14519.75
16594.0
18668.25
20742.5
22816.75
24891.0
26965.25
29039.5
31113.75
33188.0
35262.25
37336.5
39410.75
Bins


Histogram of 'balance' using .hist()
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
5000
10000
15000
20000
25000
30000
35000
40000
45000


Histogram of 'balance' using .Histogram()
4.000e+6
3.000e+6
2.000e+6
1.000e+6
0.000e+0
Count(Obs)
0
10000
20000
30000
40000
obs


100
80
60
40
20
0
numTrans
0
5000
10000
15000
20000
25000
balance

Chapter 5: Introducing MLlib

```
Out[8]: [Row(INFANT_NICU_ADMISSION='Y', INFANT_NICU_ADMISSION_RECODE=1),
         Row(INFANT_NICU_ADMISSION='Y', INFANT_NICU_ADMISSION_RECODE=1),
         Row(INFANT_NICU_ADMISSION='U', INFANT_NICU_ADMISSION_RECODE=0),
         Row(INFANT_NICU_ADMISSION='N', INFANT_NICU_ADMISSION_RECODE=0),
         Row(INFANT_NICU_ADMISSION='U', INFANT_NICU_ADMISSION_RECODE=0)]
```

```
+------------+-------------+------------+-------------+------------------+
|DIABETES_PRE|DIABETES_GEST|HYP_TENS_PRE|HYP_TENS_GEST|PREV_BIRTH_PRETERM|
+------------+-------------+------------+-------------+------------------+
|           0|            0|           0|            0|                 0|
|           0|            0|           0|            0|                 0|
|           0|            0|           0|            0|                 0|
|           0|            0|           0|            0|                 1|
|           0|            0|           0|            0|                 0|
+------------+-------------+------------+-------------+------------------+
only showing top 5 rows
```

```
MOTHER_AGE_YEARS:        28.30    6.08
FATHER_COMBINED_AGE:     44.55    27.55
CIG_BEFORE:      1.43     5.18
CIG_1_TRI:       0.91     3.83
CIG_2_TRI:       0.70     3.31
CIG_3_TRI:       0.58     3.11
MOTHER_HEIGHT_IN:        65.12    6.45
MOTHER_PRE_WEIGHT:       214.50   210.21
MOTHER_DELIVERY_WEIGHT:          223.63   180.01
MOTHER_WEIGHT_GAIN:      30.74    26.23
```

```
INFANT_ALIVE_AT_REPORT [(1, 23349), (0, 22080)]
BIRTH_PLACE [('1', 44558), ('4', 327), ('3', 224), ('2', 136), ('7', 91), ('5', 74), ('6', 11), ('9', 8)]
DIABETES_PRE [(0, 44881), (1, 548)]
DIABETES_GEST [(0, 43451), (1, 1978)]
HYP_TENS_PRE [(0, 44348), (1, 1081)]
HYP_TENS_GEST [(0, 43302), (1, 2127)]
PREV_BIRTH_PRETERM [(0, 43088), (1, 2341)]
```

```
CIG_BEFORE-to-CIG_1_TRI: 0.83
CIG_BEFORE-to-CIG_2_TRI: 0.72
CIG_BEFORE-to-CIG_3_TRI: 0.62
CIG_1_TRI-to-CIG_BEFORE: 0.83
CIG_1_TRI-to-CIG_2_TRI: 0.87
CIG_1_TRI-to-CIG_3_TRI: 0.76
CIG_2_TRI-to-CIG_BEFORE: 0.72
CIG_2_TRI-to-CIG_1_TRI: 0.87
CIG_2_TRI-to-CIG_3_TRI: 0.89
CIG_3_TRI-to-CIG_BEFORE: 0.62
CIG_3_TRI-to-CIG_1_TRI: 0.76
CIG_3_TRI-to-CIG_2_TRI: 0.89
MOTHER_PRE_WEIGHT-to-MOTHER_DELIVERY_WEIGHT: 0.54
MOTHER_PRE_WEIGHT-to-MOTHER_WEIGHT_GAIN: 0.65
MOTHER_DELIVERY_WEIGHT-to-MOTHER_PRE_WEIGHT: 0.54
MOTHER_DELIVERY_WEIGHT-to-MOTHER_WEIGHT_GAIN: 0.60
MOTHER_WEIGHT_GAIN-to-MOTHER_PRE_WEIGHT: 0.65
MOTHER_WEIGHT_GAIN-to-MOTHER_DELIVERY_WEIGHT: 0.60
```

```
DenseMatrix([[ 1.,  4.],
             [ 2.,  5.],
             [ 3.,  6.]])
```

```
BIRTH_PLACE 0.0
DIABETES_PRE 0.0
DIABETES_GEST 0.0
HYP_TENS_PRE 0.0
HYP_TENS_GEST 0.0
PREV_BIRTH_PRETERM 0.0
```

```
Area under PR: 0.85
Area under ROC: 0.63
```

```
Area under PR: 0.86
Area under ROC: 0.63
```

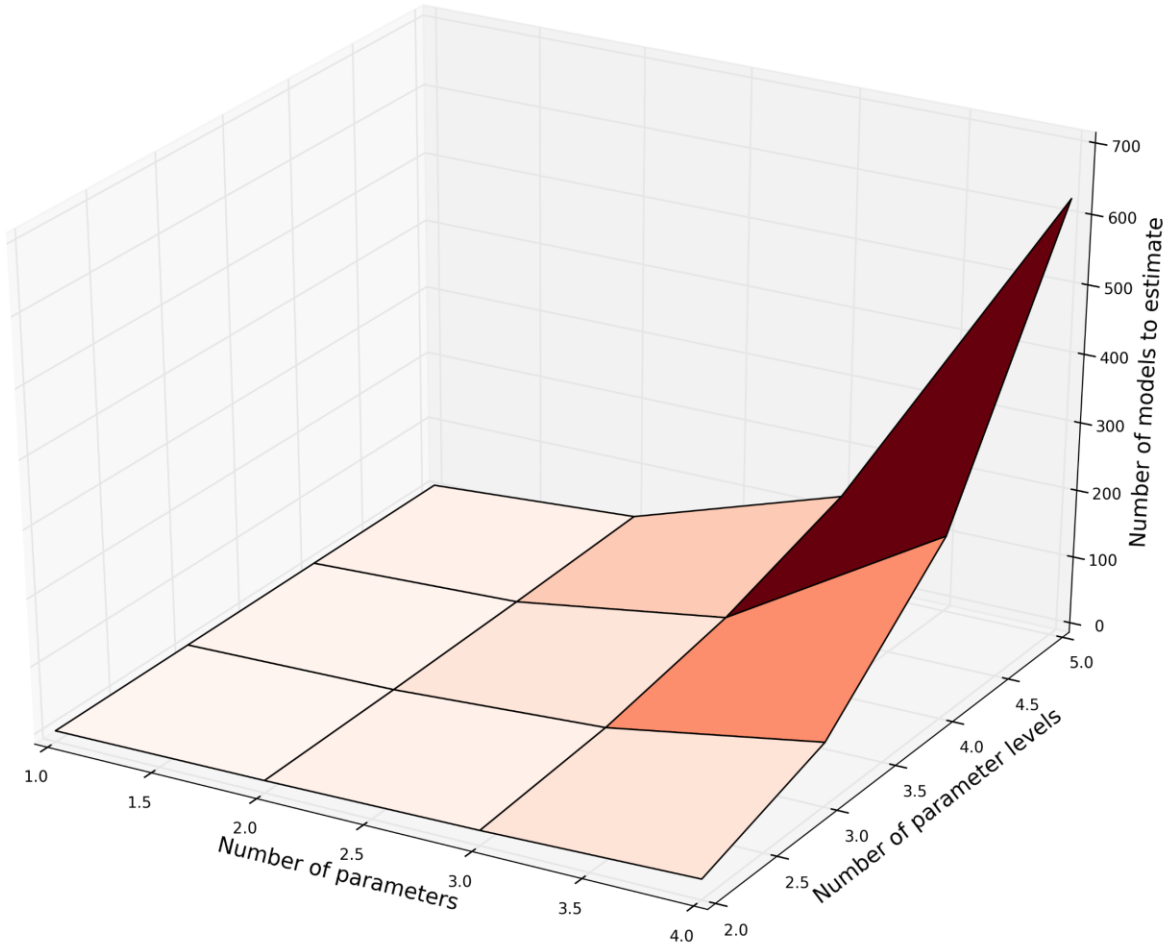
```
Area under PR: 0.85
Area under ROC: 0.63
```

Chapter 6: Introducing the ML Package


```
Out[12]: [Row(INFANT_ALIVE_AT_REPORT=0, BIRTH_PLACE='1', MOTHER_AGE_YEARS=1
         3, FATHER_COMBINED_AGE=99, CIG_BEFORE=0, CIG_1_TRI=0, CIG_2_TRI=0,
         CIG_3_TRI=0, MOTHER_HEIGHT_IN=66, MOTHER_PRE_WEIGHT=133, MOTHER_DE
         LIVERY_WEIGHT=135, MOTHER_WEIGHT_GAIN=2, DIABETES_PRE=0, DIABETES_
         GEST=0, HYP_TENS_PRE=0, HYP_TENS_GEST=0, PREV_BIRTH_PRETERM=0, BIR
         TH_PLACE_INT=1, BIRTH_PLACE_VEC=SparseVector(9, {1: 1.0}), feature
         s=SparseVector(24, {0: 13.0, 1: 99.0, 6: 66.0, 7: 133.0, 8: 135.0,
         9: 2.0, 16: 1.0}), rawPrediction=DenseVector([1.0573, -1.0573]), p
         robability=DenseVector([0.7422, 0.2578]), prediction=0.0)]
```

```
0.7401301847095617
0.7139354342365674
```

```
Out[17]: [Row(INFANT_ALIVE_AT_REPORT=0, BIRTH_PLACE='1', MOTHER_AGE_YEARS=1
         3, FATHER_COMBINED_AGE=99, CIG_BEFORE=0, CIG_1_TRI=0, CIG_2_TRI=0,
         CIG_3_TRI=0, MOTHER_HEIGHT_IN=66, MOTHER_PRE_WEIGHT=133, MOTHER_DE
         LIVERY_WEIGHT=135, MOTHER_WEIGHT_GAIN=2, DIABETES_PRE=0, DIABETES_
         GEST=0, HYP_TENS_PRE=0, HYP_TENS_GEST=0, PREV_BIRTH_PRETERM=0, BIR
         TH_PLACE_INT=1, BIRTH_PLACE_VEC=SparseVector(9, {1: 1.0}), feature
         s=SparseVector(24, {0: 13.0, 1: 99.0, 6: 66.0, 7: 133.0, 8: 135.0,
         9: 2.0, 16: 1.0}), rawPrediction=DenseVector([1.0573, -1.0573]), p
         robability=DenseVector([0.7422, 0.2578]), prediction=0.0)]
```



0.7404304424804281
0.7156729757616691

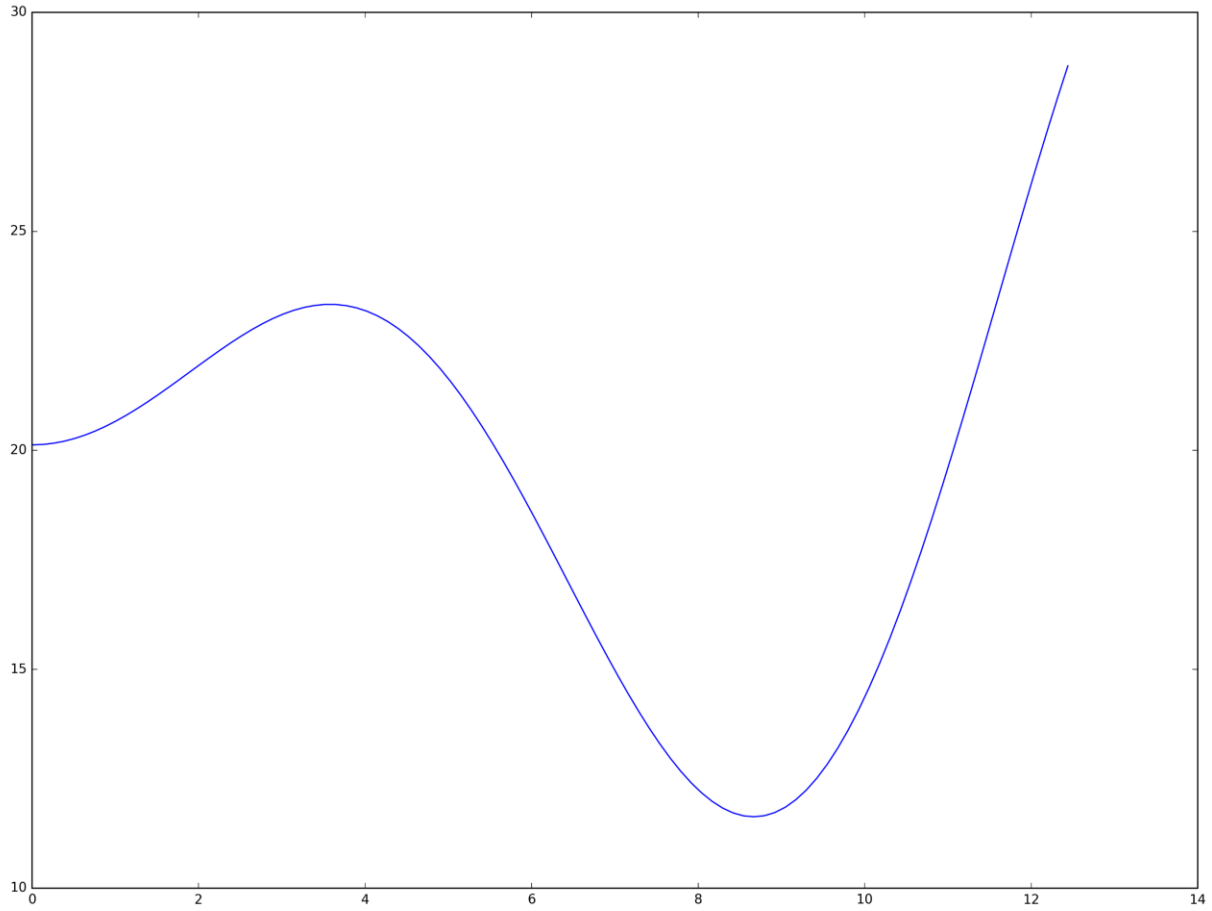
Out[27]: ([{'maxIter': 50}, {'regParam': 0.01}], 2.2158632176362274)

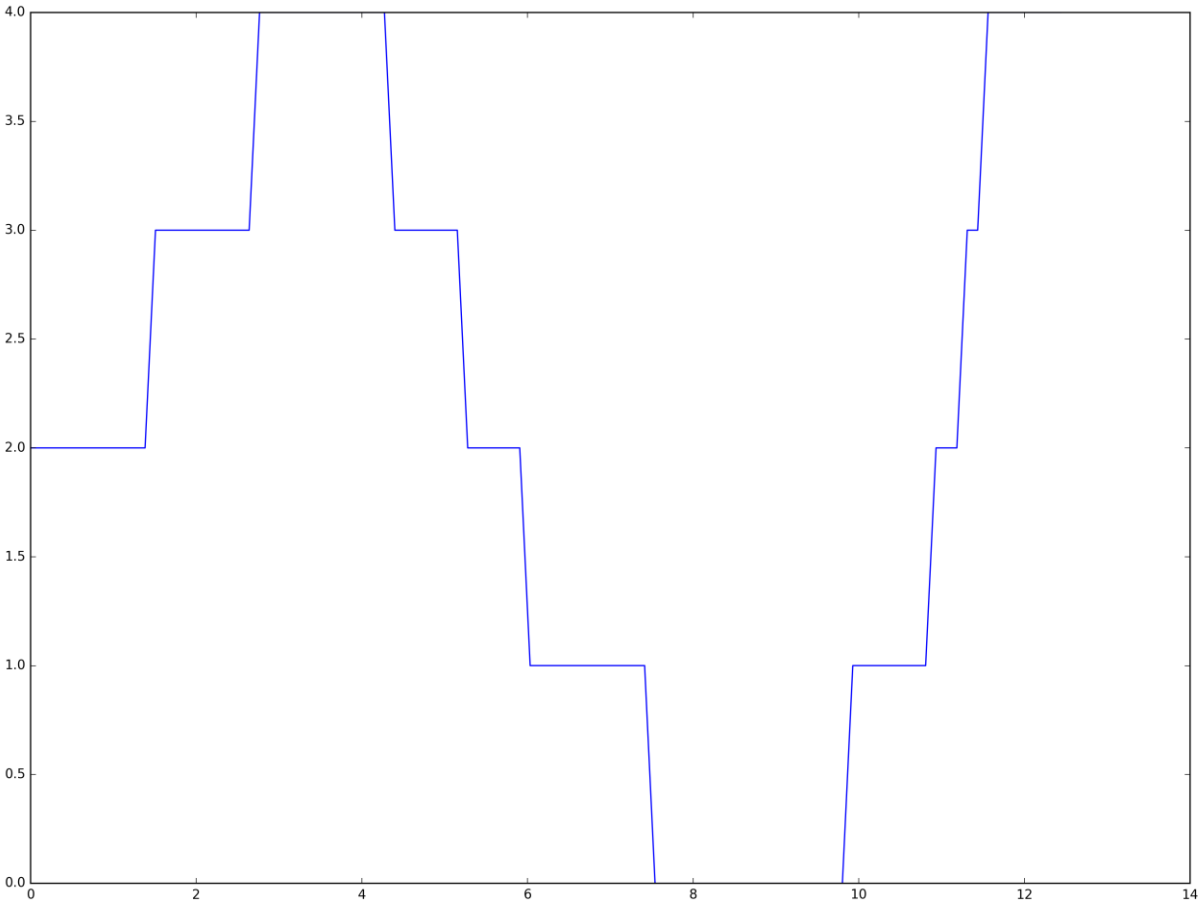
0.7334857800726642
0.7071651608758281

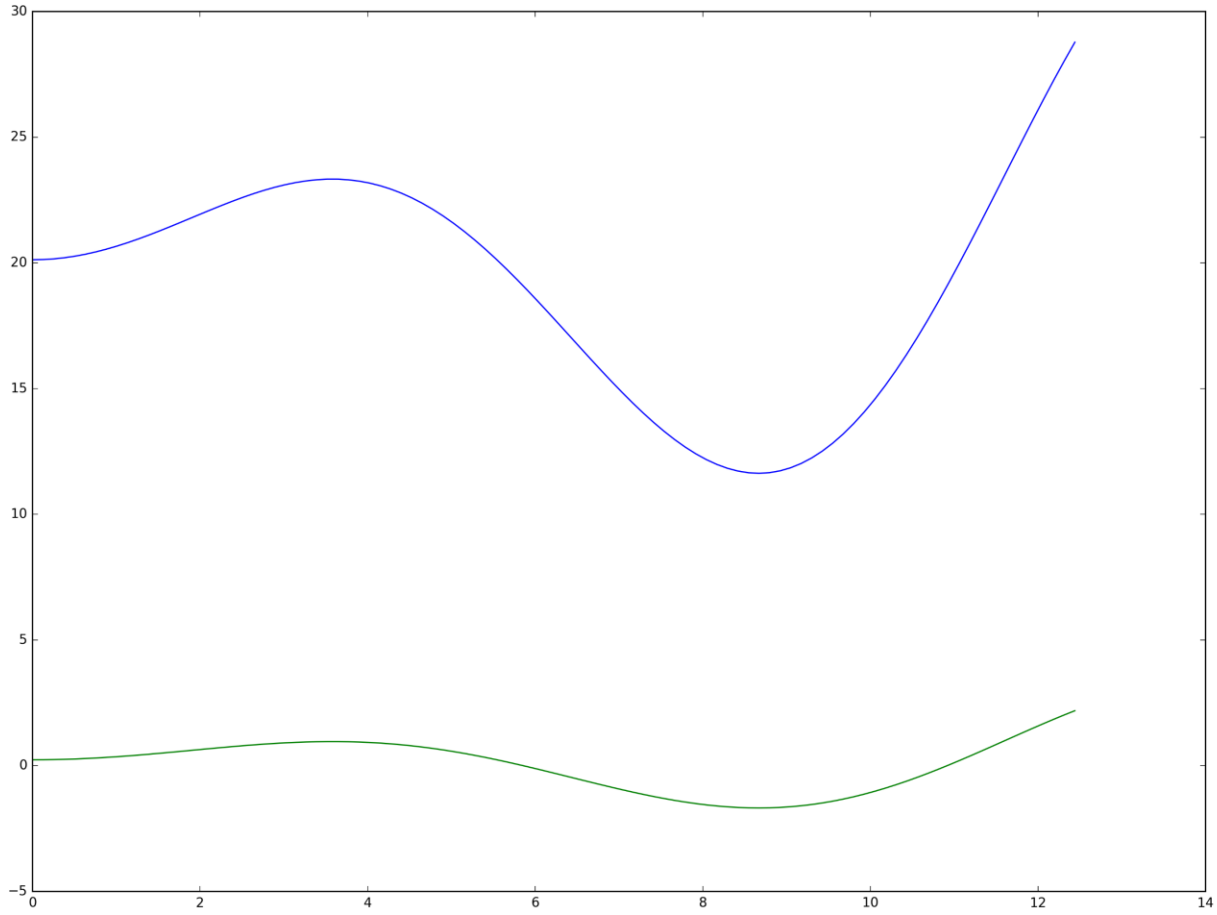
Out[35]:
```
[Row(input_arr=['machine', 'learning', 'can', 'be', 'applied', 'to
', 'a', 'wide', 'variety', 'of', 'data', 'types', 'such', 'as', 'v
ectors', 'text', 'images', 'and', 'structured', 'data', 'this', 'a
pi', 'adopts', 'the', 'dataframe', 'from', 'spark', 'sql', 'in', '
order', 'to', 'support', 'a', 'variety', 'of', 'data', 'types'])]
```

Out[37]:
```
[Row(input_stop=['machine', 'learning', 'applied', 'wide', 'variet
y', 'data', 'types', 'vectors', 'text', 'images', 'structured', 'd
ata', 'api', 'adopts', 'dataframe', 'spark', 'sql', 'order', 'supp
ort', 'variety', 'data', 'types'])]
```

Out[39]:
```
[Row(nGrams=['machine learning', 'learning applied', 'applied wide
', 'wide variety', 'variety data', 'data types', 'types vectors',
'vectors text', 'text images', 'images structured', 'structured da
ta', 'data api', 'api adopts', 'adopts dataframe', 'dataframe spar
k', 'spark sql', 'sql order', 'order support', 'support variety',
'variety data', 'data types'])]
```


30
25
20
15
10
0
2
4
6
8
10
12
14


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
2
4
6
8
10
12
14




```
0.7736428008521183  0.7582781726635287
0.7415879154340478  0.7787580540118526
```

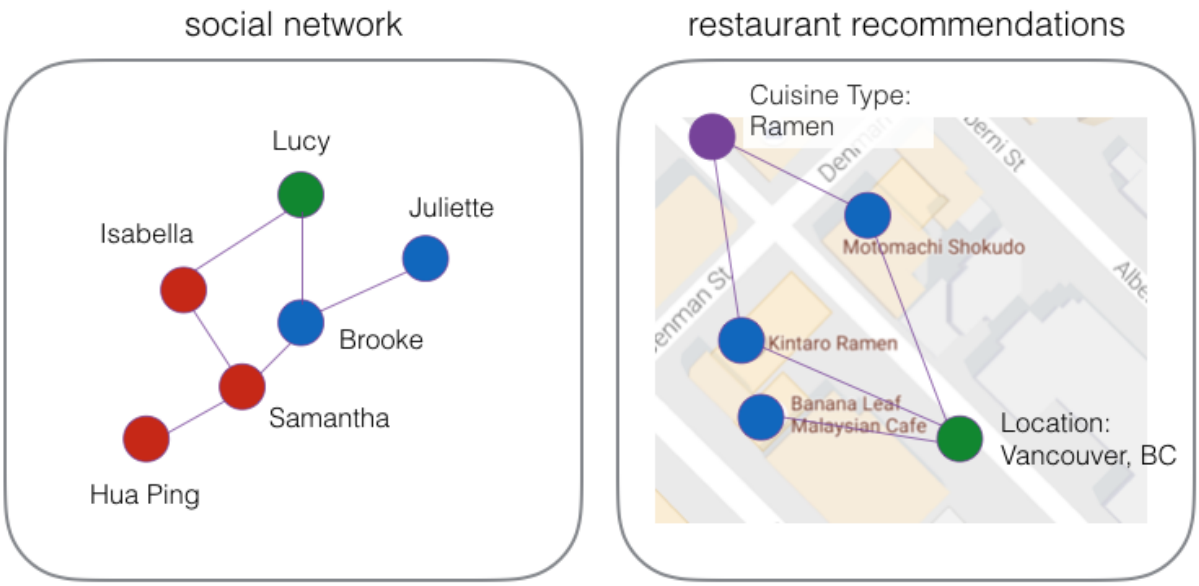
```
Out[58]: [Row(prediction=1, avg(MOTHER_HEIGHT_IN)=66.64658634538152, count(
         1)=249),
          Row(prediction=3, avg(MOTHER_HEIGHT_IN)=67.69473684210526, count(
         1)=475),
          Row(prediction=4, avg(MOTHER_HEIGHT_IN)=65.38934651290499, count(
         1)=3642),
          Row(prediction=2, avg(MOTHER_HEIGHT_IN)=83.91154791154791, count(
         1)=407),
          Row(prediction=0, avg(MOTHER_HEIGHT_IN)=63.90958873491283, count(
         1)=8948)]
```

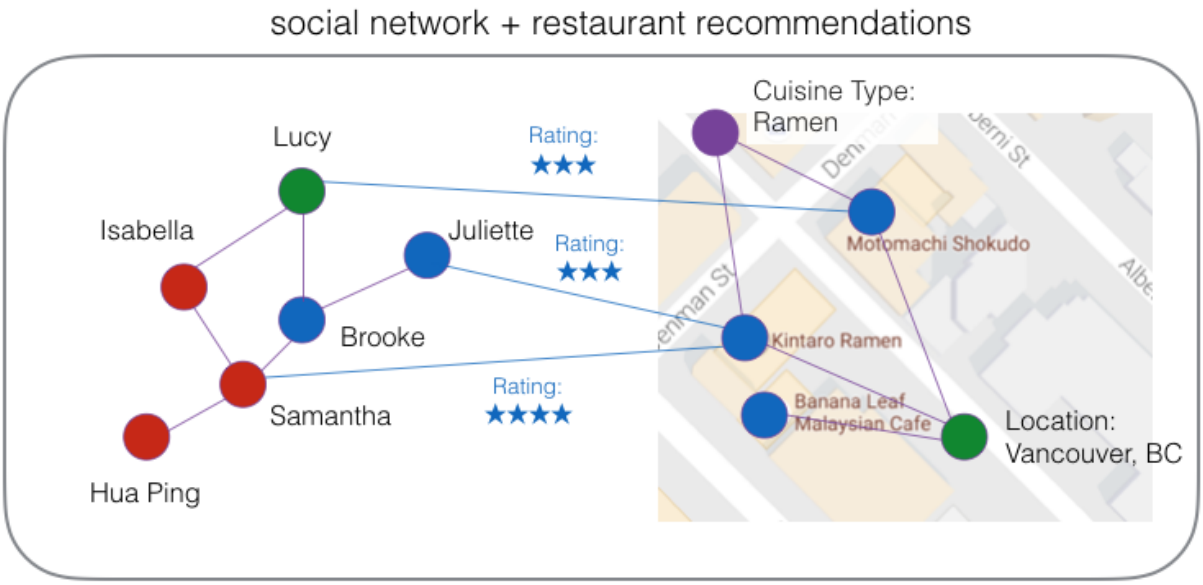
```
Out[61]: [Row(input_indexed=SparseVector(262, {2: 7.0, 6: 1.0, 8: 3.0, 10:
         3.0, 12: 3.0, 19: 1.0, 20: 1.0, 29: 1.0, 38: 1.0, 39: 2.0, 41: 2.0
         , 44: 1.0, 50: 1.0, 60: 1.0, 65: 1.0, 87: 1.0, 108: 1.0, 110: 1.0,
         112: 1.0, 114: 1.0, 116: 1.0, 139: 1.0, 149: 1.0, 150: 1.0, 162: 1
         .0, 181: 1.0, 182: 1.0, 190: 1.0, 193: 1.0, 218: 1.0, 226: 1.0, 23
         0: 1.0, 232: 1.0, 249: 1.0, 251: 1.0, 256: 1.0})),
          Row(input_indexed=SparseVector(262, {20: 1.0, 21: 1.0, 22: 2.0, 3
         2: 2.0, 33: 2.0, 36: 2.0, 48: 1.0, 49: 1.0, 55: 1.0, 63: 1.0, 72:
         1.0, 73: 1.0, 77: 1.0, 83: 1.0, 88: 1.0, 90: 1.0, 93: 1.0, 102: 1.
         0, 105: 1.0, 111: 1.0, 122: 1.0, 128: 1.0, 130: 1.0, 140: 1.0, 145
         : 1.0, 146: 1.0, 170: 1.0, 173: 1.0, 195: 1.0, 196: 1.0, 202: 1.0,
         203: 1.0, 207: 1.0, 209: 1.0, 212: 1.0, 213: 1.0, 216: 1.0, 221: 1
         .0, 224: 1.0, 225: 1.0, 228: 1.0, 231: 1.0, 237: 1.0, 241: 1.0, 24
         6: 1.0, 247: 1.0, 255: 1.0, 260: 1.0}))]
```

```
Out[65]: [Row(topicDistribution=DenseVector([0.0221, 0.9779])),
          Row(topicDistribution=DenseVector([0.0171, 0.9829])),
          Row(topicDistribution=DenseVector([0.0199, 0.9801])),
          Row(topicDistribution=DenseVector([0.9923, 0.0077])),
          Row(topicDistribution=DenseVector([0.9925, 0.0075])),
          Row(topicDistribution=DenseVector([0.9904, 0.0096]))]
```

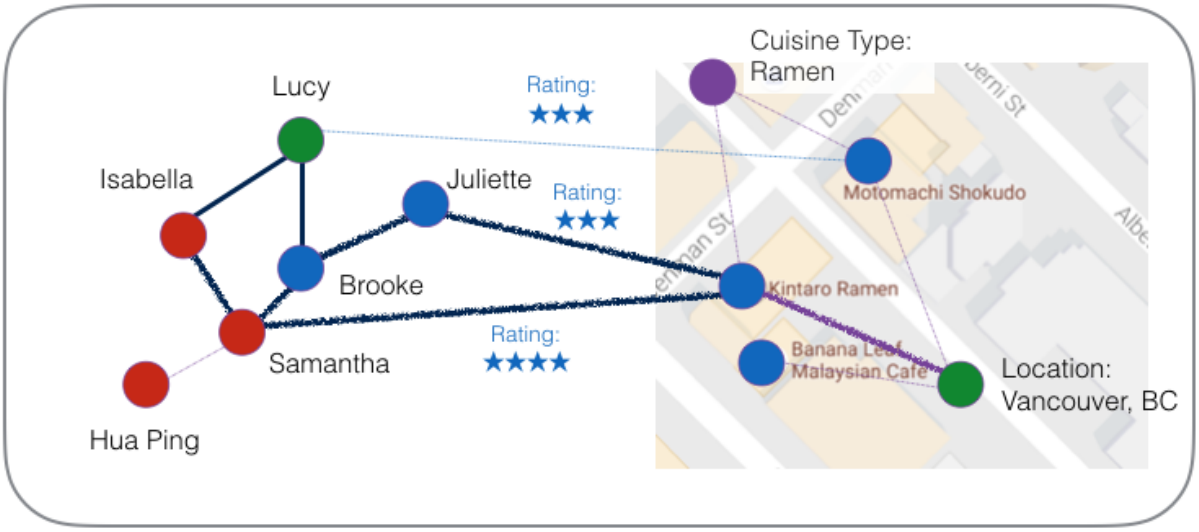
```
0.48862170400240335
```

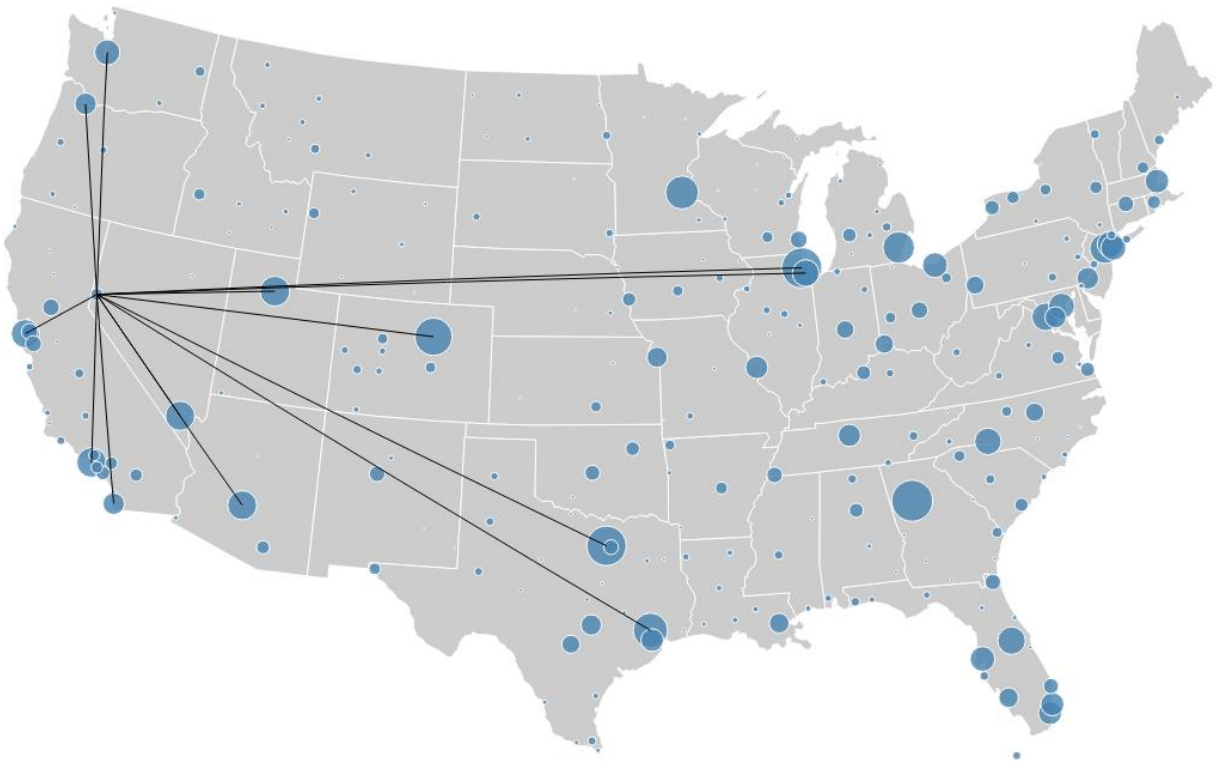
Chapter 7: GraphFrames

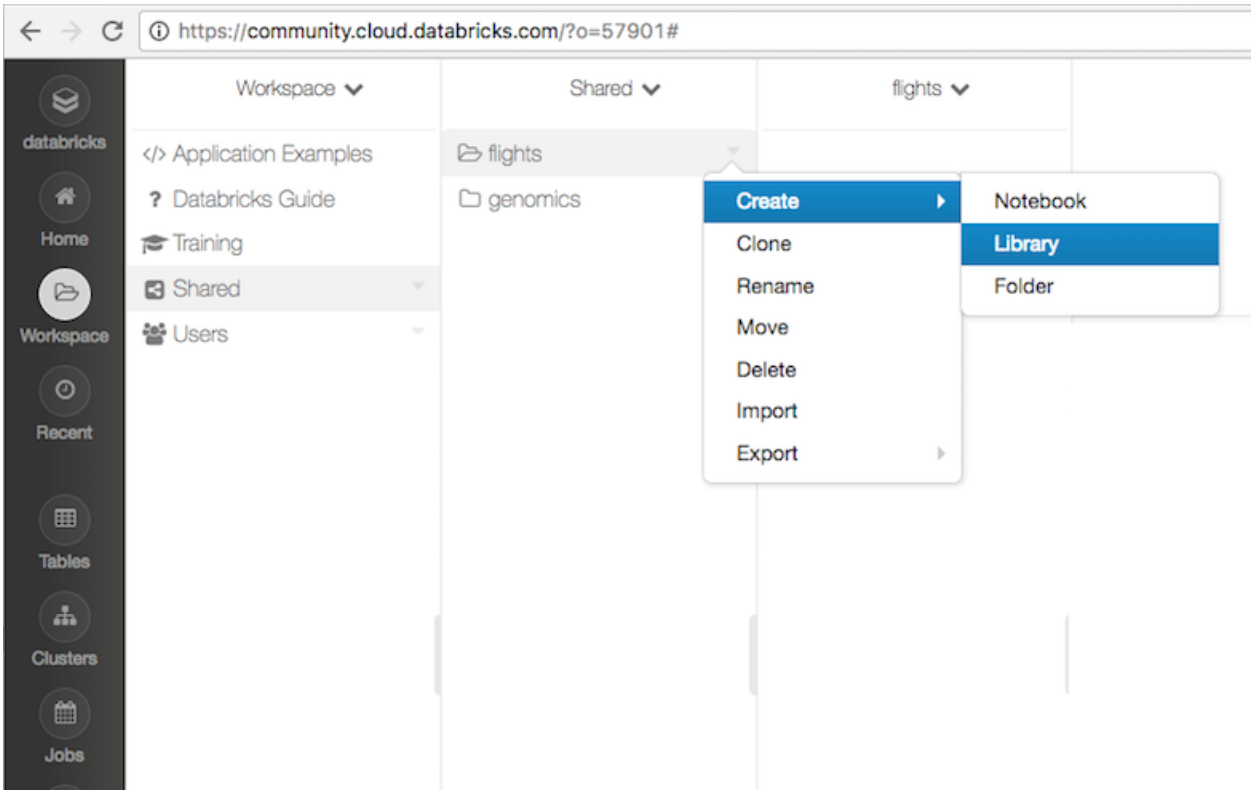


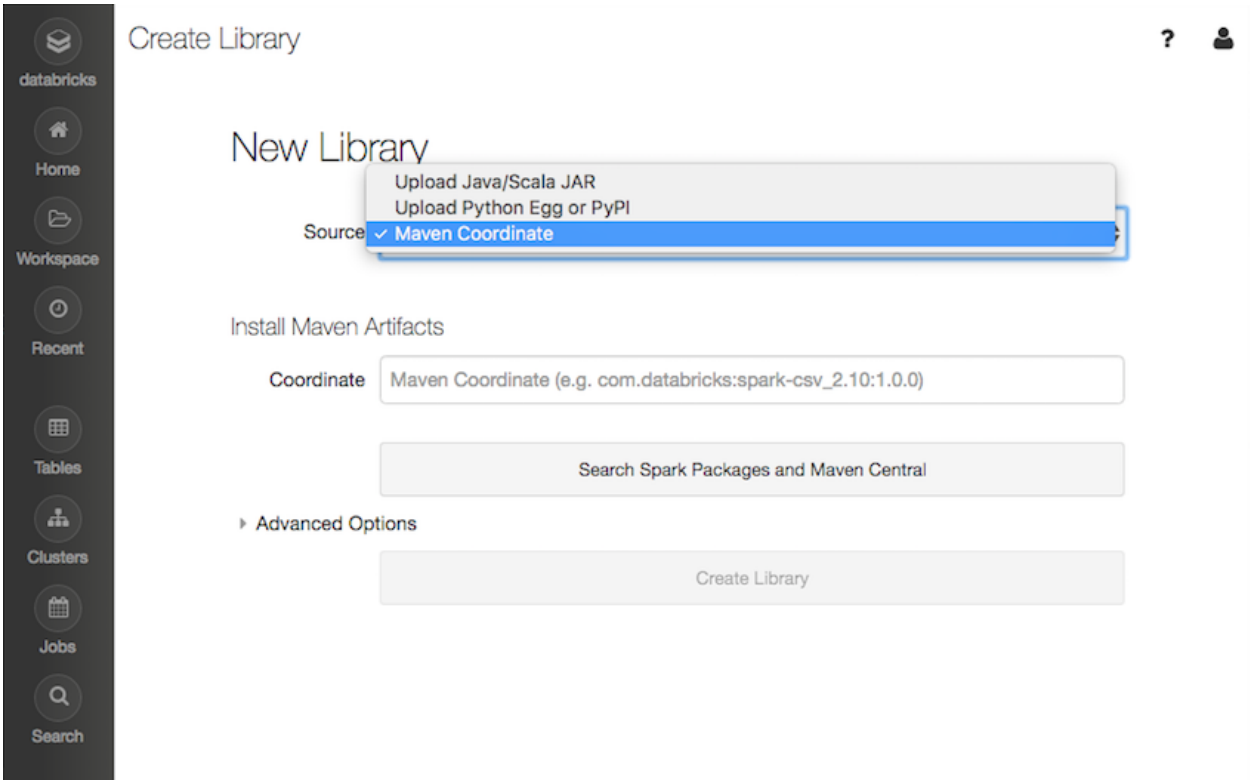


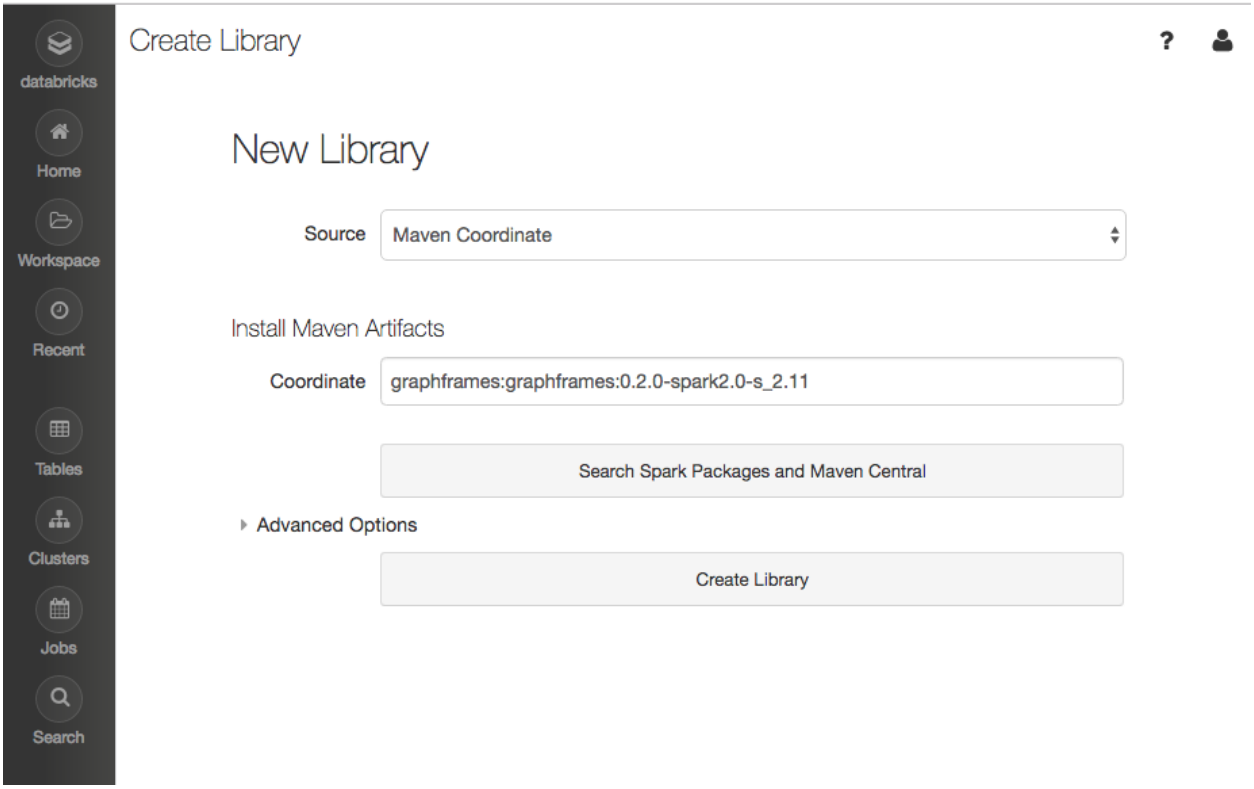
social network + restaurant recommendations

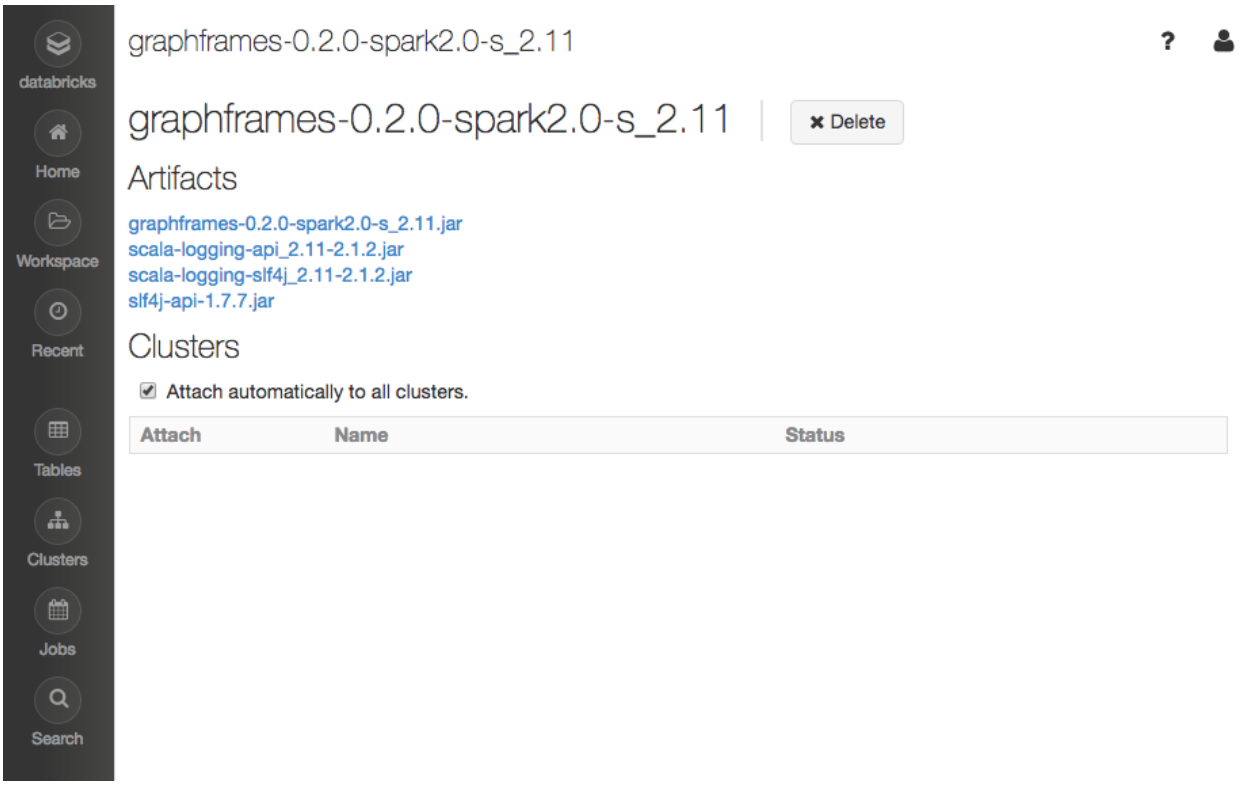





https://community.cloud.databricks.com/?o=57901#
databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Workspace
Shared
flights
Application Examples
Databricks Guide
Training
Shared
Users
flights
genomics
Create
Clone
Rename
Move
Delete
Import
Export
Notebook
Library
Folder


databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
Create Library
New Library
Upload Java/Scala JAR
Upload Python Egg or PyPI
Source
Maven Coordinate
Install Maven Artifacts
Coordinate
Maven Coordinate (e.g. com.databricks:spark-csv_2.10:1.0.0)
Search Spark Packages and Maven Central
Advanced Options
Create Library


databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
Create Library
New Library
Source
Maven Coordinate
Install Maven Artifacts
Coordinate
graphframes:graphframes:0.2.0-spark2.0-s_2.11
Search Spark Packages and Maven Central
Advanced Options
Create Library


databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
graphframes-0.2.0-spark2.0-s_2.11
graphframes-0.2.0-spark2.0-s_2.11
Delete
Artifacts
graphframes-0.2.0-spark2.0-s_2.11.jar
scala-logging-api_2.11-2.1.2.jar
scala-logging-slf4j_2.11-2.1.2.jar
slf4j-api-1.7.7.jar
Clusters
Attach automatically to all clusters.
Attach
Name
Status

▸ (2) Spark Jobs

```
+-------+-------------------+-----+--------+---+---+----------+------------------+---------+---------+
| tripid|          localdate|delay|distance|src|dst|  city_src|          city_dst|state_src|state_dst|
+-------+-------------------+-----+--------+---+---+----------+------------------+---------+---------+
|1011111|2014-01-01 11:11:...|   -5|     221|MSP|INL|Minneapolis|International Falls|       MN|       MN|
|1021111|2014-01-02 11:11:...|    7|     221|MSP|INL|Minneapolis|International Falls|       MN|       MN|
|1031111|2014-01-03 11:11:...|    0|     221|MSP|INL|Minneapolis|International Falls|       MN|       MN|
|1041925|2014-01-04 19:25:...|    0|     221|MSP|INL|Minneapolis|International Falls|       MN|       MN|
|1061115|2014-01-06 11:15:...|   33|     221|MSP|INL|Minneapolis|International Falls|       MN|       MN|
|1071115|2014-01-07 11:15:...|   23|     221|MSP|INL|Minneapolis|International Falls|       MN|       MN|
|1081115|2014-01-08 11:15:...|   -9|     221|MSP|INL|Minneapolis|International Falls|       MN|       MN|
|1091115|2014-01-09 11:15:...|   11|     221|MSP|INL|Minneapolis|International Falls|       MN|       MN|
|1101115|2014-01-10 11:15:...|   -3|     221|MSP|INL|Minneapolis|International Falls|       MN|       MN|
|1112015|2014-01-11 20:15:...|   -7|     221|MSP|INL|Minneapolis|International Falls|       MN|       MN|
+-------+-------------------+-----+--------+---+---+----------+------------------+---------+---------+
only showing top 10 rows
```

▸ (2) Spark Jobs

tripid	delay	src	dst	city_dst	state_dst
1011111	-5	MSP	INL	International Falls	MN
1021111	7	MSP	INL	International Falls	MN
1031111	0	MSP	INL	International Falls	MN
1041925	0	MSP	INL	International Falls	MN
1061115	33	MSP	INL	International Falls	MN
1071115	23	MSP	INL	International Falls	MN
1081115	-9	MSP	INL	International Falls	MN
1091115	11	MSP	INL	International Falls	MN
1101115	-3	MSP	INL	International Falls	MN

Showing the first 1000 rows.

▾ (2) Spark Jobs

- ▸ Job 16 View (Stages: 2/2, 4 skipped)
- ▸ Job 17 View (Stages: 2/2, 7 skipped)

```
Airports: 279
Trips: 1361141
```

▾ (2) Spark Jobs

- ▸ Job 18 View (Stages: 2/2, 7 skipped)
- ▸ Job 19 View (Stages: 2/2, 7 skipped)

```
On-time / Early Flights: 780469
Delayed Flights: 580672
```

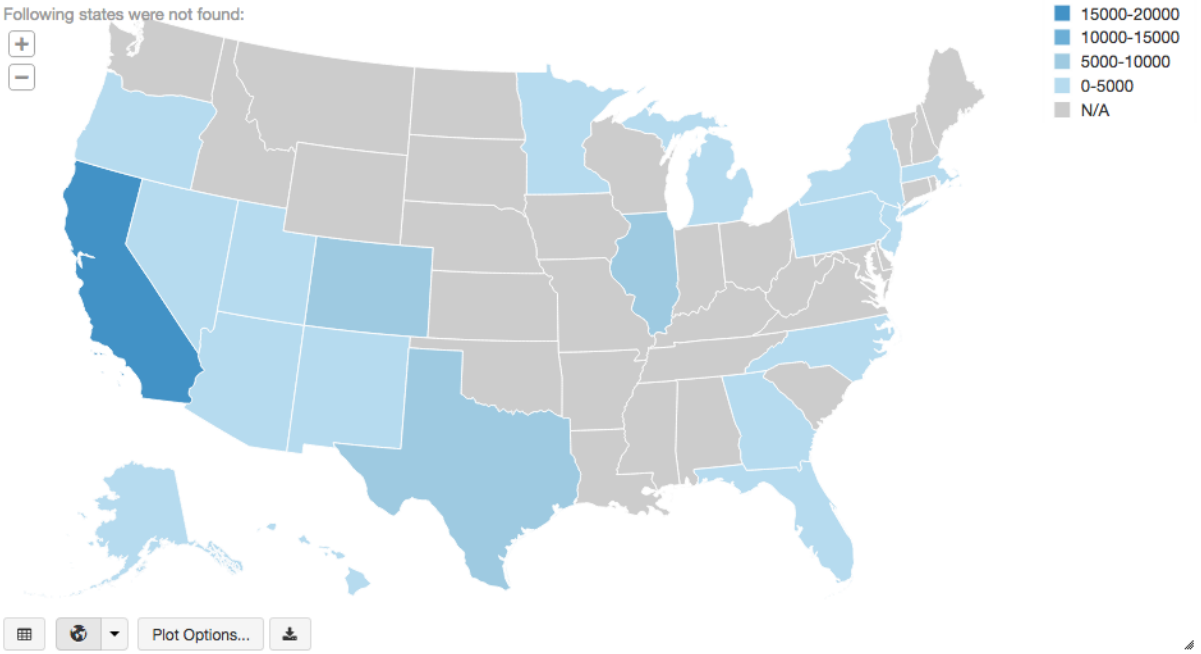
▸ (1) Spark Jobs

```
+---+---+------------------+
|src|dst|        avg(delay)|
+---+---+------------------+
|SEA|PHL|55.666666666666664|
|SEA|COS| 43.53846153846154|
|SEA|FAT| 43.03846153846154|
|SEA|LGB| 39.39705882352941|
|SEA|IAD|37.733333333333334|
+---+---+------------------+
only showing top 5 rows
```

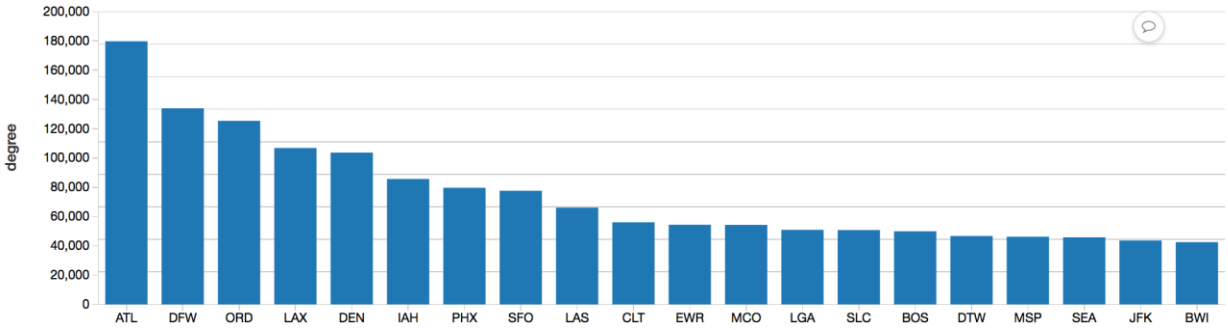
▸ (2) Spark Jobs

tripid	delay	src	dst	city_dst	state_dst
3201938	108	SEA	BUR	Burbank	CA
3201655	107	SEA	SNA	Orange County	CA
1011950	123	SEA	OAK	Oakland	CA
1021950	194	SEA	OAK	Oakland	CA
1021615	317	SEA	OAK	Oakland	CA
1021755	385	SEA	OAK	Oakland	CA
1031950	283	SEA	OAK	Oakland	CA
1031615	364	SEA	OAK	Oakland	CA
1031325	130	SEA	OAK	Oakland	CA
1061755	107	SEA	OAK	Oakland	CA

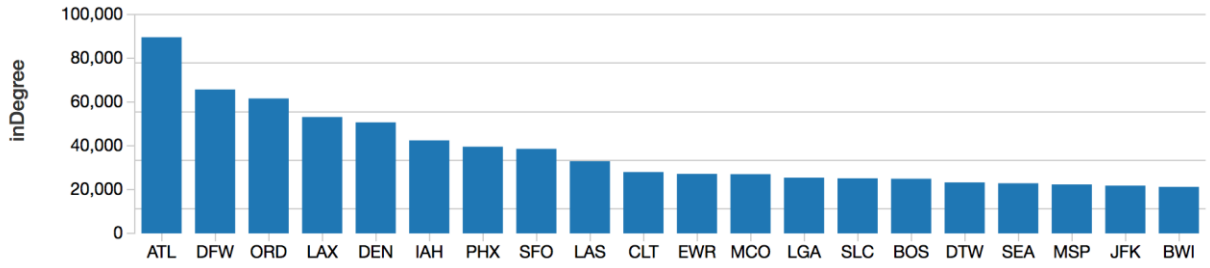
▸ (2) Spark Jobs



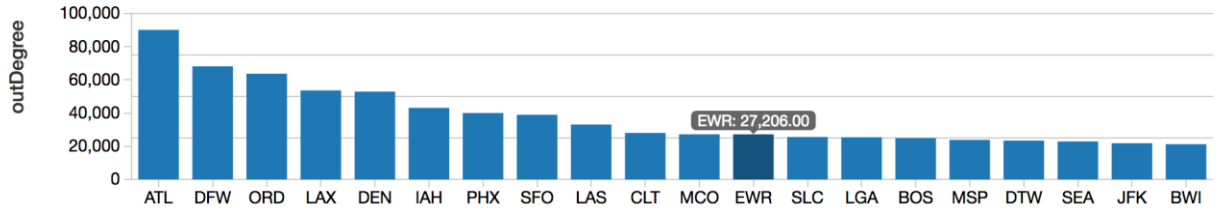
▸ (1) Spark Jobs



▸ (1) Spark Jobs



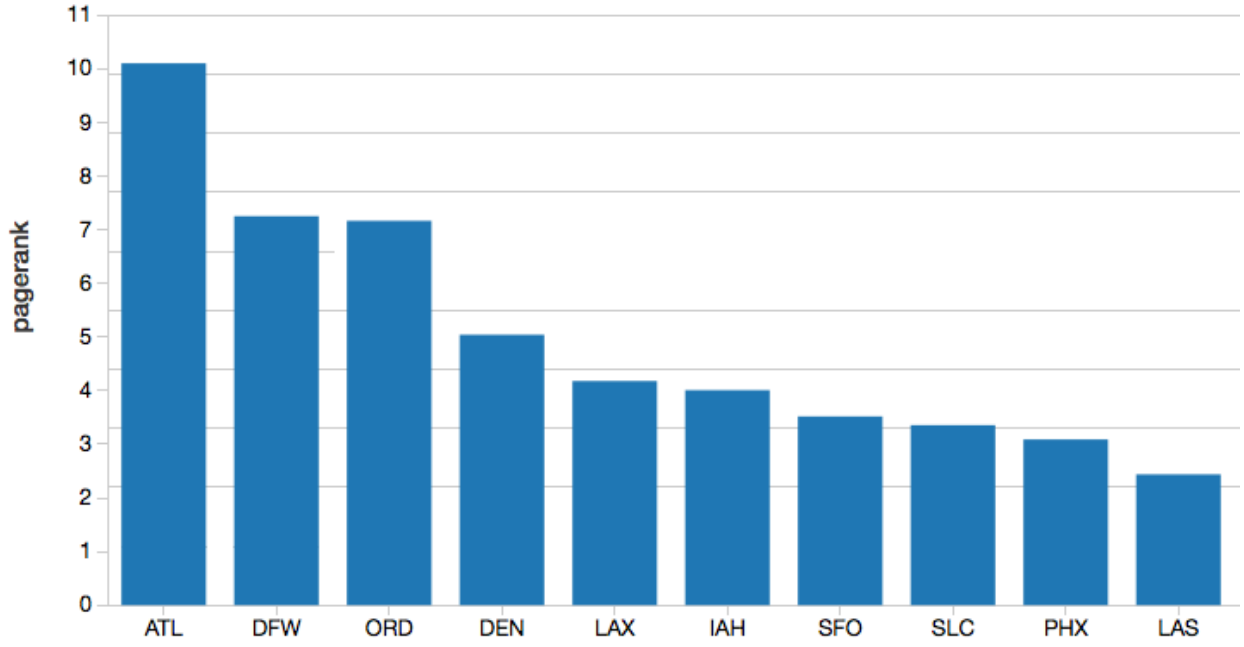
▸ (1) Spark Jobs



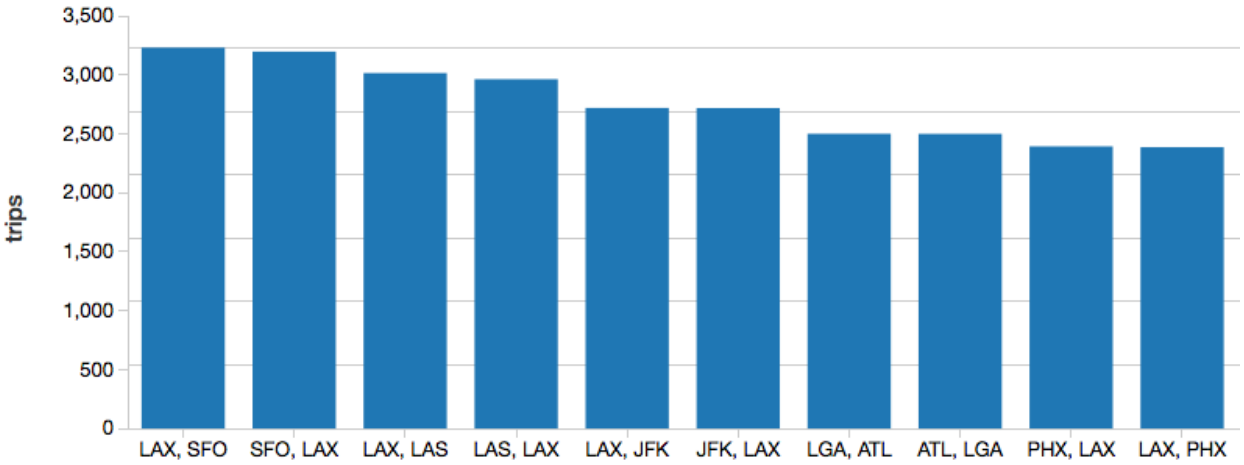
▸ (8) Spark Jobs

a	ab	b	bc	c
[MSY,New Orleans,...]	[1011751,-4,MSY,SFO]	[SFO,San Francisc...]	[1021507,536,SFO,...]	[JFK,New York,NY,...]
[MSY,New Orleans,...]	[1201725,2,MSY,SFO]	[SFO,San Francisc...]	[1211508,593,SFO,...]	[JFK,New York,NY,...]
[MSY,New Orleans,...]	[2091725,87,MSY,SFO]	[SFO,San Francisc...]	[2092110,740,SFO,...]	[MIA,Miami,FL,USA]
[MSY,New Orleans,...]	[2091725,87,MSY,SFO]	[SFO,San Francisc...]	[2092230,636,SFO,...]	[JFK,New York,NY,...]
[MSY,New Orleans,...]	[2121725,15,MSY,SFO]	[SFO,San Francisc...]	[2131420,504,SFO,...]	[SAN,San Diego,CA...]
[BUR,Burbank,CA,USA]	[1011828,88,BUR,SFO]	[SFO,San Francisc...]	[1021507,536,SFO,...]	[JFK,New York,NY,...]
[BUR,Burbank,CA,USA]	[1020941,-17,BUR,...	[SFO,San Francisc...]	[1021507,536,SFO,...]	[JFK,New York,NY,...]
[BUR,Burbank,CA,USA]	[1020705,6,BUR,SFO]	[SFO,San Francisc...]	[1021507,536,SFO,...]	[JFK,New York,NY,...]
[BUR,Burbank,CA,USA]	[1021320,-5,BUR,SFO]	[SFO,San Francisc...]	[1021507,536,SFO,...]	[JFK,New York,NY,...]
[BUR,Burbank,CA,USA]	[1202011,-3,BUR,SFO]	[SFO,San Francisc...]	[1211508,593,SFO,...]	[JFK,New York,NY,...]

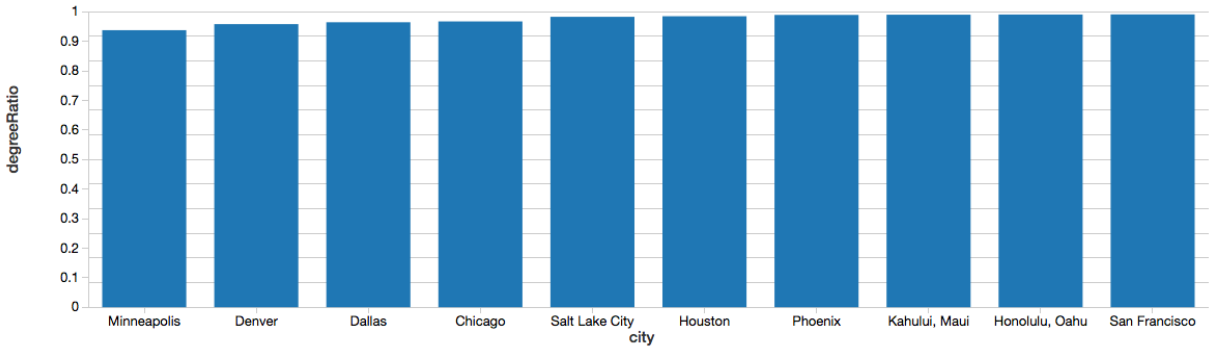
▸ (6) Spark Jobs



▸ (1) Spark Jobs

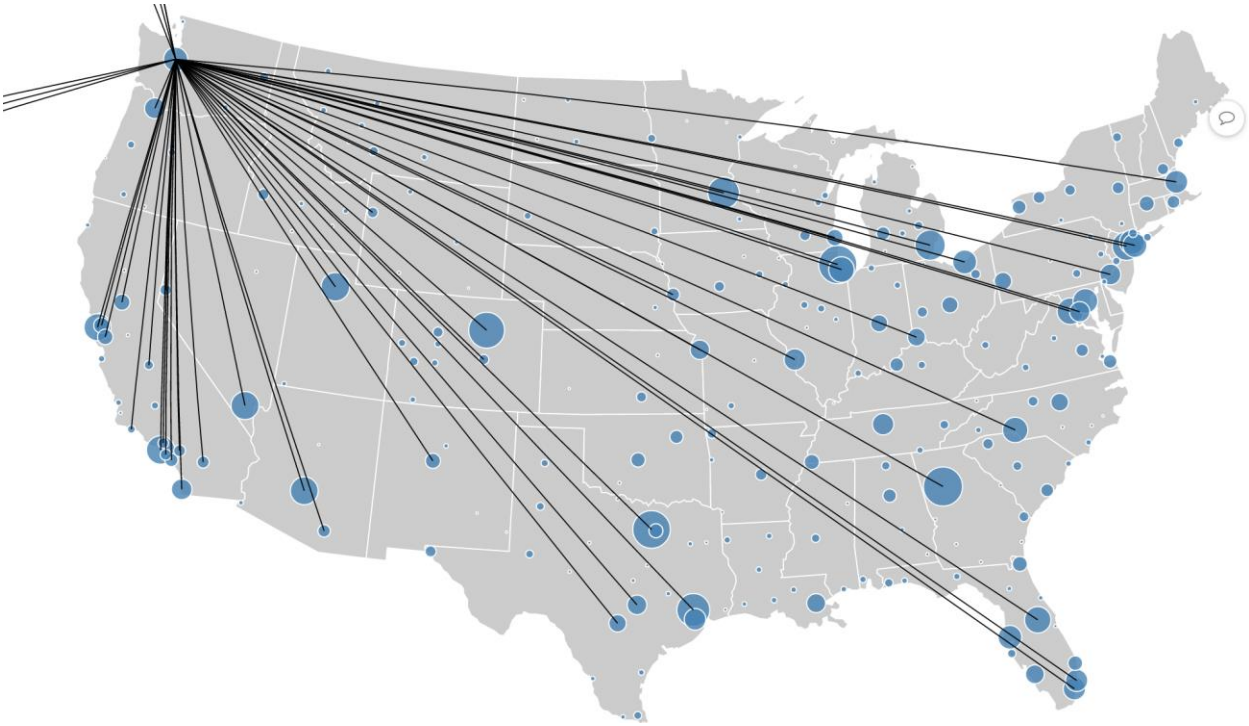


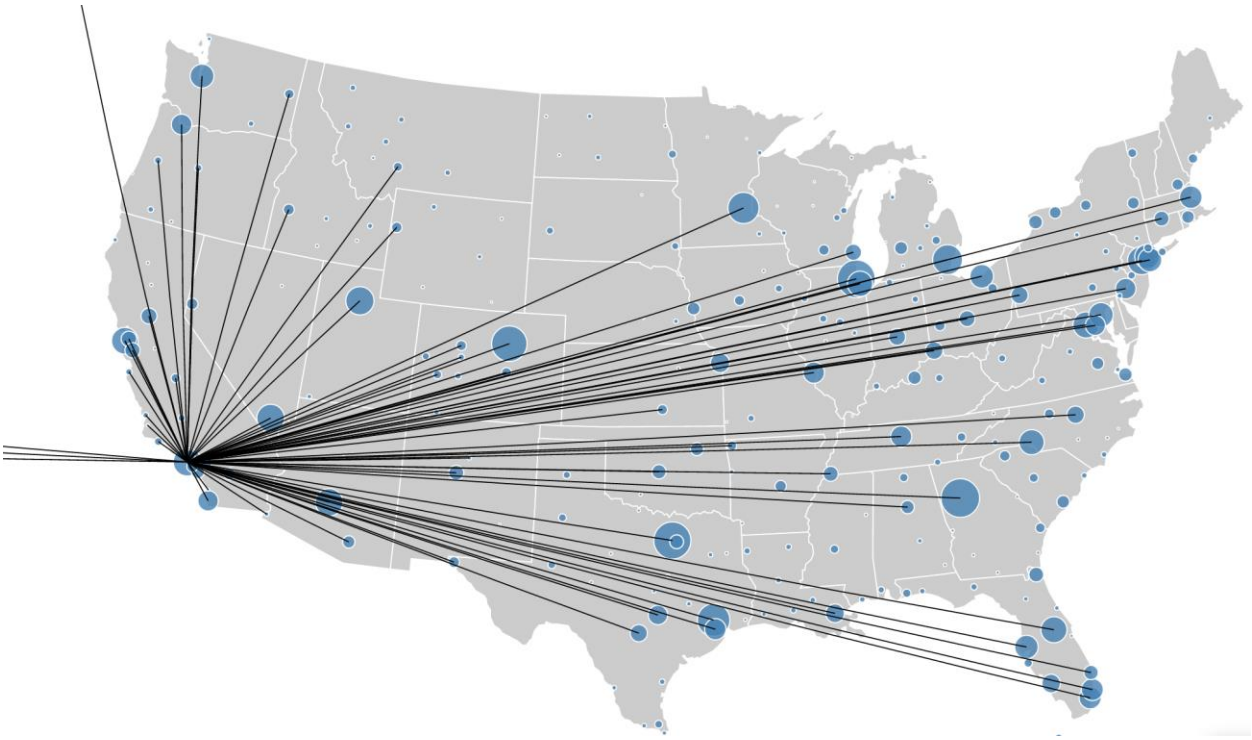
▸ (2) Spark Jobs

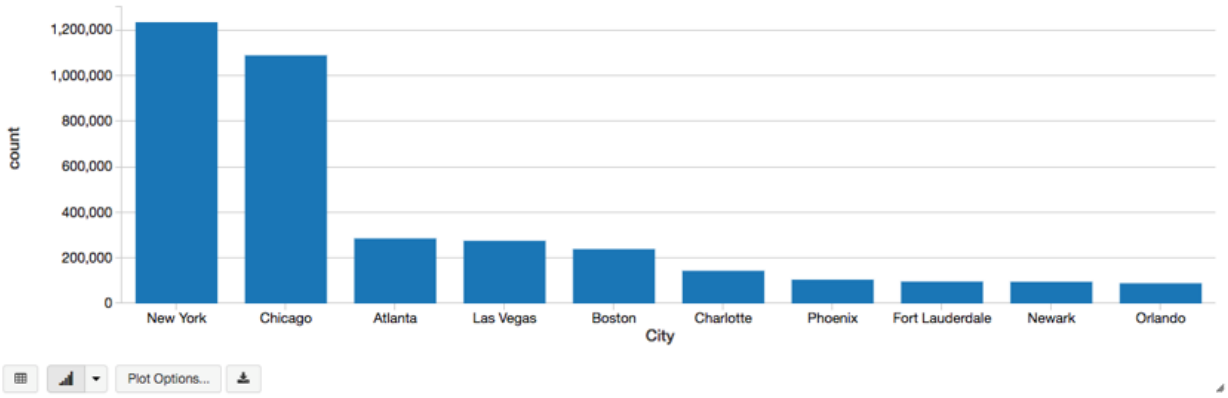


▸ (14) Spark Jobs

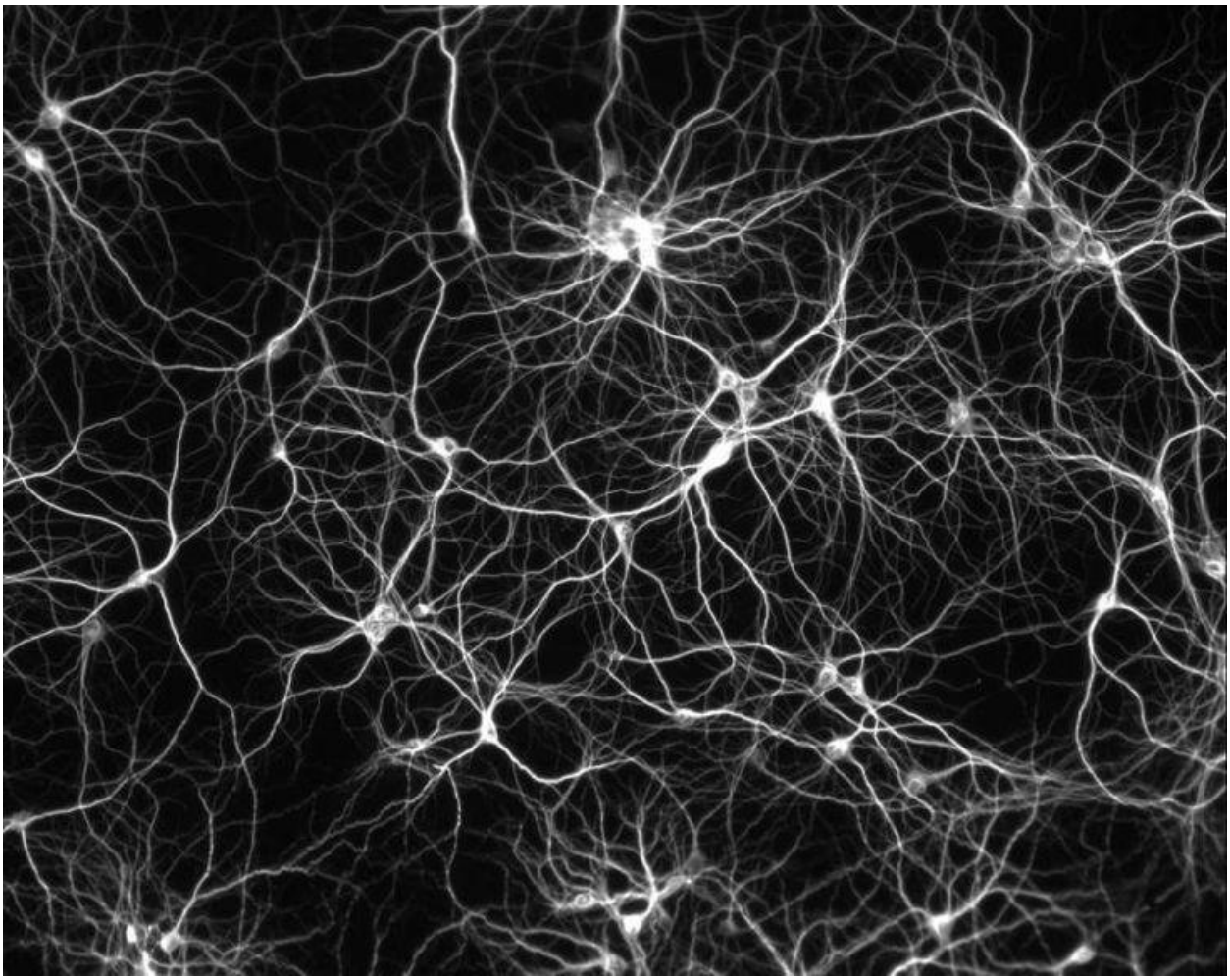
from	e0	to
▸ {"id":"SEA","City":"Seattle","State":"WA","Country":"USA"}	▸ {"tripid":1010710,"delay":31,"src":"SEA","dst":"SFO","city_dst":"San Francisco","state_dst":"CA"}	▸ {"id":"SFO","City":"San Francisco","State":"CA","Country":"USA"}
▸ {"id":"SEA","City":"Seattle","State":"WA","Country":"USA"}	▸ {"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"}

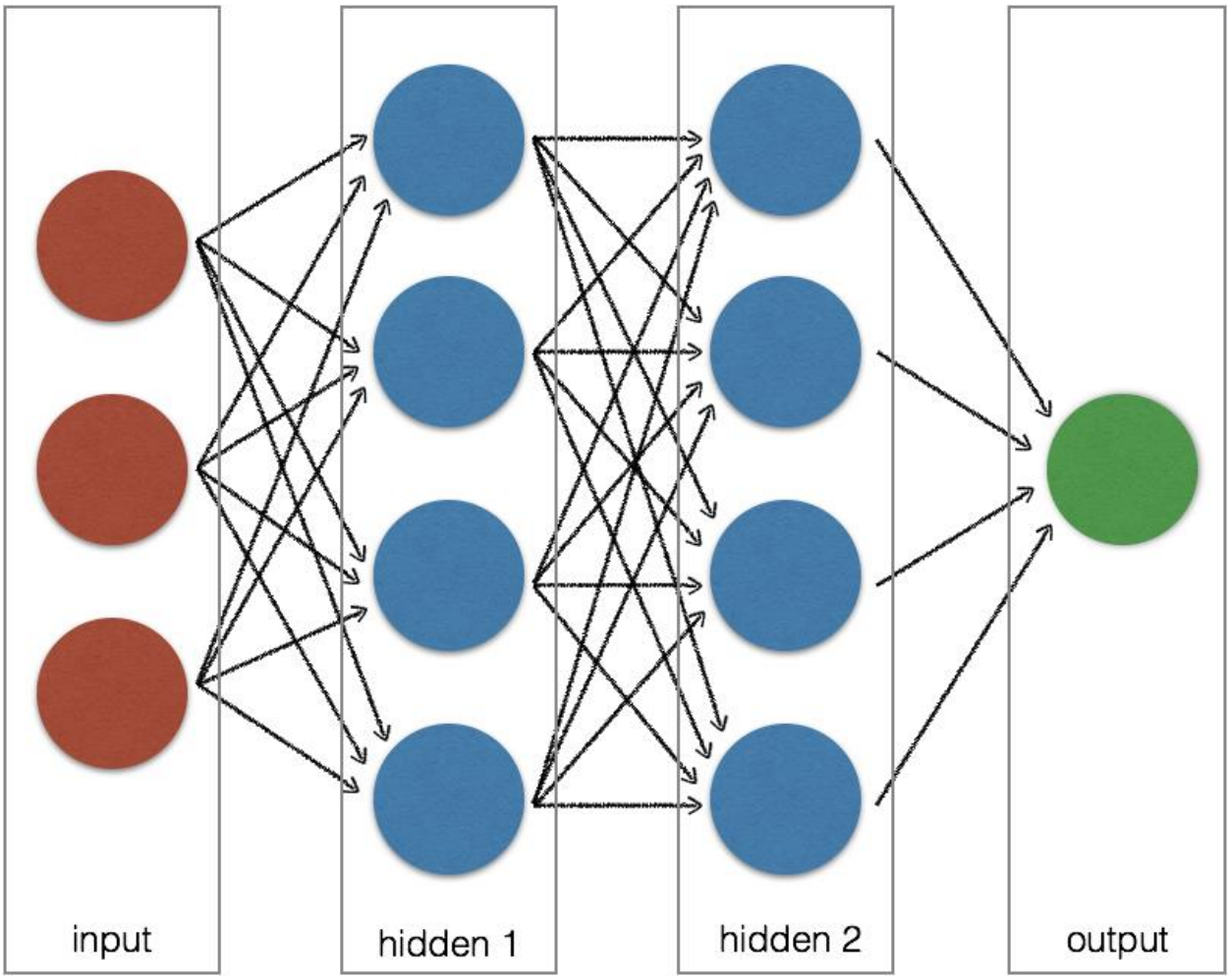


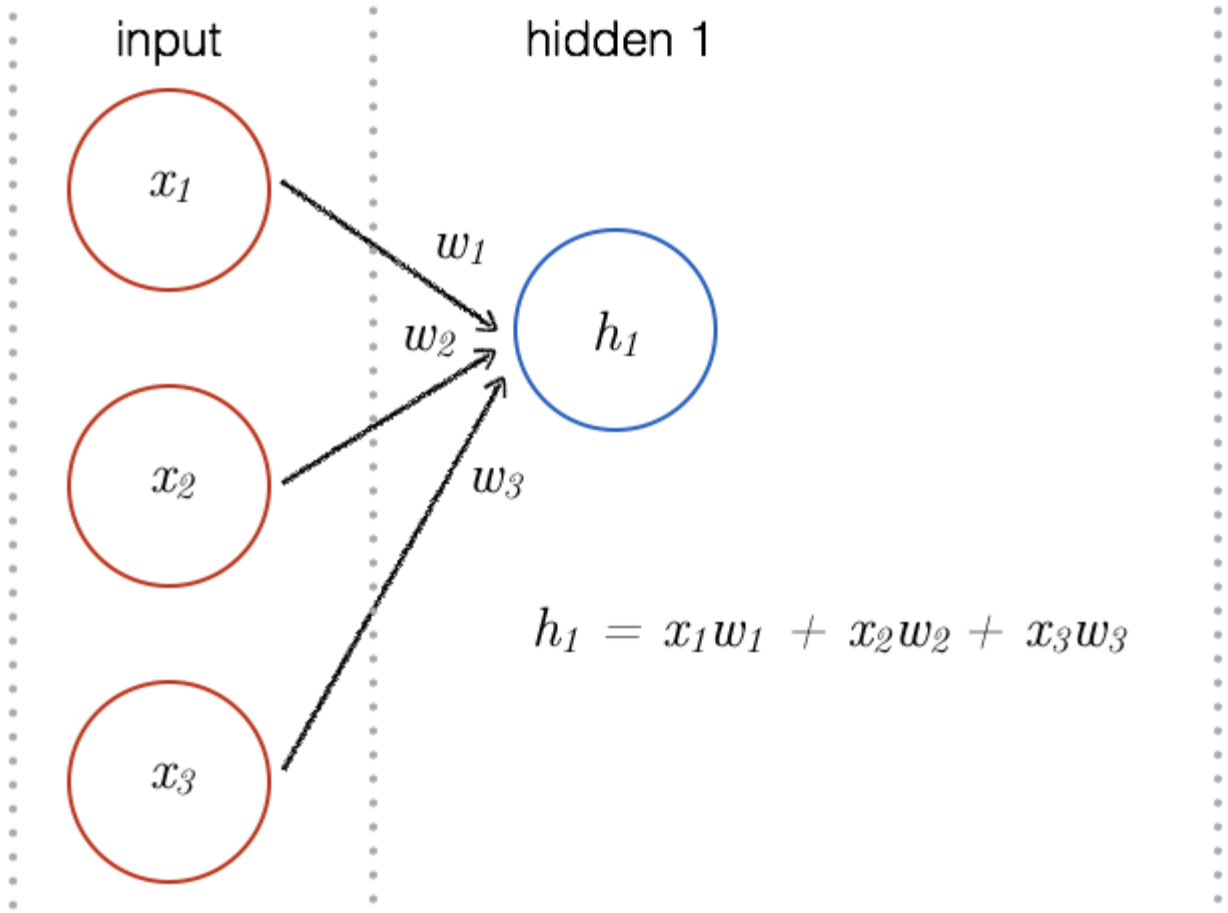



1,200,000
1,000,000
800,000
600,000
400,000
200,000
0
count
New York
Chicago
Atlanta
Las Vegas
Boston
Charlotte
Phoenix
Fort Lauderdale
Newark
Orlando
City
Plot Options...

Chapter 8: TensorFrames




input
hidden 1
hidden 2
output



$$h_1 = x_1 w_1 + x_2 w_2 + x_3 w_3 \, y_i = a + bx_2 + \ldots + bx_n + e_i$$

Potential Features

location
cuisine type
location size
takeout?
good for kids?
good for business?
reservations?

How to choose which features?

Analysis via BI tools or spreadsheets?
Use machine learning algorithms?

classification

algorithm

regression

Potential Features

location
cuisine type
location size
takeout?
good for kids?
good for business?
reservations?

Segment: High End Restaurant

1

location
cuisine type

3

location size
takeout?
good for kids?
good for business?
reservations?

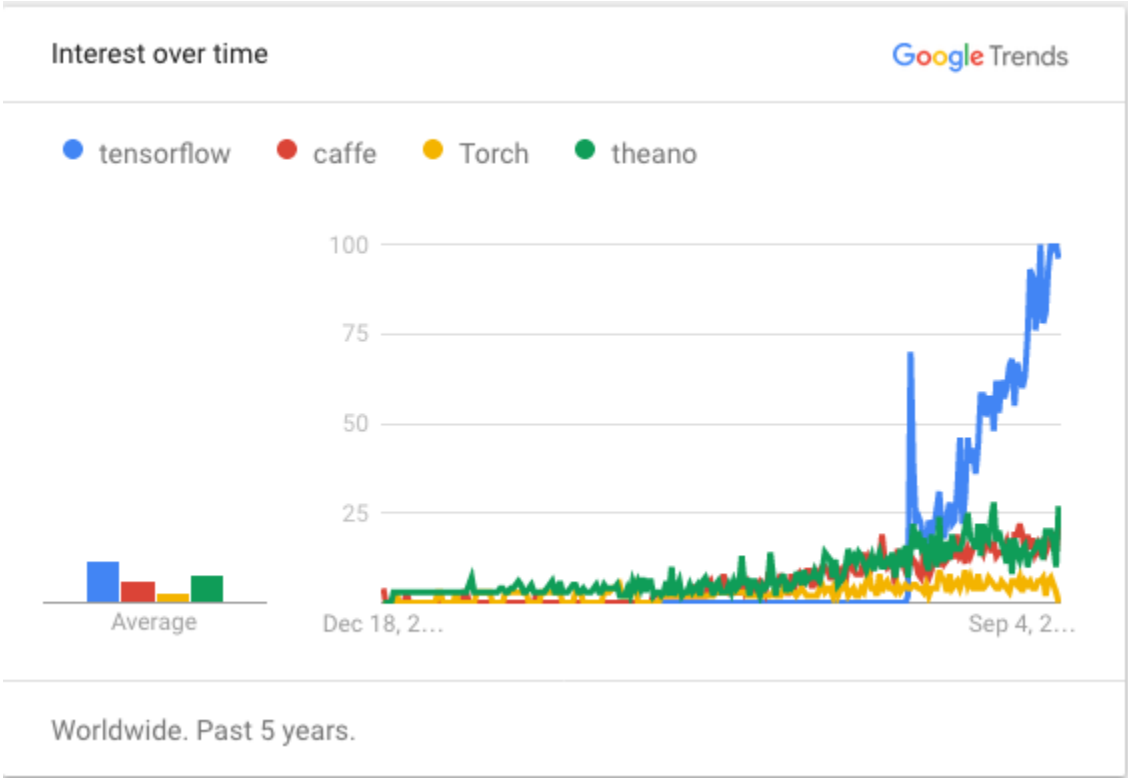
2

Potential Features

location
cuisine type
takeout?
good for kids?
reservations?
ratings
reviews
buzz

Segment: Speciality Restaurant

location
cuisine type 1
takeout?
good for kids?
reservations? 4
ratings
3 *reviews*
buzz 2

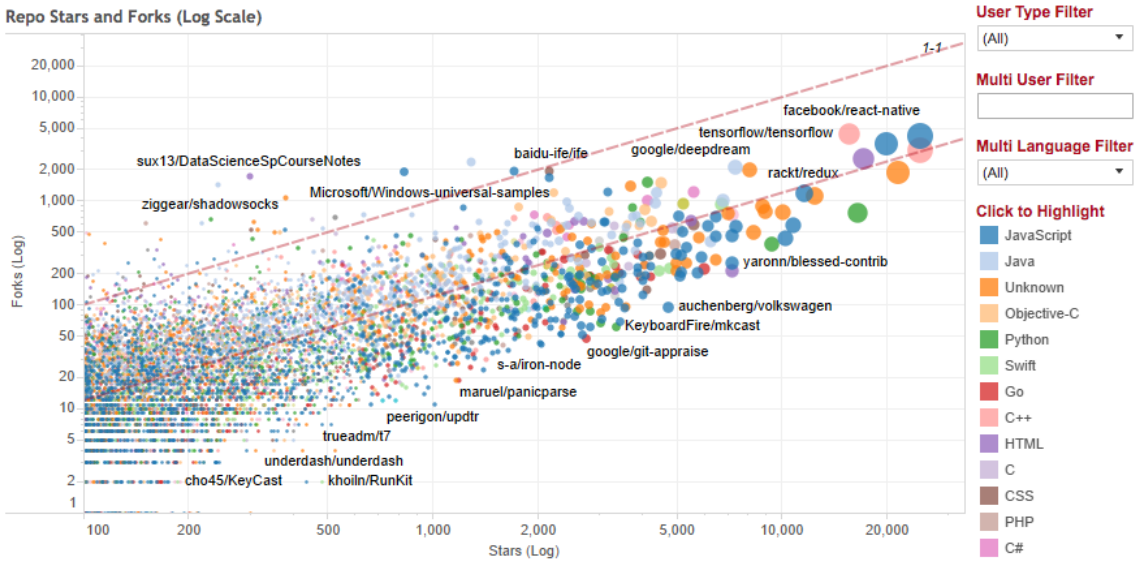


GitHub Repositories Created in 2015

Interactive Visualizations of GitHub's Newest, Most Popular Repos

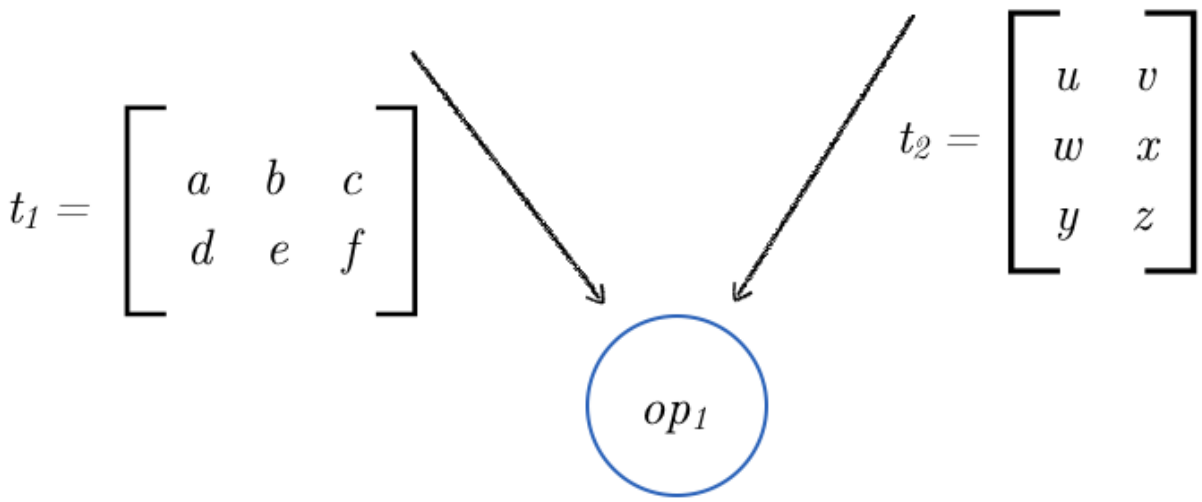
Author: https://www.github.com/donnemartin

1 2 3 4 5 6 7 8 9 A



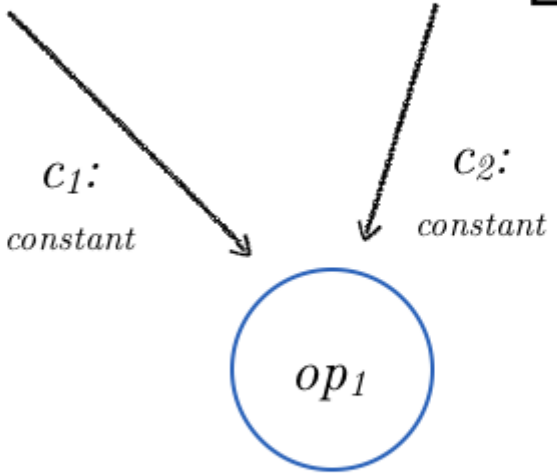
Hover: *View info* | ***Click***: *View repo url*
Interact *with the* ***filters***

Data: *Repos created in 2015, >= 100 stars* | ***Date Range***: *1/1/2015 to 1/1/2016*
FAQ: *See the final tab "A" for more info*



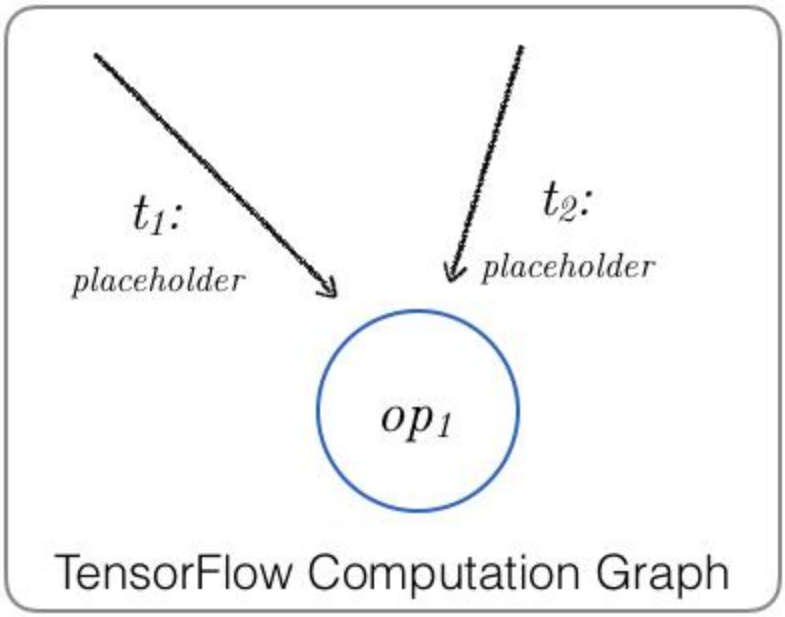
$$op_1 = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \; x \; \begin{bmatrix} u & v \\ w & x \\ y & z \end{bmatrix}$$

$$c_1 = \begin{bmatrix} 3. & 2. & 1. \end{bmatrix} \qquad c_2 = \begin{bmatrix} -1. \\ 2. \\ 1. \end{bmatrix}$$



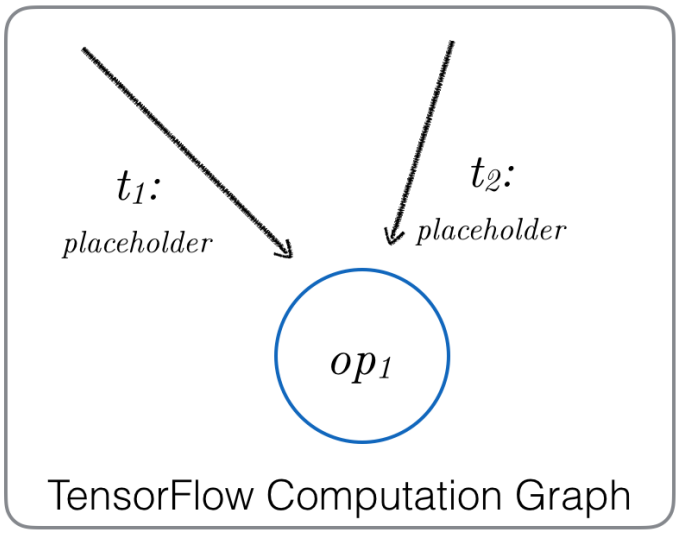
$$\begin{aligned} op_1 &= c_1 \; x \; c_2 \\ &= 2. \end{aligned}$$

$$m_1 = \begin{bmatrix} 3. & 2. & 1. \end{bmatrix} \qquad m_2 = \begin{bmatrix} -1. & 2. & 1. \end{bmatrix}^{-1}$$

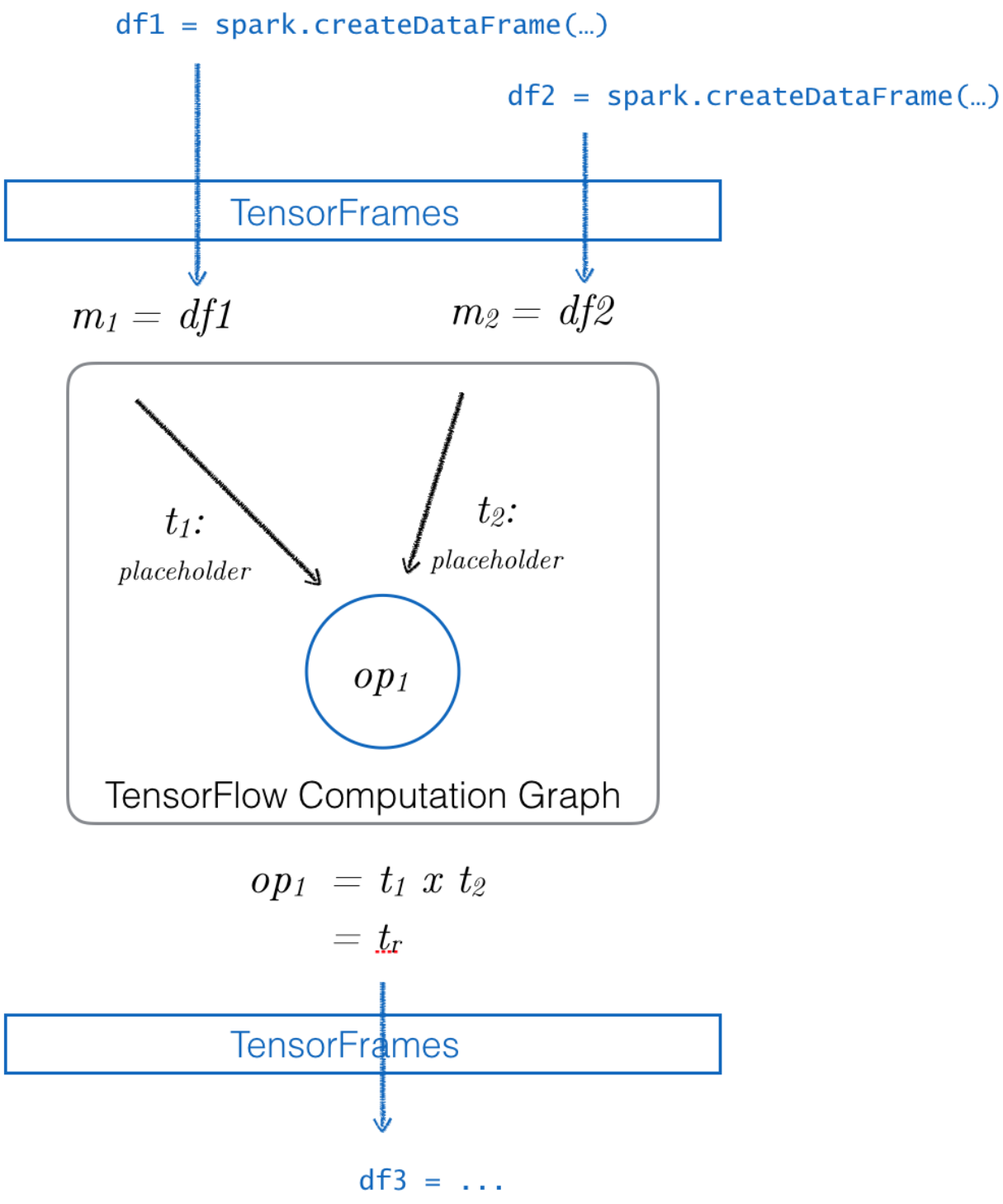


$$\begin{aligned} op_1 &= t_1 \; x \; t_2 \\ &= m_1 \; x \; m_2 \\ &= 2. \end{aligned}$$

$$m_1 = \begin{bmatrix} 3. & 2. & 1. & 0. \end{bmatrix} \qquad m_2 = \begin{bmatrix} -5. & -4. & -3. & -2. \end{bmatrix}^{-1}$$



$$\begin{aligned} op_1 &= t_1 \; x \; t_2 \\ &= m_1 \; x \; m_2 \\ &= -26. \end{aligned}$$


df1 = spark.createDataFrame(…)
df2 = spark.createDataFrame(…)
TensorFrames
$m_1 = df1$
$m_2 = df2$
t_1:
placeholder
t_2:
placeholder
op_1
TensorFlow Computation Graph
$op_1 = t_1 \; x \; t_2$
$= t_r$
TensorFrames
df3 = ...

▸ (2) Spark Jobs

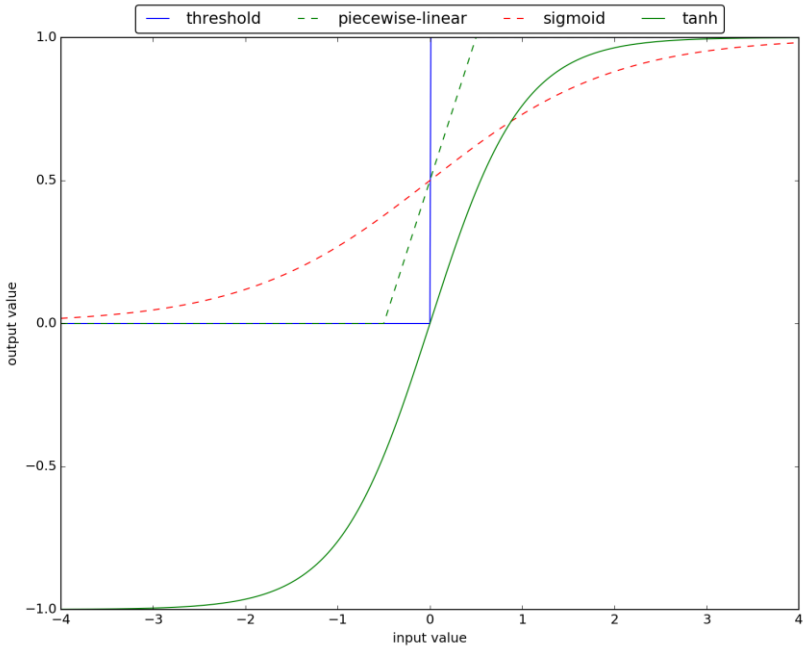
```
+---+
|  x|
+---+
|0.0|
|1.0|
|2.0|
|3.0|
|4.0|
|5.0|
|6.0|
|7.0|
|8.0|
|9.0|
+---+
```

▸ (2) Spark Jobs

```
+----+---+
|   z|  x|
+----+---+
| 3.0|0.0|
| 4.0|1.0|
| 5.0|2.0|
| 6.0|3.0|
| 7.0|4.0|
| 8.0|5.0|
| 9.0|6.0|
|10.0|7.0|
|11.0|8.0|
|12.0|9.0|
+----+---+
```

▸ (2) Spark Jobs

```
+-----------+
|          y|
+-----------+
| [0.0, 0.0]|
|[1.0, -1.0]|
|[2.0, -2.0]|
|[3.0, -3.0]|
|[4.0, -4.0]|
|[5.0, -5.0]|
|[6.0, -6.0]|
|[7.0, -7.0]|
|[8.0, -8.0]|
|[9.0, -9.0]|
+-----------+
```



$$10^{15}$$

x_i

w_i

h_1

Chapter 9: Polyglot Persistence with Blaze

```
Fetching package metadata .......
Solving package specifications: ..........

Package plan for installation in environment /Users/drabast/anaconda:

The following NEW packages will be INSTALLED:

    blaze: 0.10.1-py35_0

Proceed ([y]/n)? y

Linking packages ...
[      COMPLETE      ]|#################################################################| 100%
```

```
Out[4]: array([[1, 2, 3],
               [4, 5, 6]])
```

Out[6]:

	None
0	1
1	2
2	3

Out[7]:

	None
0	1
1	4

Out[9]:

	b
0	2
1	5

Out[13]:

	a
0	1
1	4

```
['Stop_month', 'Stop_day', 'Stop_year', 'Stop_hr', 'Stop_min', 'Sto
p_sec', 'Agency', 'SubAgency', 'Description', 'Location', 'Latitude
', 'Longitude', 'Accident', 'Belts', 'Personal_Injury', 'Property_D
amage', 'Fatal', 'Commercial_License', 'HAZMAT', 'Commercial_Vehicl
e', 'Alcohol', 'Work_Zone', 'State', 'VehicleType', 'Year', 'Make',
'Model', 'Color', 'Violation_Type', 'Charge', 'Article', 'Contribut
ed_To_Accident', 'Race', 'Gender', 'Driver_City', 'Driver_State', '
DL_State', 'Arrest_Type', 'Geolocation']
```

Out[17]:

	Stop_month	Stop_day	Stop_year	Stop_hr	Stop_min	Stop_sec	Agency
0	9	30	2014	23	51	0	MCP
1	3	31	2015	23	59	0	MCP

Out[19]:

	Stop_month	Stop_day	Stop_year	Stop_hr	Stop_min	Stop_sec	Agency
0	3	29	2013	17	34	0	MCP
1	8	12	2013	8	41	0	MCP

Out[28]:

	Year
0	2014.0
1	2003.0

Out[29]:

	Location	Year	Accident	Fatal	Alcohol
0	PARK RD AT HUNGERFORD DR	2014.0	No	No	No
1	CONNECTICUT AT METROPOLITAN AVE	2003.0	No	No	No

Out[33]:

	Stop_year	Arrest_Type	Color	Charge
73	2013	A - Marked Patrol	SILVER	13-409(b)
215	2013	B - Unmarked Patrol	BLACK	21-309(b)

```
[2013 'A - Marked Patrol' 'SILVER' '13-409(b)']
[2013 'B - Unmarked Patrol' 'BLACK' '21-309(b)']
```

Out[35]:

	Stop_year
2	2013
0	2014
1	2015
3	2016

Out[36]:

	Stop_year
0	14
1	15

Out[37]:

	Stop_year
0	7.607878
1	7.608374

Out[38]: 2016

Out[9]:

	Stop_year	Year	Age_of_car
0	2014	2014.0	0.0
1	2015	2003.0	12.0

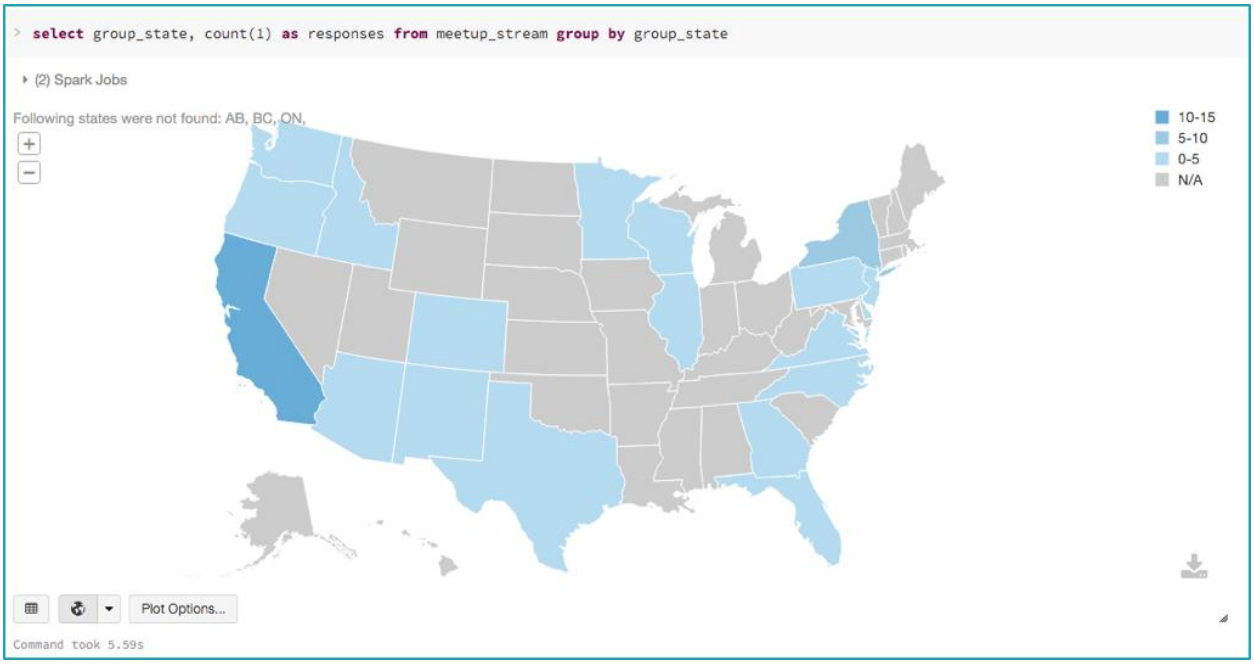
Out[40]:

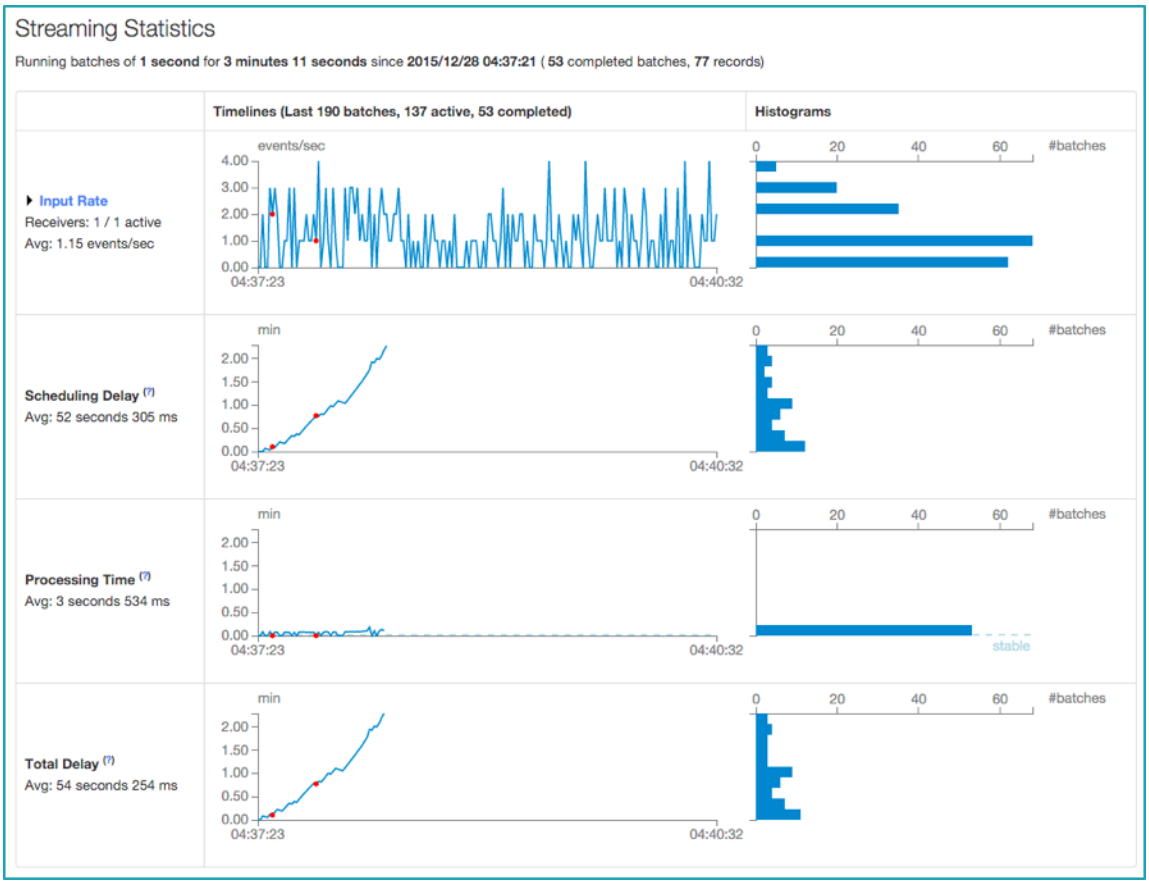
	Fatal	Fatal_AvgAge	Fatal_Count
0	No	9.580998	404418
1	Yes	8.798246	116

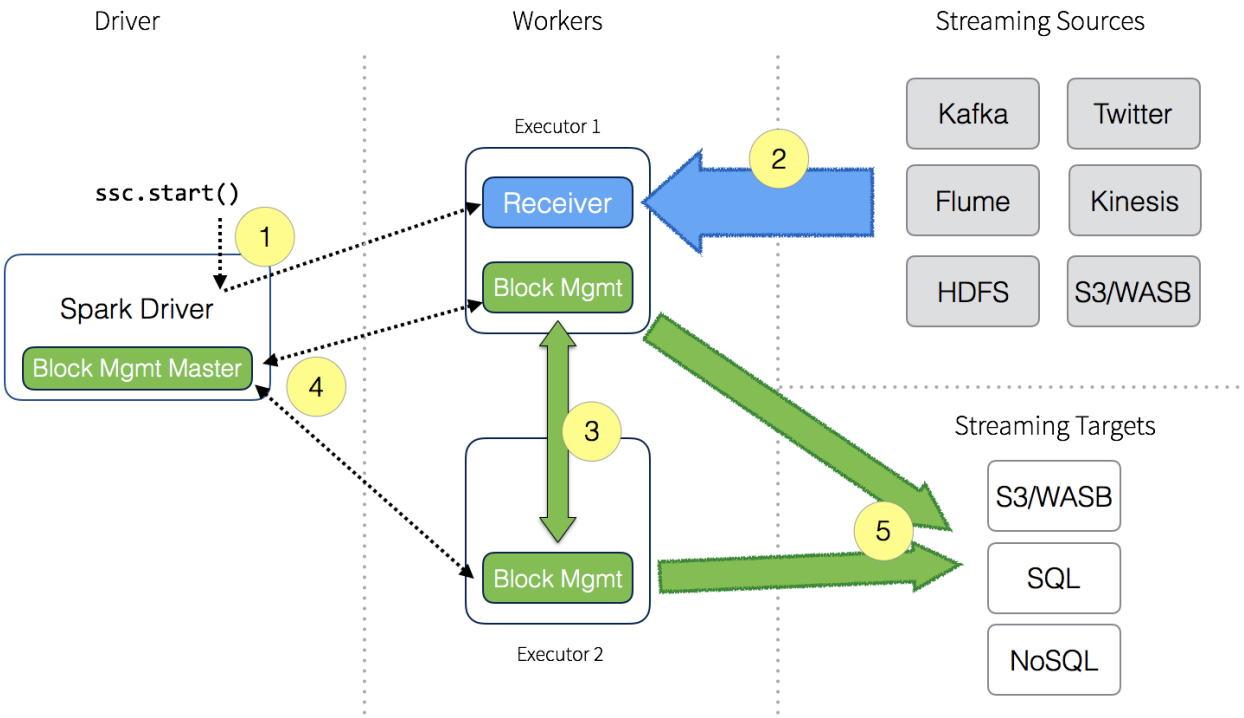
Out[43]:

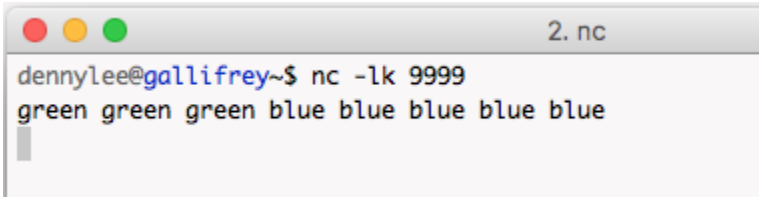
	Violation_Type	Belts	Violation_count
0	Citation	No	989728
5	Warning	No	439490
2	ESERO	No	56447
1	Citation	Yes	35596
6	Warning	Yes	12245
3	ESERO	Yes	1327
4	SERO	No	3

Chapter 10: Structures Streaming

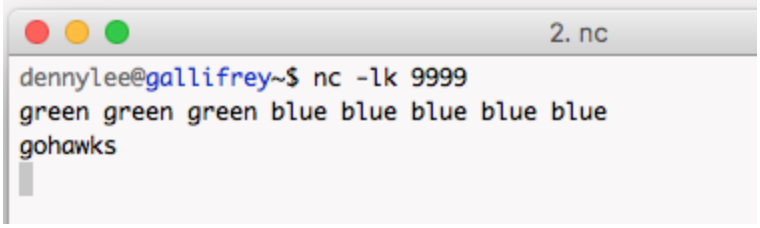



Streaming Statistics
Running batches of 1 second for 3 minutes 11 seconds since 2015/12/28 04:37:21 (53 completed batches, 77 records)
Timelines (Last 190 batches, 137 active, 53 completed)
Histograms
Input Rate
Receivers: 1 / 1 active
Avg: 1.15 events/sec
events/sec
4.00
3.00
2.00
1.00
0.00
04:37:23
04:40:32
0
20
40
60
#batches
Scheduling Delay (?)
Avg: 52 seconds 305 ms
min
2.00
1.50
1.00
0.50
0.00
Processing Time (?)
Avg: 3 seconds 534 ms
stable
Total Delay (?)
Avg: 54 seconds 254 ms





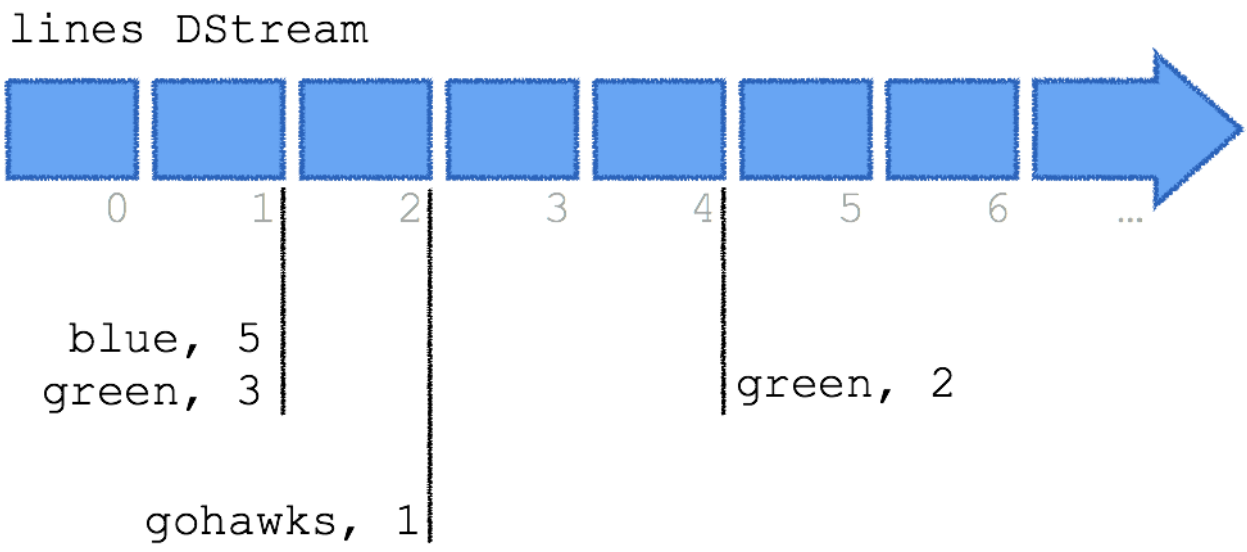

```
-------------------------------------------
Time: 2017-01-14 13:30:31
-------------------------------------------

-------------------------------------------
Time: 2017-01-14 13:30:32
-------------------------------------------
(u'blue', 5)
(u'green', 3)

-------------------------------------------
Time: 2017-01-14 13:30:33
-------------------------------------------
```

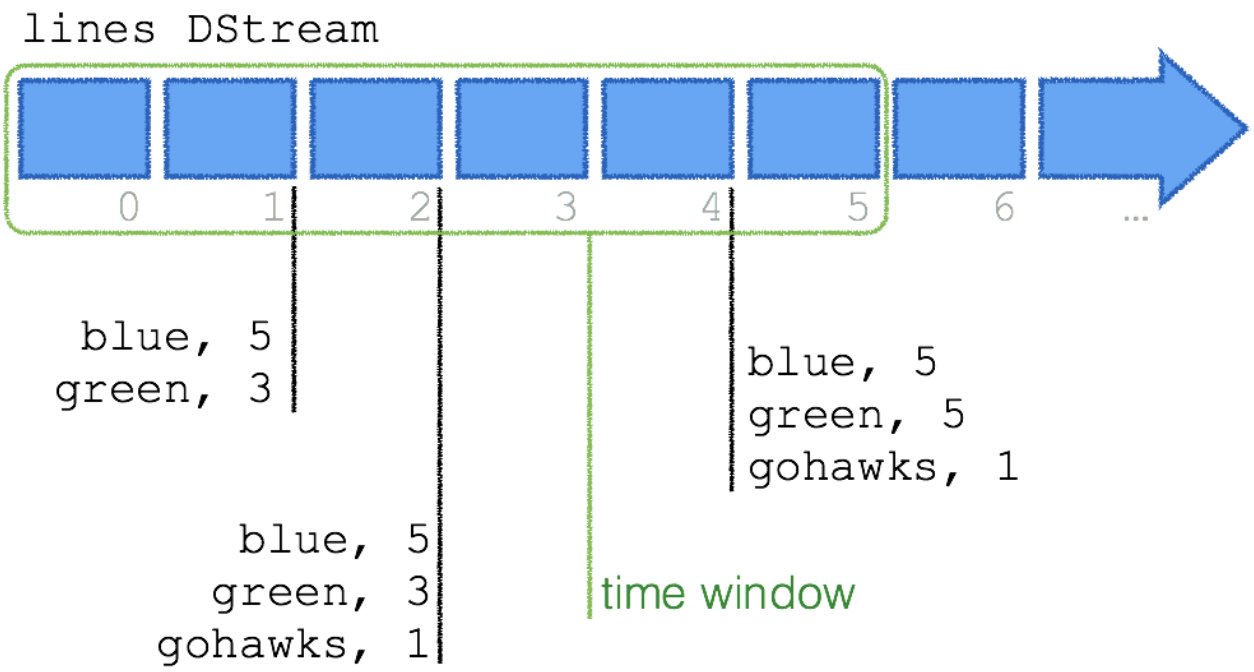


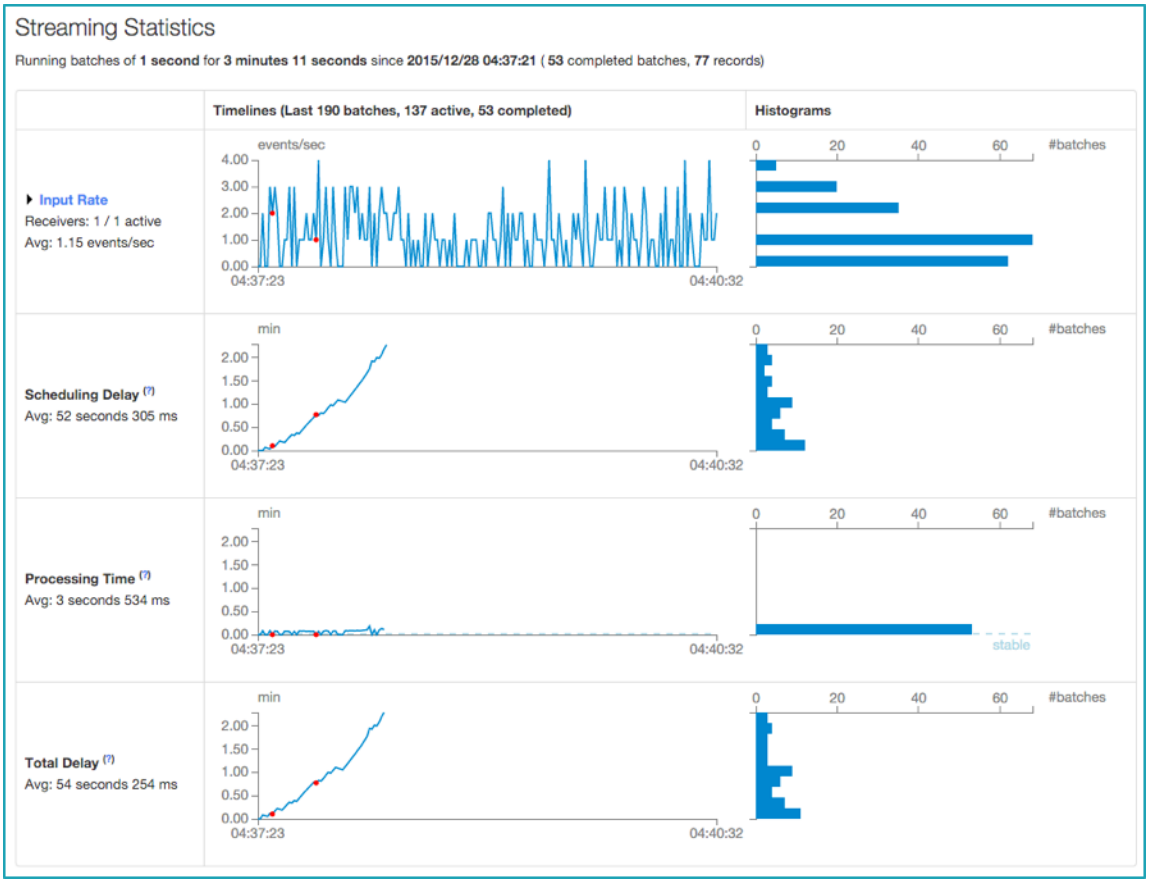

```
-------------------------------------------
Time: 2017-01-14 13:30:31
-------------------------------------------

-------------------------------------------
Time: 2017-01-14 13:30:32
-------------------------------------------
(u'blue', 5)
(u'green', 3)

-------------------------------------------
Time: 2017-01-14 13:30:33
-------------------------------------------

-------------------------------------------
Time: 2017-01-14 13:30:34
-------------------------------------------

-------------------------------------------
Time: 2017-01-14 13:30:35
-------------------------------------------
(u'gohawks', 1)

-------------------------------------------
Time: 2017-01-14 13:30:36
-------------------------------------------

-------------------------------------------
Time: 2017-01-14 13:30:37
-------------------------------------------
```

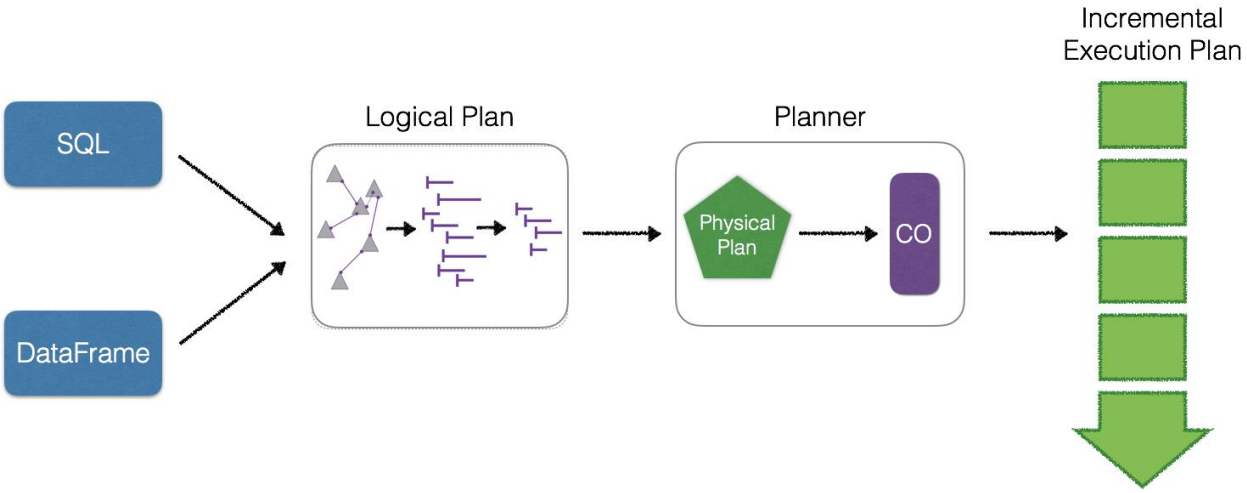
```
dennylee@gallifrey~$ nc -lk 9999
green green blue blue blue blue blue
gohawks
green green
```

```
-------------------------------------------
Time: 2017-01-16 17:19:38
-------------------------------------------

-------------------------------------------
Time: 2017-01-16 17:19:39
-------------------------------------------
(u'blue', 5)
(u'green', 2)

-------------------------------------------
Time: 2017-01-16 17:19:40
-------------------------------------------
(u'gohawks', 1)

-------------------------------------------
Time: 2017-01-16 17:19:41
-------------------------------------------

-------------------------------------------
Time: 2017-01-16 17:19:42
-------------------------------------------

-------------------------------------------
Time: 2017-01-16 17:19:43
-------------------------------------------
(u'green', 2)
```



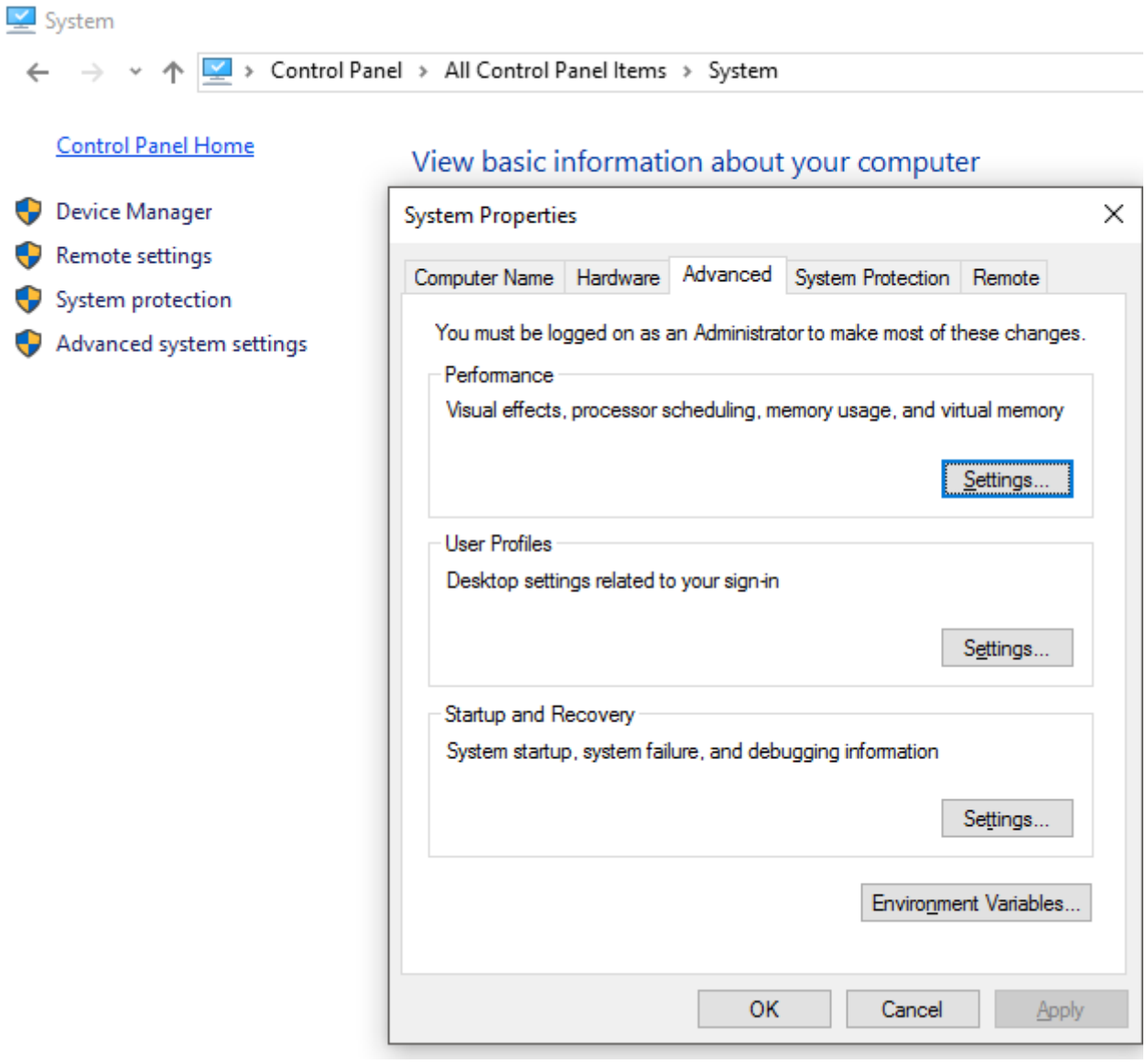

lines DStream
0
1
2
3
4
5
6
...
blue, 5
green, 3
blue, 5
green, 3
gohawks, 1
blue, 5
green, 5
gohawks, 1
time window

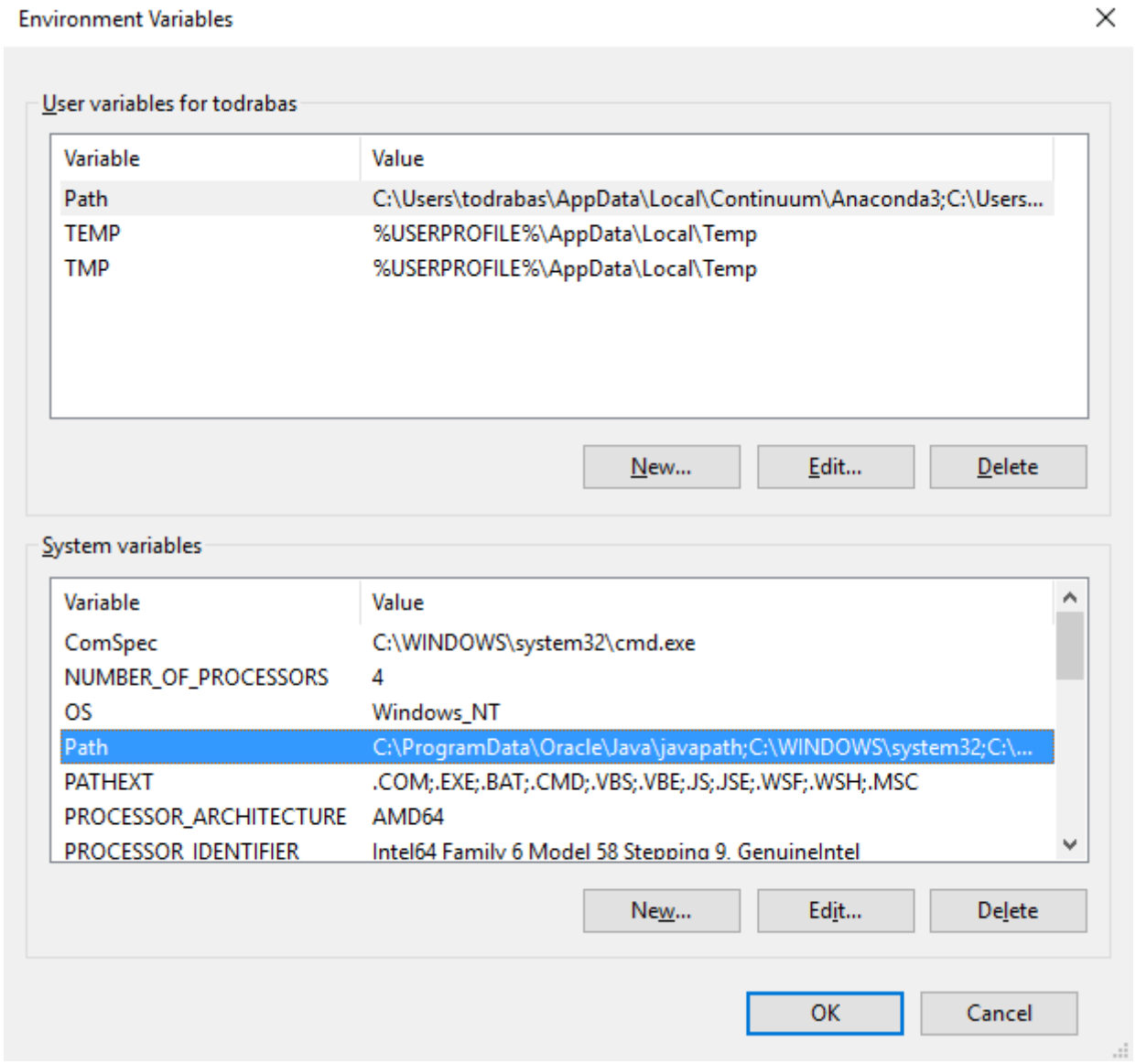

Streaming Statistics
Running batches of 1 second for 3 minutes 11 seconds since 2015/12/28 04:37:21 (53 completed batches, 77 records)
Timelines (Last 190 batches, 137 active, 53 completed)
Histograms
Input Rate
Receivers: 1 / 1 active
Avg: 1.15 events/sec
events/sec
#batches
Scheduling Delay
Avg: 52 seconds 305 ms
Processing Time
Avg: 3 seconds 534 ms
stable
Total Delay
Avg: 54 seconds 254 ms
min
04:37:23
04:40:32

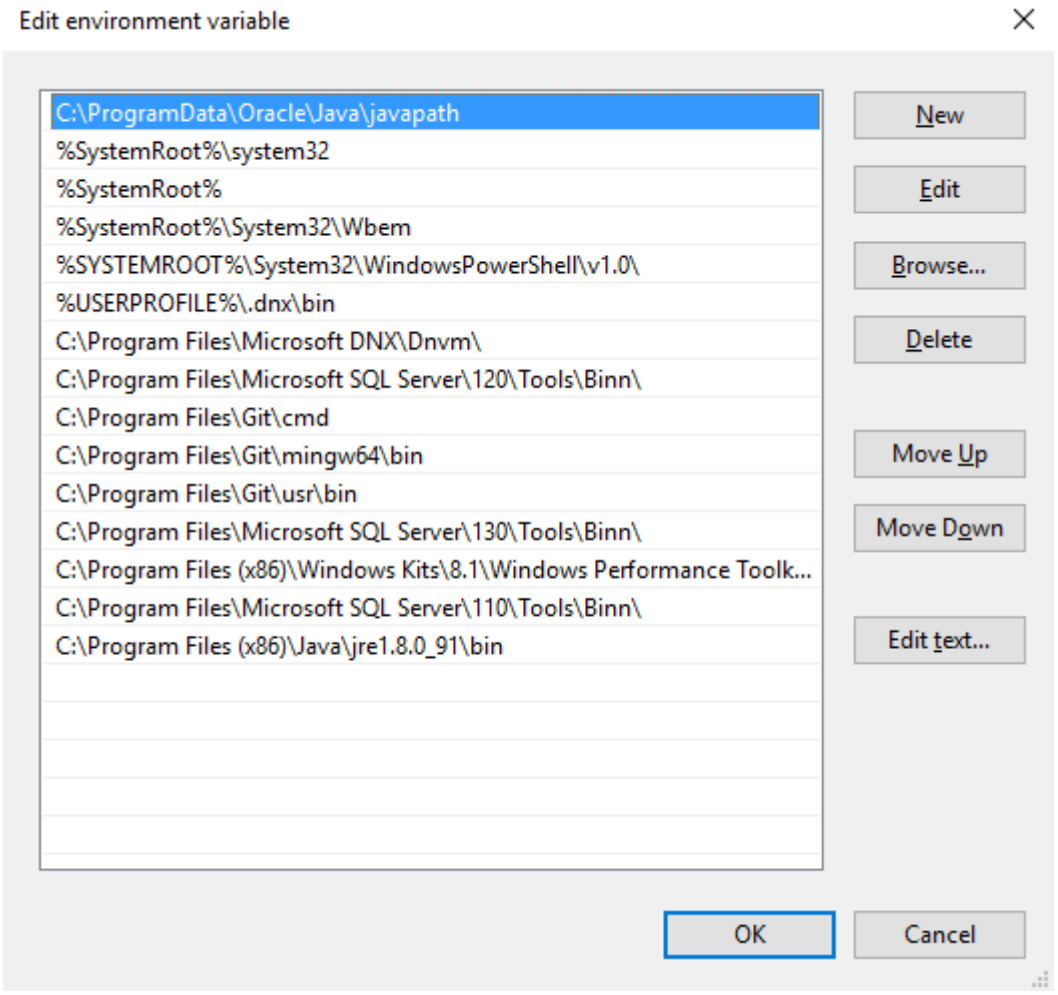



```
-------------------------------------------
Batch: 0
-------------------------------------------
+-----+-----+
| word|count|
+-----+-----+
|green|    3|
| blue|    5|
+-----+-----+

-------------------------------------------
Batch: 1
-------------------------------------------
+-------+-----+
|   word|count|
+-------+-----+
|  green|    3|
|   blue|    5|
|gohawks|    1|
+-------+-----+

-------------------------------------------
Batch: 2
-------------------------------------------
+-------+-----+
|   word|count|
+-------+-----+
|  green|    5|
|   blue|    5|
|gohawks|    1|
+-------+-----+
```

Bonus Chapter 1: Installing Spark




Environment Variables
User variables for todrabas
Variable
Value
Path
C:\Users\todrabas\AppData\Local\Continuum\Anaconda3;C:\Users...
TEMP
%USERPROFILE%\AppData\Local\Temp
TMP
%USERPROFILE%\AppData\Local\Temp
New...
Edit...
Delete
System variables
Variable
Value
ComSpec
C:\WINDOWS\system32\cmd.exe
NUMBER_OF_PROCESSORS
4
OS
Windows_NT
Path
C:\ProgramData\Oracle\Java\javapath;C:\WINDOWS\system32;C:\...
PATHEXT
.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE
AMD64
PROCESSOR_IDENTIFIER
Intel64 Family 6 Model 58 Stepping 9, GenuineIntel
New...
Edit...
Delete
OK
Cancel


Edit environment variable
C:\ProgramData\Oracle\Java\javapath
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%USERPROFILE%\.dnx\bin
C:\Program Files\Microsoft DNX\Dnvm\
C:\Program Files\Microsoft SQL Server\120\Tools\Binn\
C:\Program Files\Git\cmd
C:\Program Files\Git\mingw64\bin
C:\Program Files\Git\usr\bin
C:\Program Files\Microsoft SQL Server\130\Tools\Binn\
C:\Program Files (x86)\Windows Kits\8.1\Windows Performance Toolk...
C:\Program Files\Microsoft SQL Server\110\Tools\Binn\
C:\Program Files (x86)\Java\jre1.8.0_91\bin
New
Edit
Browse...
Delete
Move Up
Move Down
Edit text...
OK
Cancel

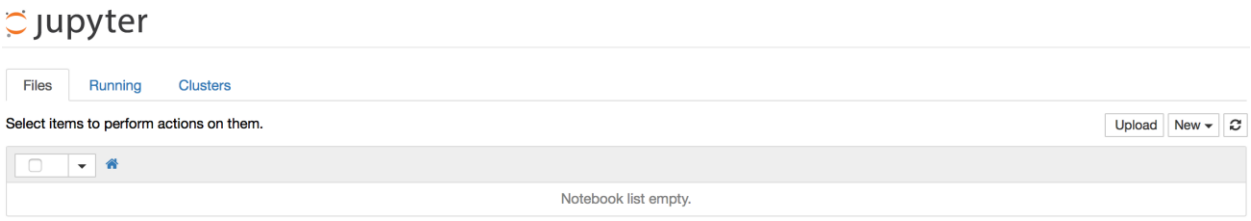
```
[INFO] ------------------------------------------------------------------------
[INFO] Reactor Summary:
[INFO]
[INFO] Spark Project Parent POM ........................... SUCCESS [  2.612 s]
[INFO] Spark Project Tags ................................. SUCCESS [  5.155 s]
[INFO] Spark Project Sketch ............................... SUCCESS [  6.345 s]
[INFO] Spark Project Networking ........................... SUCCESS [  8.141 s]
[INFO] Spark Project Shuffle Streaming Service ............ SUCCESS [  4.775 s]
[INFO] Spark Project Unsafe ............................... SUCCESS [  6.784 s]
[INFO] Spark Project Launcher ............................. SUCCESS [  7.271 s]
[INFO] Spark Project Core ................................. SUCCESS [01:50 min]
[INFO] Spark Project ML Local Library ..................... SUCCESS [  6.066 s]
[INFO] Spark Project GraphX ............................... SUCCESS [ 11.841 s]
[INFO] Spark Project Streaming ............................ SUCCESS [ 24.800 s]
[INFO] Spark Project Catalyst ............................. SUCCESS [ 59.887 s]
[INFO] Spark Project SQL .................................. SUCCESS [01:21 min]
[INFO] Spark Project ML Library ........................... SUCCESS [01:02 min]
[INFO] Spark Project Tools ................................ SUCCESS [  0.886 s]
[INFO] Spark Project Hive ................................. SUCCESS [ 38.901 s]
[INFO] Spark Project REPL ................................. SUCCESS [  3.463 s]
[INFO] Spark Project YARN Shuffle Service ................. SUCCESS [  5.193 s]
[INFO] Spark Project YARN ................................. SUCCESS [  8.081 s]
[INFO] Spark Project Hive Thrift Server ................... SUCCESS [ 16.256 s]
[INFO] Spark Project Assembly ............................. SUCCESS [  2.667 s]
[INFO] Spark Project External Flume Sink .................. SUCCESS [  4.421 s]
[INFO] Spark Project External Flume ....................... SUCCESS [  9.387 s]
[INFO] Spark Project External Flume Assembly .............. SUCCESS [  2.294 s]
[INFO] Spark Integration for Kafka 0.8 .................... SUCCESS [  8.363 s]
[INFO] Spark Project Examples ............................. SUCCESS [ 14.318 s]
[INFO] Spark Project External Kafka Assembly .............. SUCCESS [  3.098 s]
[INFO] Spark Integration for Kafka 0.10 ................... SUCCESS [  6.825 s]
[INFO] Spark Integration for Kafka 0.10 Assembly .......... SUCCESS [  2.987 s]
[INFO] Kafka 0.10 Source for Structured Streaming ......... SUCCESS [  7.260 s]
[INFO] Spark Project Java 8 Tests ......................... SUCCESS [  3.987 s]
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time: 08:57 min
[INFO] Finished at: 2017-01-15T16:29:36-08:00
[INFO] Final Memory: 92M/952M
[INFO] ------------------------------------------------------------------------
```

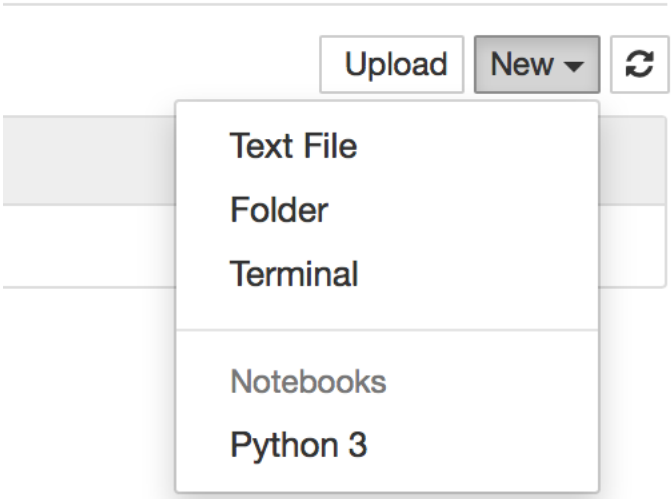
```
[info] Packaging /Users/drabast/Downloads/spark-2.1.0/examples/target/scala-2.11/jars/spark-examples_2.11-2.1.0.jar ...
[info] Done packaging.
[success] Total time: 238 s, completed Jan 16, 2017 8:40:00 PM
```

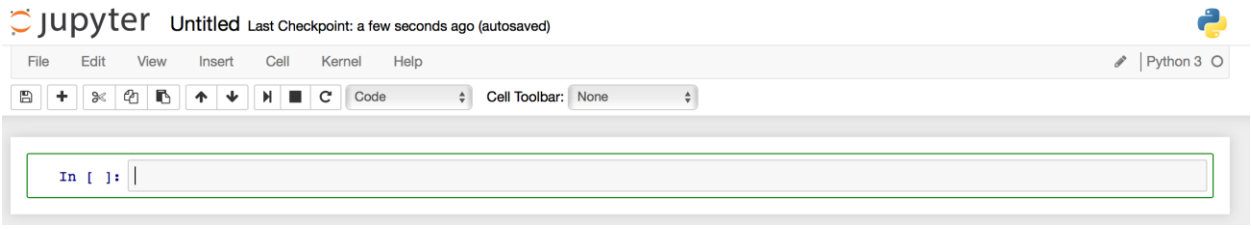
```
Running PySpark tests. Output is in /Users/drabast/Downloads/spark-2.1.0/python/unit-tests.log
Will test against the following Python executables: ['python']
Will test the following Python modules: ['pyspark-core', 'pyspark-ml', 'pyspark-mllib', 'pyspark-sql', 'pyspark-streaming']
Finished test(python): pyspark.sql.tests (63s)
Finished test(python): pyspark.accumulators (8s)
Finished test(python): pyspark.broadcast (5s)
Finished test(python): pyspark.conf (4s)
Finished test(python): pyspark.context (19s)
Finished test(python): pyspark.ml.classification (30s)
Finished test(python): pyspark.tests (140s)
Finished test(python): pyspark.ml.clustering (23s)
Finished test(python): pyspark.ml.evaluation (14s)
Finished test(python): pyspark.ml.linalg.__init__ (0s)
Finished test(python): pyspark.ml.recommendation (18s)
Finished test(python): pyspark.ml.feature (31s)
Finished test(python): pyspark.streaming.tests (187s)
Finished test(python): pyspark.ml.regression (25s)
Finished test(python): pyspark.ml.tuning (23s)
Finished test(python): pyspark.mllib.tests (214s)
Finished test(python): pyspark.mllib.classification (26s)
Finished test(python): pyspark.mllib.evaluation (20s)
Finished test(python): pyspark.mllib.feature (26s)
Finished test(python): pyspark.mllib.clustering (42s)
Finished test(python): pyspark.mllib.linalg.__init__ (0s)
Finished test(python): pyspark.mllib.fpm (21s)
Finished test(python): pyspark.mllib.random (10s)
Finished test(python): pyspark.ml.tests (89s)
Finished test(python): pyspark.mllib.stat.KernelDensity (0s)
Finished test(python): pyspark.mllib.recommendation (27s)
Finished test(python): pyspark.mllib.linalg.distributed (31s)
Finished test(python): pyspark.mllib.regression (27s)
Finished test(python): pyspark.mllib.stat._statistics (14s)
Finished test(python): pyspark.mllib.util (11s)
Finished test(python): pyspark.profiler (9s)
Finished test(python): pyspark.mllib.tree (17s)
Finished test(python): pyspark.shuffle (1s)
Finished test(python): pyspark.serializers (15s)
Finished test(python): pyspark.rdd (21s)
Finished test(python): pyspark.sql.conf (5s)
Finished test(python): pyspark.sql.catalog (18s)
Finished test(python): pyspark.sql.column (19s)
Finished test(python): pyspark.sql.context (21s)
Finished test(python): pyspark.sql.group (34s)
Finished test(python): pyspark.sql.dataframe (39s)
Finished test(python): pyspark.sql.functions (41s)
Finished test(python): pyspark.sql.types (9s)
Finished test(python): pyspark.sql.window (5s)
Finished test(python): pyspark.streaming.util (0s)
Finished test(python): pyspark.sql.readwriter (33s)
Finished test(python): pyspark.sql.session (16s)
Tests passed in 372 seconds
```

```
Python 3.5.1 |Anaconda 2.4.1 (x86_64)| (default, Dec  7 2015, 11:24:55)
[GCC 4.2.1 (Apple Inc. build 5577)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.1.0
      /_/

Using Python version 3.5.1 (default, Dec  7 2015 11:24:55)
SparkSession available as 'spark'.
>>>
```

```
Command Prompt - pyspark

C:\Users\todrabas>pyspark
Python 3.5.1 |Anaconda 2.4.1 (64-bit)| (default, Dec  7 2015, 15:00:12) [MSC v.1900 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel).
16/07/18 20:57:04 ERROR Shell: Failed to locate the winutils binary in the hadoop binary path
java.io.IOException: Could not locate executable null\bin\winutils.exe in the Hadoop binaries.
        at org.apache.hadoop.util.Shell.getQualifiedBinPath(Shell.java:356)
        at org.apache.hadoop.util.Shell.getWinUtilsPath(Shell.java:371)
        at org.apache.hadoop.util.Shell.<clinit>(Shell.java:364)
        at org.apache.hadoop.util.StringUtils.<clinit>(StringUtils.java:80)
        at org.apache.hadoop.security.SecurityUtil.getAuthenticationMethod(SecurityUtil.java:611)
        at org.apache.hadoop.security.UserGroupInformation.initialize(UserGroupInformation.java:272)
        at org.apache.hadoop.security.UserGroupInformation.ensureInitialized(UserGroupInformation.java:260)
        at org.apache.hadoop.security.UserGroupInformation.loginUserFromSubject(UserGroupInformation.java:790)
        at org.apache.hadoop.security.UserGroupInformation.getLoginUser(UserGroupInformation.java:760)
        at org.apache.hadoop.security.UserGroupInformation.getCurrentUser(UserGroupInformation.java:633)
        at org.apache.spark.util.Utils$$anonfun$getCurrentUserName$1.apply(Utils.scala:2181)
        at org.apache.spark.util.Utils$$anonfun$getCurrentUserName$1.apply(Utils.scala:2181)
        at scala.Option.getOrElse(Option.scala:121)
        at org.apache.spark.util.Utils$.getCurrentUserName(Utils.scala:2181)
        at org.apache.spark.SparkContext.<init>(SparkContext.scala:299)
        at org.apache.spark.api.java.JavaSparkContext.<init>(JavaSparkContext.scala:58)
        at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
        at sun.reflect.NativeConstructorAccessorImpl.newInstance(Unknown Source)
        at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(Unknown Source)
        at java.lang.reflect.Constructor.newInstance(Unknown Source)
        at py4j.reflection.MethodInvoker.invoke(MethodInvoker.java:240)
        at py4j.reflection.ReflectionEngine.invoke(ReflectionEngine.java:357)
        at py4j.Gateway.invoke(Gateway.java:236)
        at py4j.commands.ConstructorCommand.invokeConstructor(ConstructorCommand.java:80)
        at py4j.commands.ConstructorCommand.execute(ConstructorCommand.java:69)
        at py4j.GatewayConnection.run(GatewayConnection.java:211)
        at java.lang.Thread.run(Unknown Source)
16/07/18 20:57:04 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/07/18 20:57:05 WARN AbstractHandler: No Server set for org.spark_project.jetty.server.handler.ErrorHandler@17f9462
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.1.0
      /_/

Using Python version 3.5.1 (default, Dec  7 2015 15:00:12)
SparkSession available as 'spark'.
>>>
```

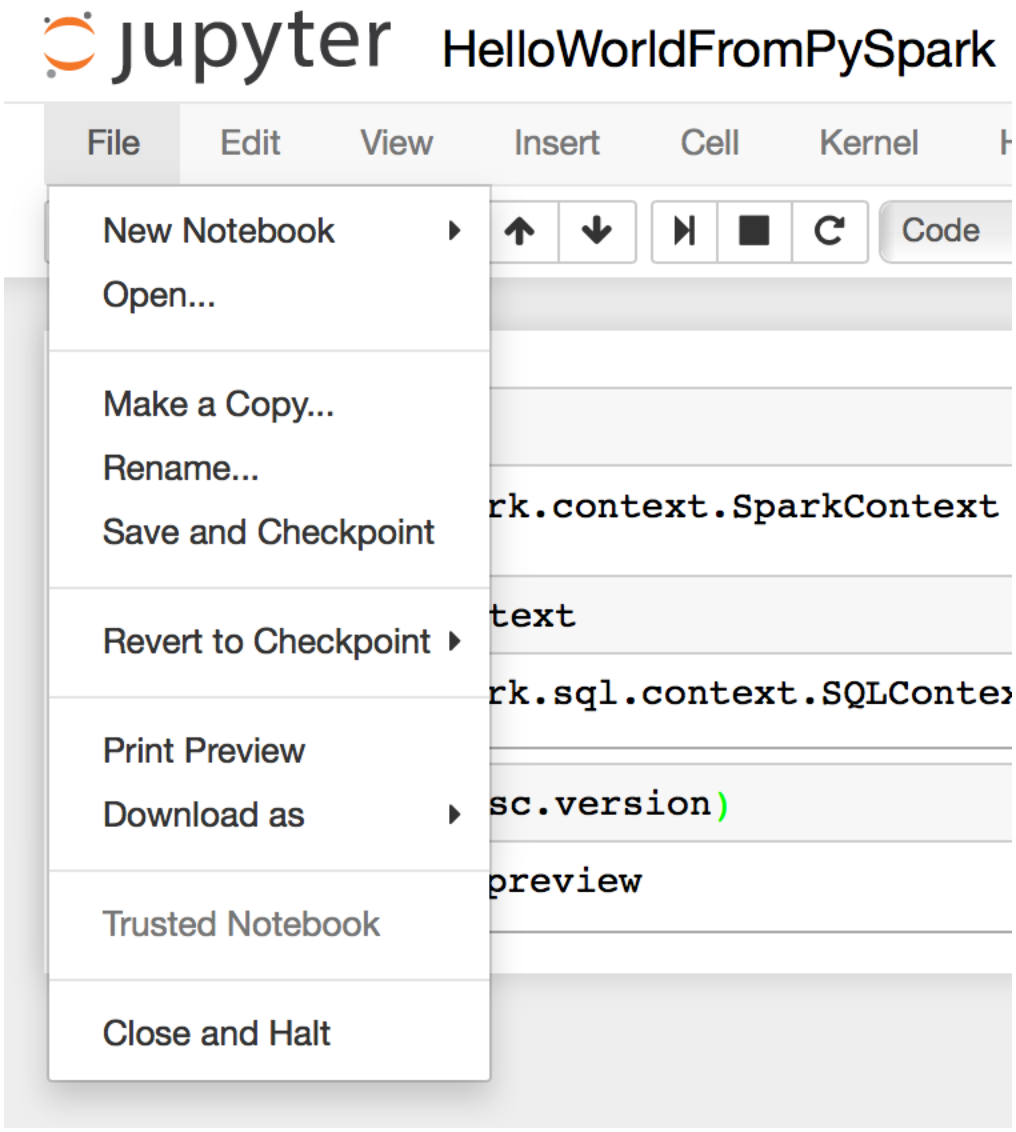




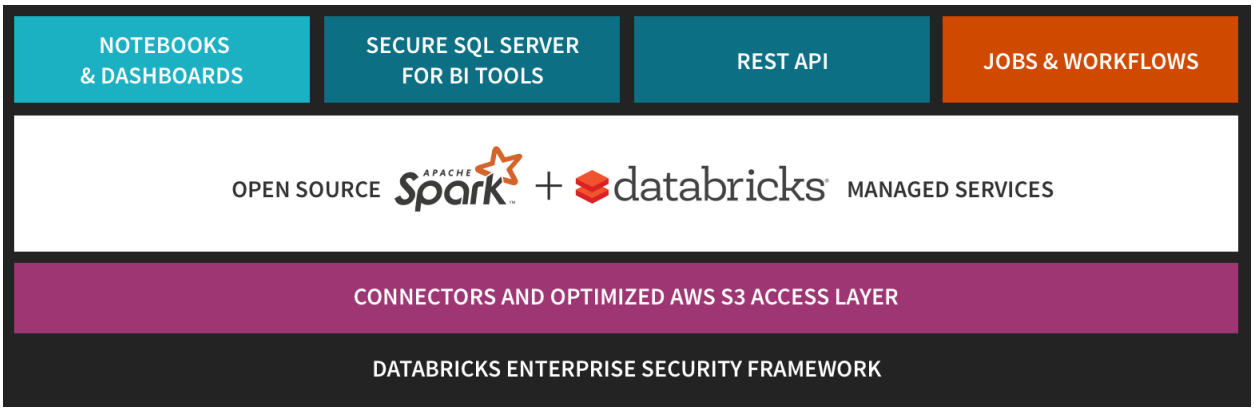



```
Out[1]: <pyspark.context.SparkContext at 0x1050456a0>
```

```
Out[2]: <pyspark.sql.context.SQLContext at 0x10b832a58>
```


jupyter
HelloWorldFromPySpark
File
Edit
View
Insert
Cell
Kernel
Code
New Notebook
Open...
Make a Copy...
Rename...
Save and Checkpoint
Revert to Checkpoint
Print Preview
Download as
Trusted Notebook
Close and Halt
rk.context.SparkContext
text
rk.sql.context.SQLContex
sc.version)
preview

Bonus Chapter 2: Free Spark Cloud Offering



Select a version to get started.

FULL-PLATFORM TRIAL

Put Apache Spark to work

- Unlimited clusters
- Notebooks, dashboards, production jobs, RESTful APIs
- Interactive guide to Spark and Databricks
- Deployed to your AWS VPC
- BI tools integration
- 14-day free trial (excludes AWS charges)

START TODAY

COMMUNITY EDITION

Learn Apache Spark

- Mini 6GB cluster
- Interactive notebooks and dashboards
- Public environment to share your work

START TODAY

databricks

Sign Up for Databricks Community Edition

First Name *
Doctor

Last Name *
Who

Company Name *
Tardis

Work Email *
put-your-email-here

Password *
••••••••••••••••••••

Confirm Password *
••••••••••••••••••••

Phone Number
425-555-1212

What is your intended use case? *
Personal - Learning Spark

How would you describe your role? *
Data Scientist

I'm not a robot
reCAPTCHA
Privacy - Terms

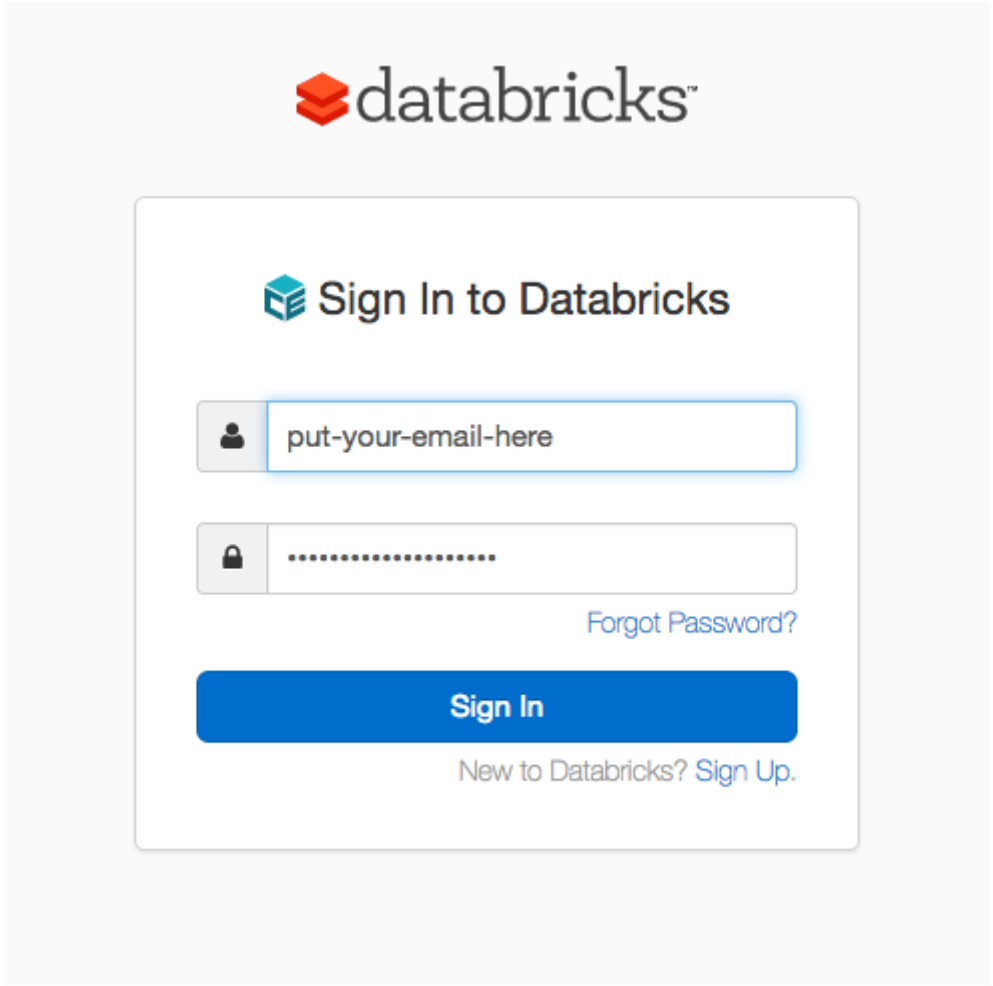
Sign Up

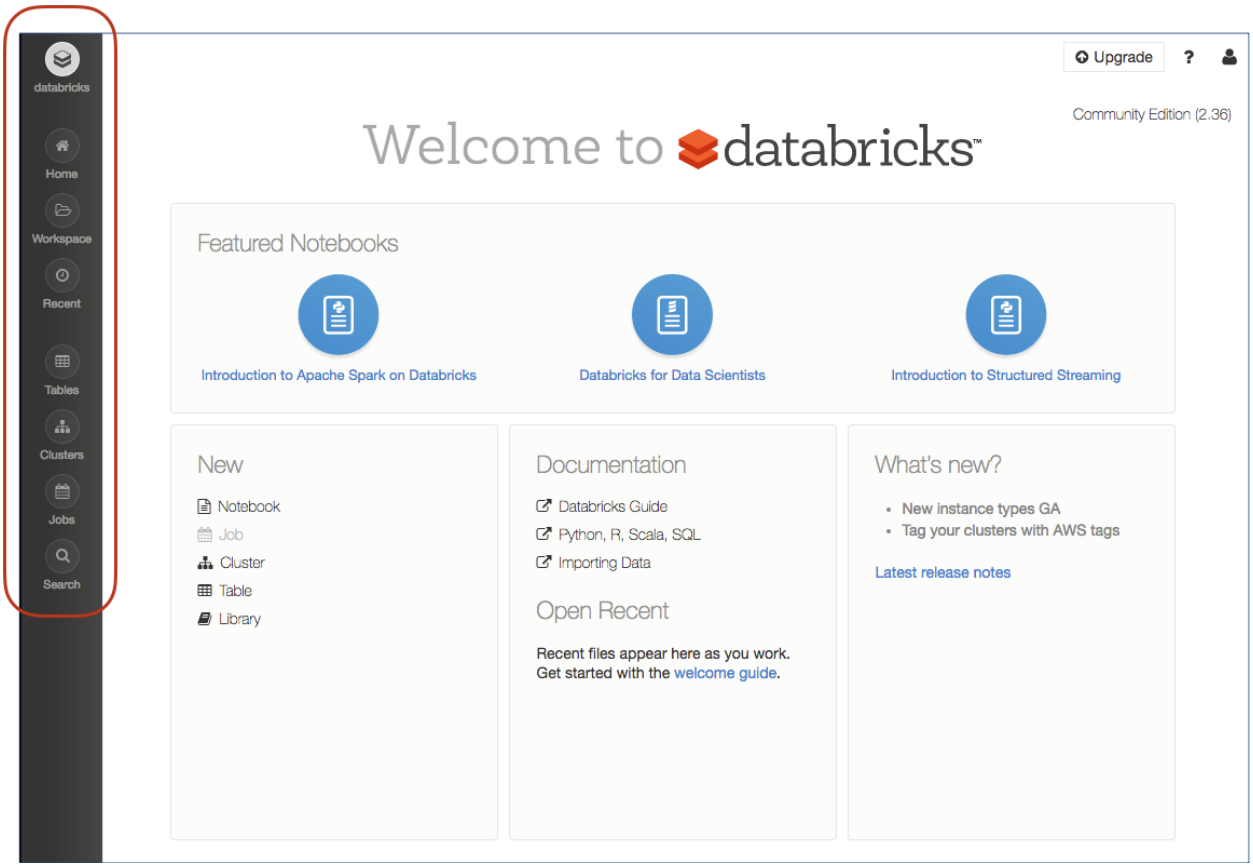
databricks

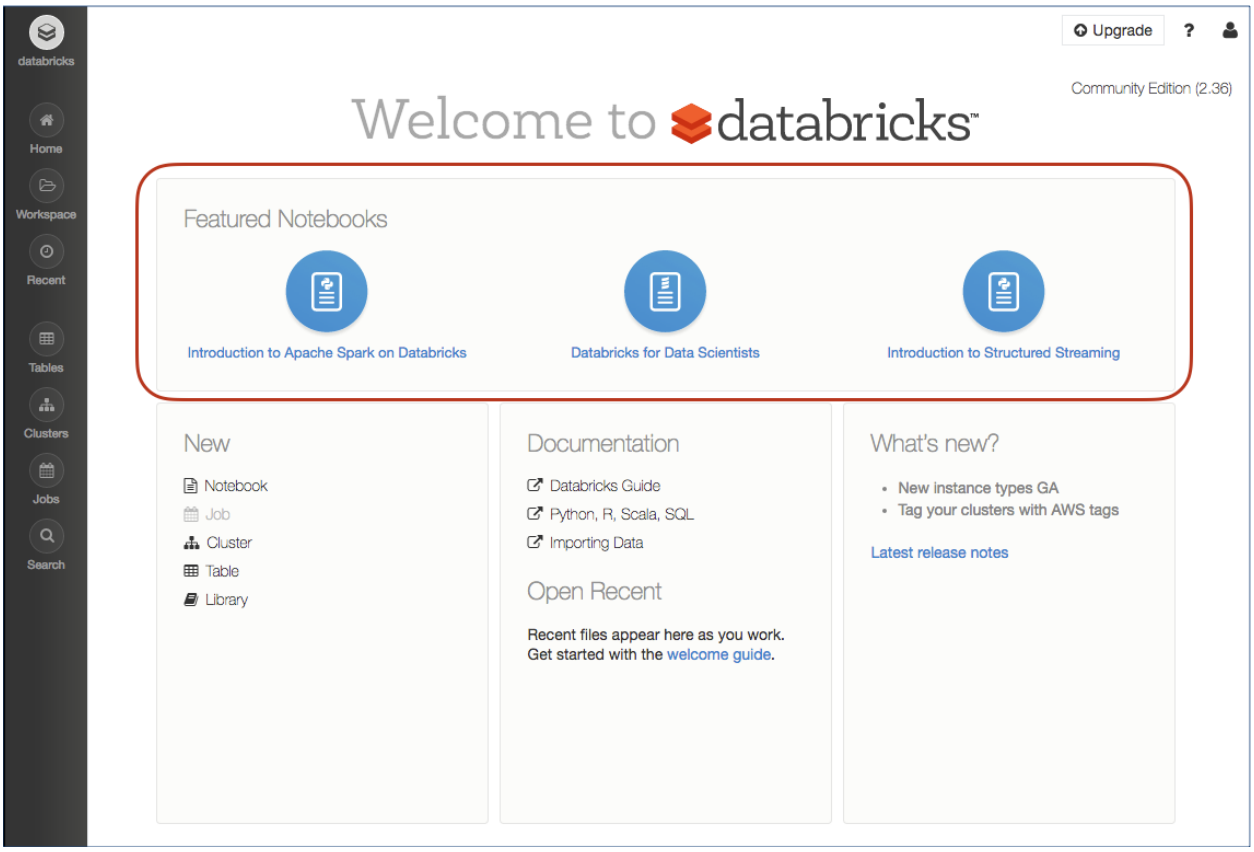
Please Confirm Your Email Address

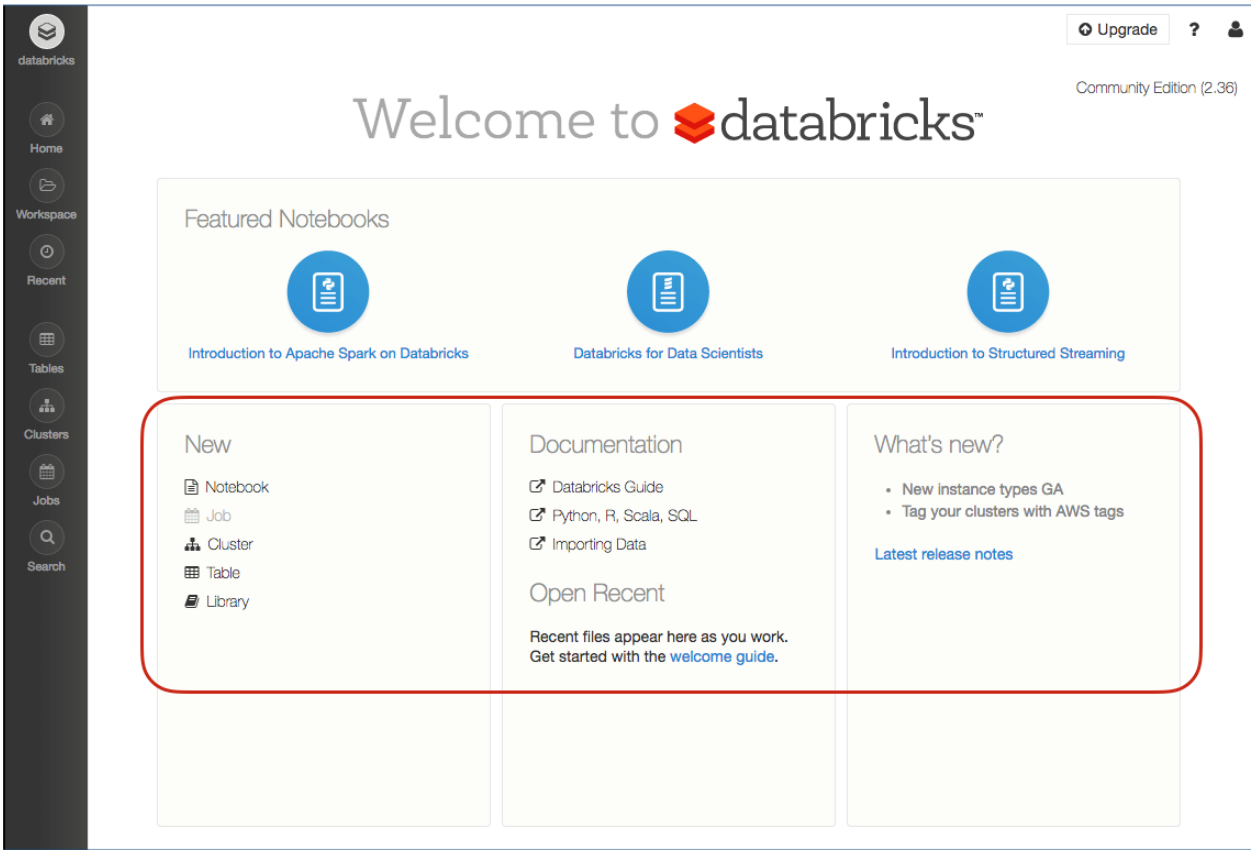
You will receive an email with a link to confirm your email address. Please click the link to complete the signup process. **If you haven't received the email, please check your spam folder.**

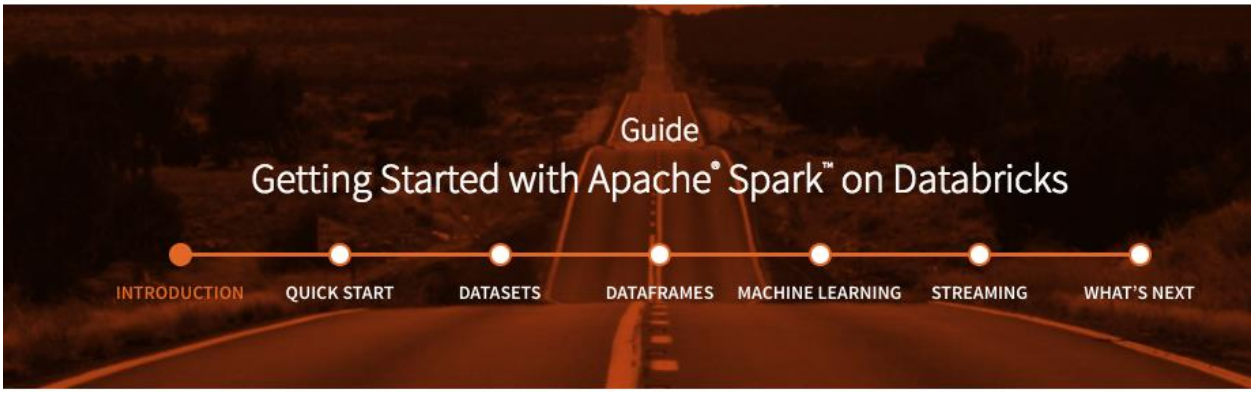
Contact feedback@databricks.com if you have any questions.


databricks
Sign In to Databricks
put-your-email-here
Forgot Password?
Sign In
New to Databricks? Sign Up.


databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
Upgrade
Community Edition (2.36)
Welcome to databricks
Featured Notebooks
Introduction to Apache Spark on Databricks
Databricks for Data Scientists
Introduction to Structured Streaming
New
Notebook
Job
Cluster
Table
Library
Documentation
Databricks Guide
Python, R, Scala, SQL
Importing Data
Open Recent
Recent files appear here as you work.
Get started with the welcome guide.
What's new?
New instance types GA
Tag your clusters with AWS tags
Latest release notes


databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
Upgrade
Community Edition (2.36)
Welcome to databricks
Featured Notebooks
Introduction to Apache Spark on Databricks
Databricks for Data Scientists
Introduction to Structured Streaming
New
Notebook
Job
Cluster
Table
Library
Documentation
Databricks Guide
Python, R, Scala, SQL
Importing Data
Open Recent
Recent files appear here as you work.
Get started with the welcome guide.
What's new?
New instance types GA
Tag your clusters with AWS tags
Latest release notes


databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
Upgrade
Community Edition (2.36)
Welcome to databricks
Featured Notebooks
Introduction to Apache Spark on Databricks
Databricks for Data Scientists
Introduction to Structured Streaming
New
Notebook
Job
Cluster
Table
Library
Documentation
Databricks Guide
Python, R, Scala, SQL
Importing Data
Open Recent
Recent files appear here as you work.
Get started with the welcome guide.
What's new?
New instance types GA
Tag your clusters with AWS tags
Latest release notes



Welcome

This self-paced guide is the "Hello World" tutorial of Apache Spark using Databricks (try Databricks here). In the following chapters, you will familiarize yourself with the Spark UI, learn how to create Spark jobs, load data and work with Datasets, get familiar with Spark's DataFrames API, run machine learning algorithms, and understand the basic concepts behind Spark Streaming. Instead of worrying about spinning up clusters, maintaining clusters, maintaining code history, or Spark versions, you can start writing Spark queries instantly and focus on your data problems.

Microsoft Azure

SALES 1-800-867-1389 | MY ACCOUNT | PORTAL | Search

Why Azure | Solutions | Products | Documentation | Pricing | Partners | Blog | Resources | Support

Create your free Azure account today

Get $200 free credit

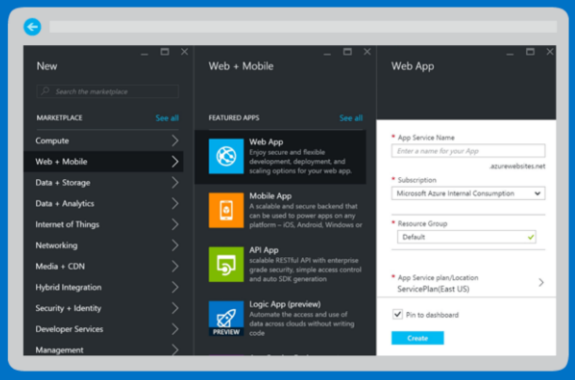
Start free with $200 in credit, and keep going with free options.

Try any Azure services

Explore our cloud by trying out any combination of Azure services for 30 days.

Pay nothing at the end

We use your credit card information for identity verification, but you'll never be charged unless you choose to subscribe.

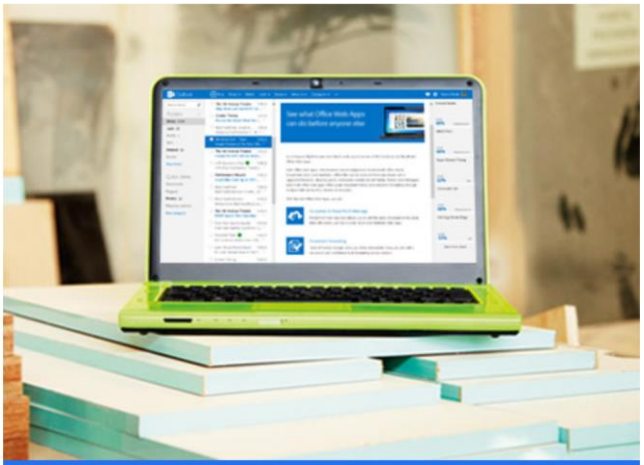


Start free

Or buy now >

Frequently asked questions >

Call sales 1-800-867-1389



Your account, our priority

Adding security information helps protect your account

Sign in

Microsoft account What's this?

[redacted]

Password

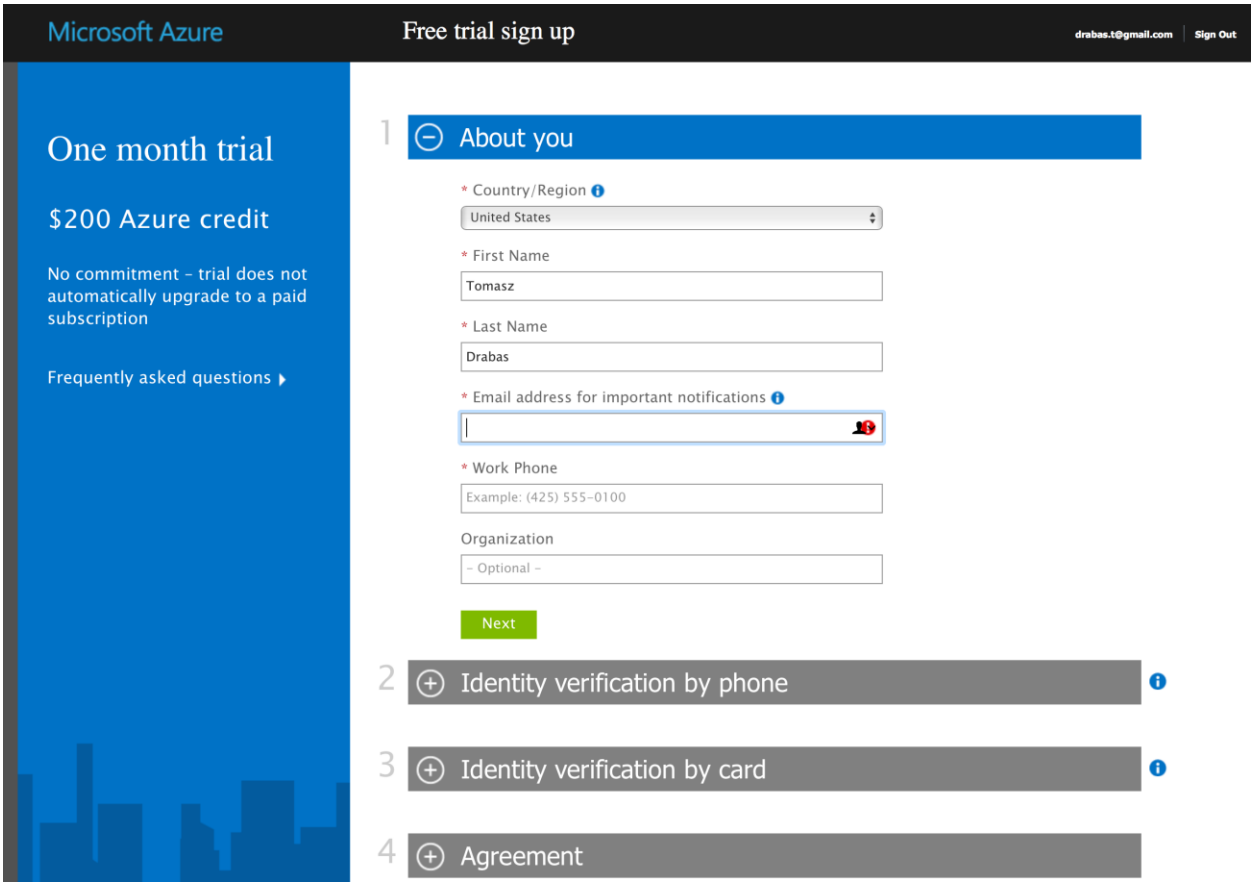
Keep me signed in

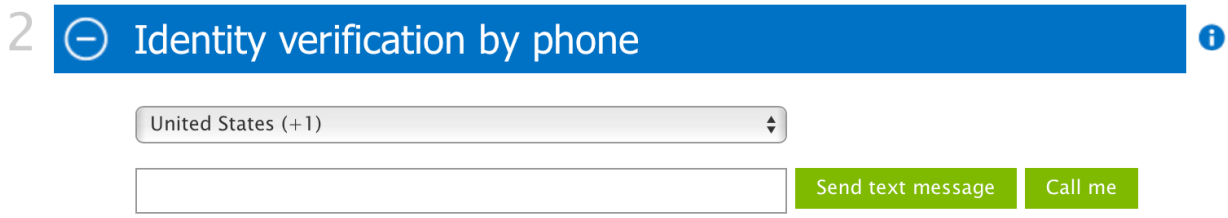
Sign in

Can't access your account?

Sign in with a single-use code

Don't have a Microsoft account? Sign up now


Microsoft Azure
Free trial sign up
drabas.t@gmail.com
Sign Out
One month trial
$200 Azure credit
No commitment - trial does not automatically upgrade to a paid subscription
Frequently asked questions
1 About you
* Country/Region
United States
* First Name
Tomasz
* Last Name
Drabas
* Email address for important notifications
* Work Phone
Example: (425) 555-0100
Organization
- Optional -
Next
2 Identity verification by phone
3 Identity verification by card
4 Agreement

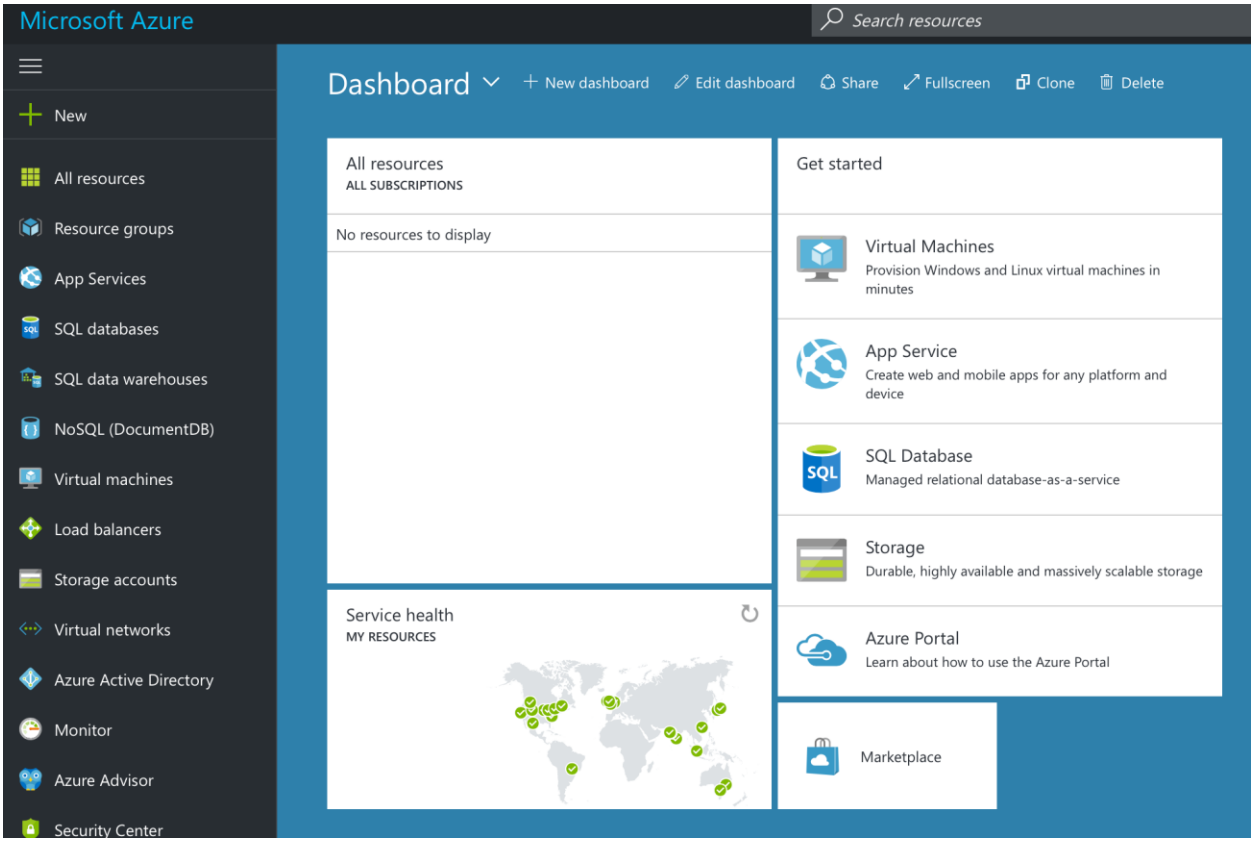

2 Identity verification by phone
United States (+1)
Send text message
Call me

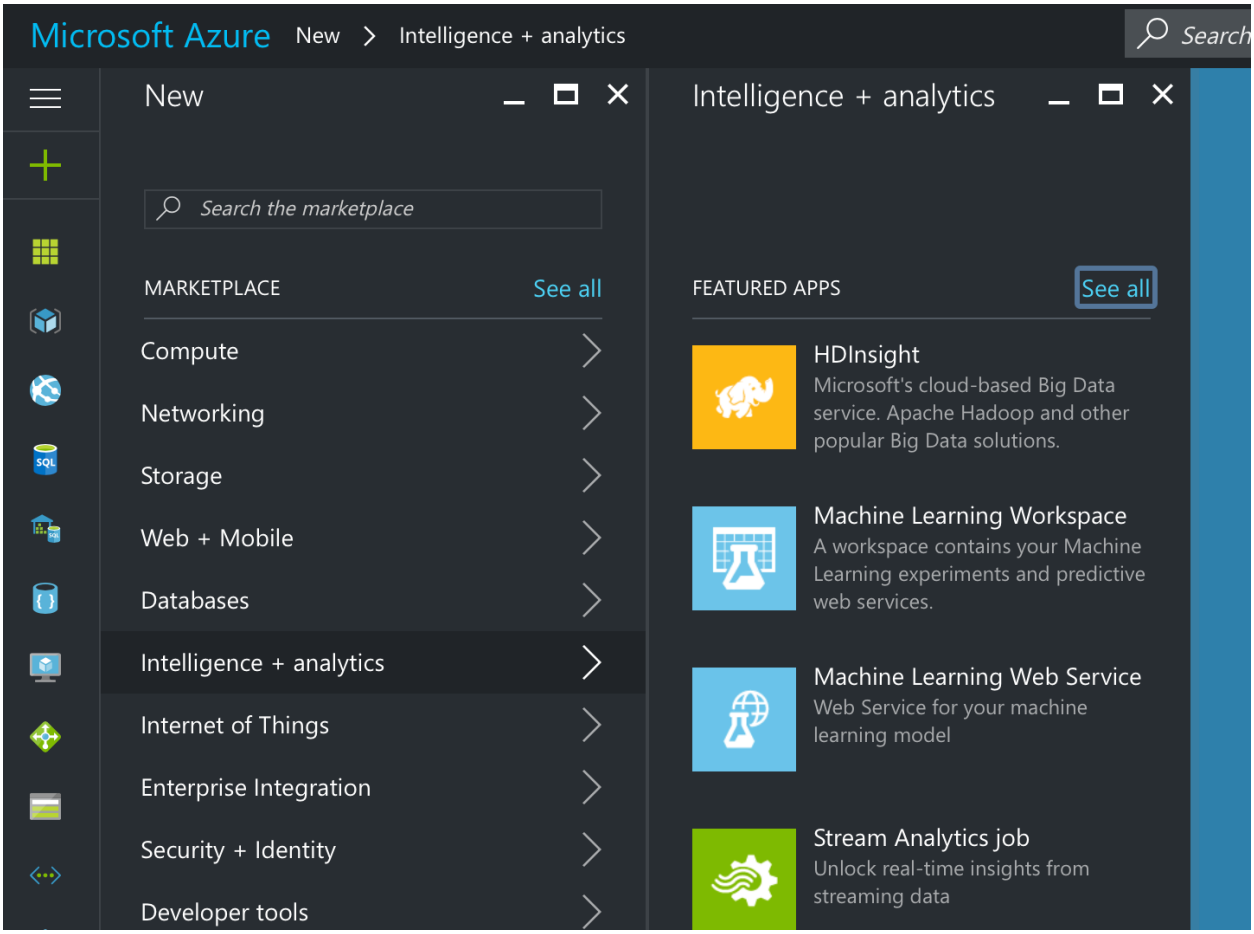
Welcome to Microsoft Azure!

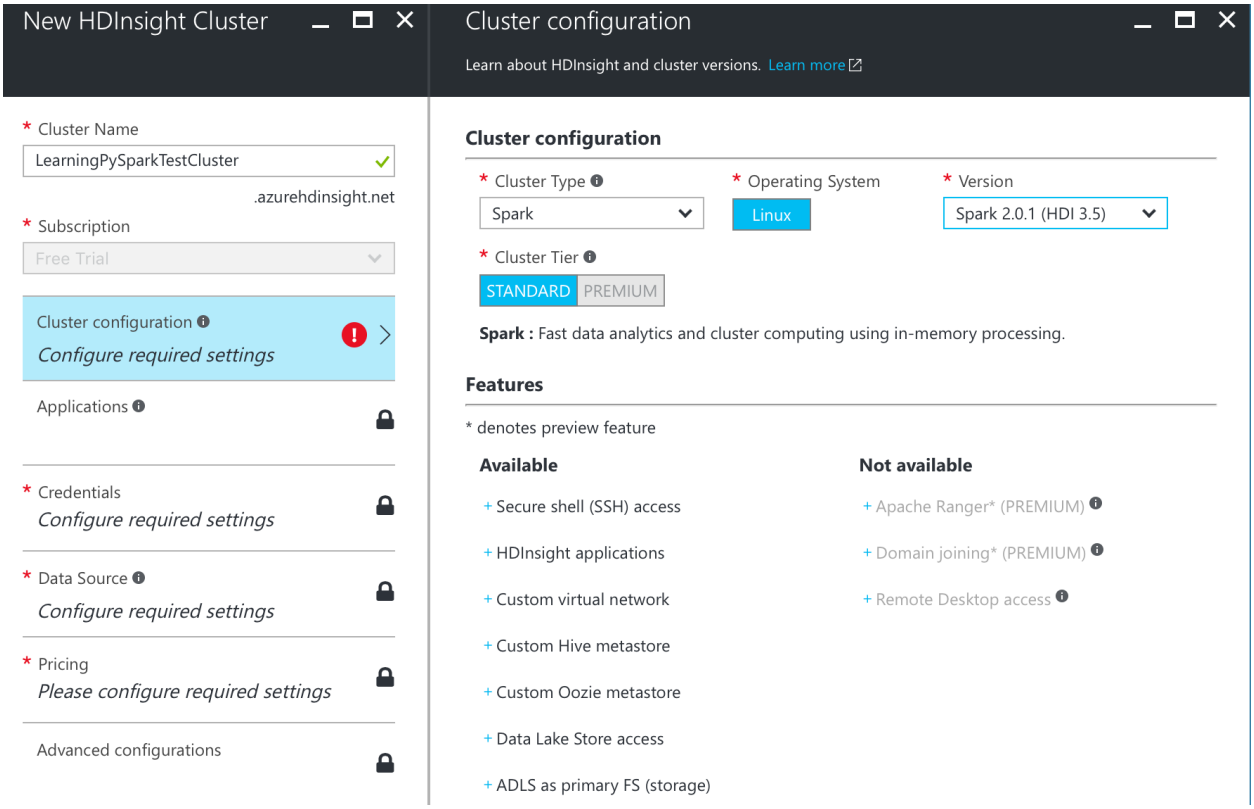
Your subscription - Free Trial

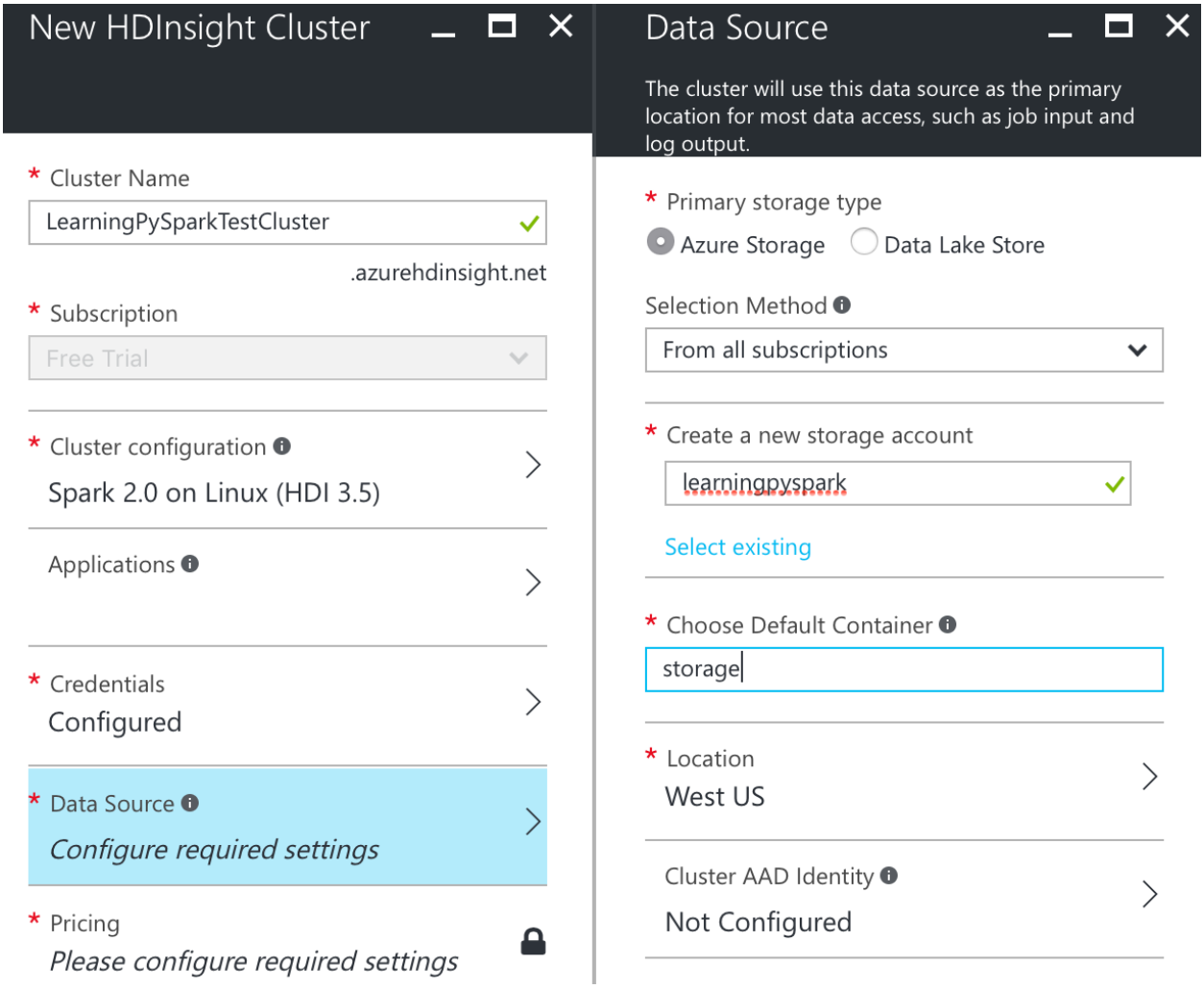
Your subscription is ready for you!

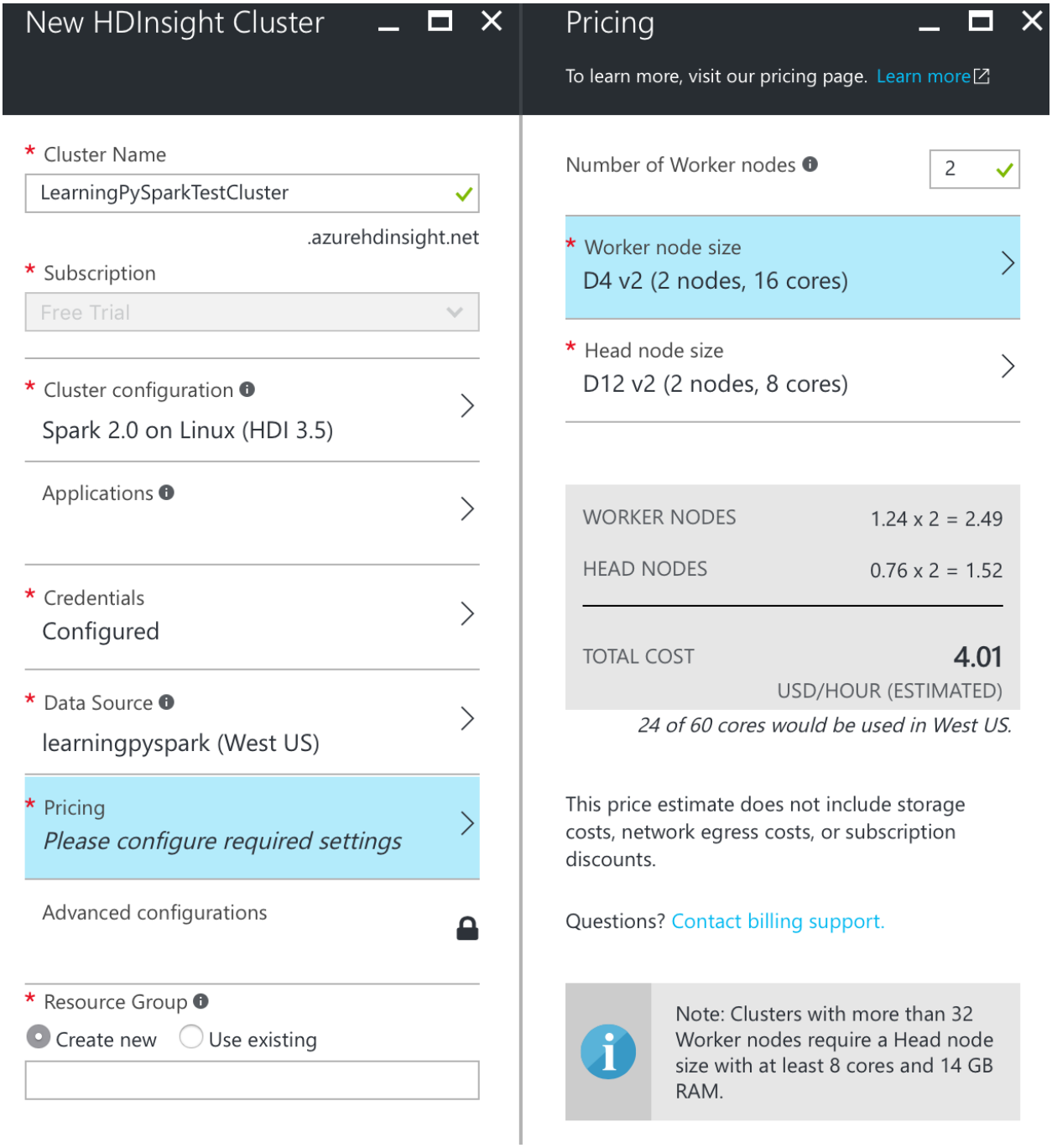
Start managing my service

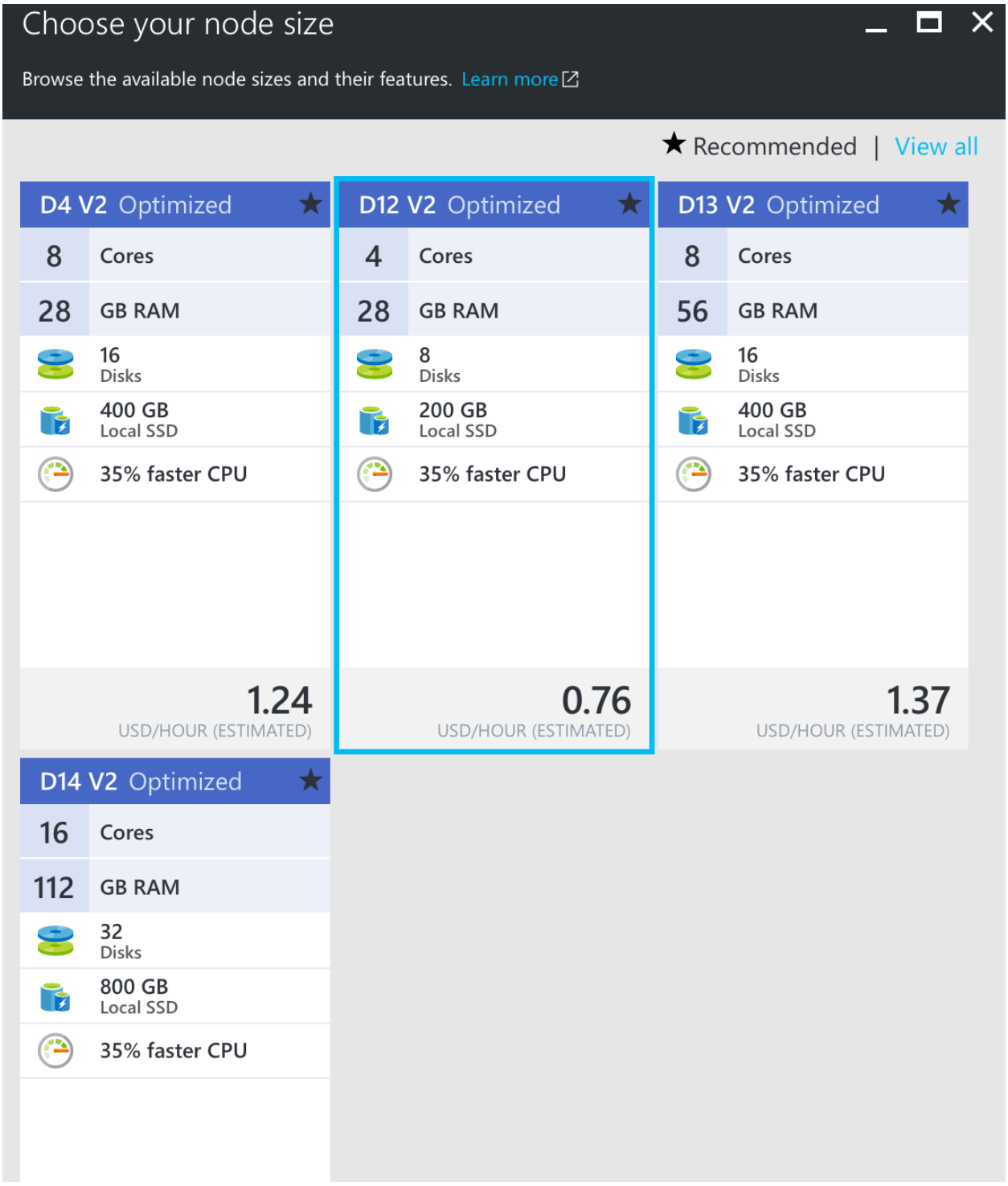


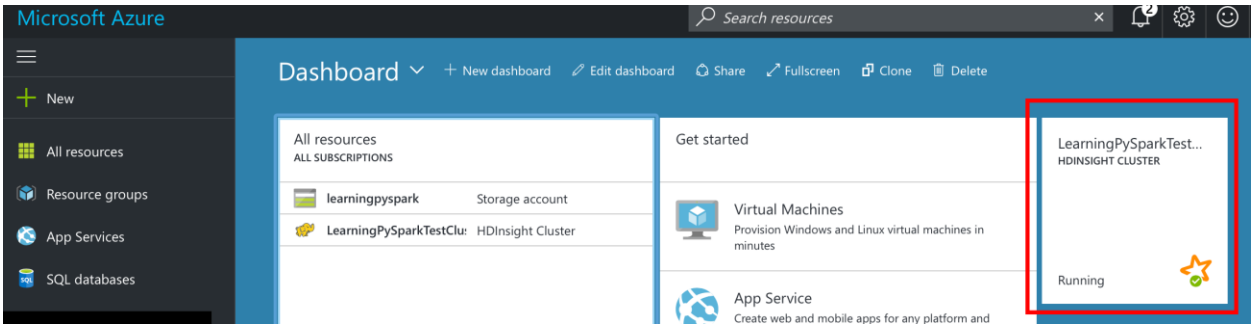

Microsoft Azure
New
Intelligence + analytics
Search
New
Search the marketplace
MARKETPLACE
See all
Compute
Networking
Storage
Web + Mobile
Databases
Intelligence + analytics
Internet of Things
Enterprise Integration
Security + Identity
Developer tools
Intelligence + analytics
FEATURED APPS
See all
HDInsight
Microsoft's cloud-based Big Data service. Apache Hadoop and other popular Big Data solutions.
Machine Learning Workspace
A workspace contains your Machine Learning experiments and predictive web services.
Machine Learning Web Service
Web Service for your machine learning model
Stream Analytics job
Unlock real-time insights from streaming data

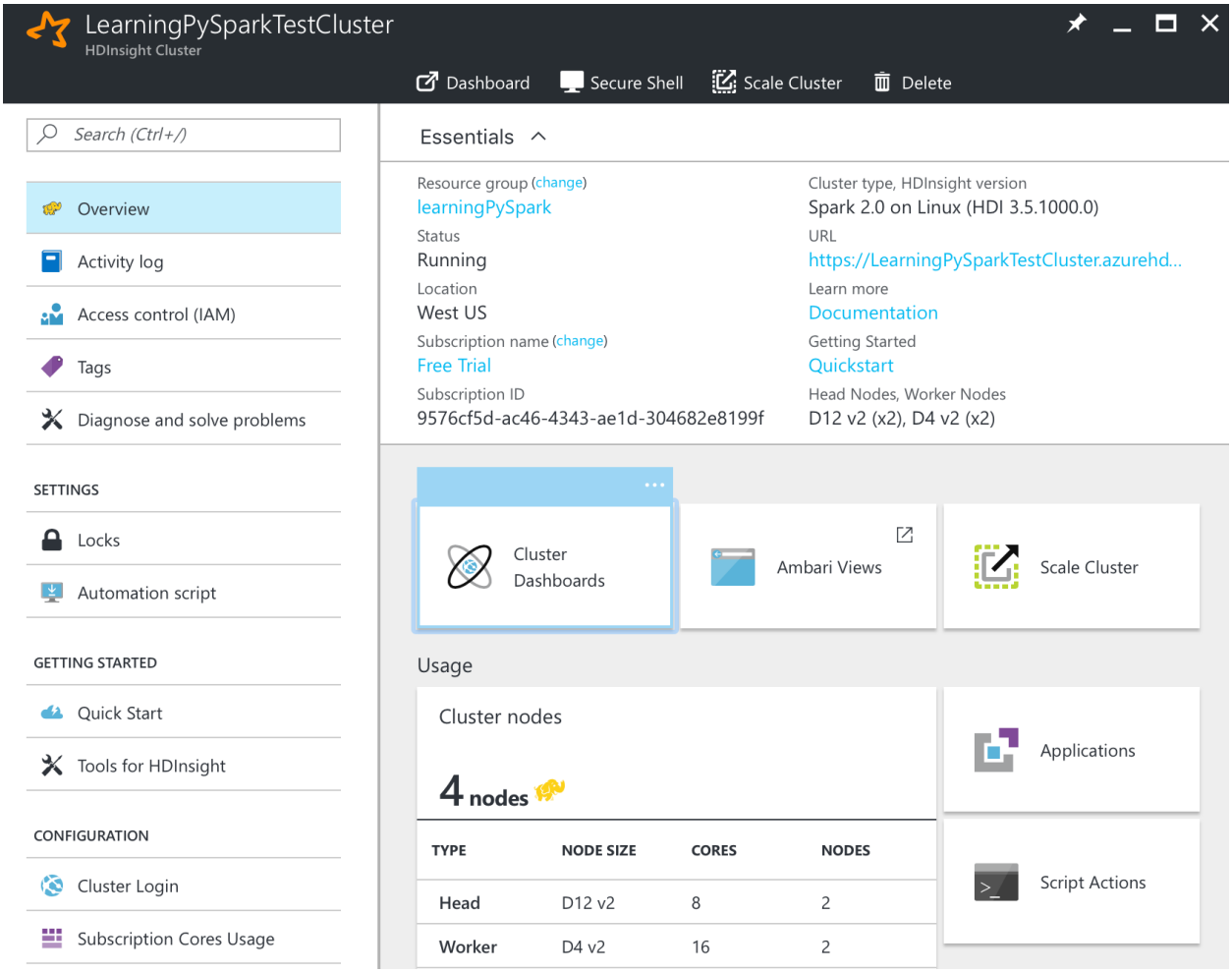

New HDInsight Cluster
* Cluster Name
LearningPySparkTestCluster
.azurehdinsight.net
* Subscription
Free Trial
Cluster configuration
Configure required settings
Applications
* Credentials
Configure required settings
* Data Source
Configure required settings
* Pricing
Please configure required settings
Advanced configurations
Cluster configuration
Learn about HDInsight and cluster versions. Learn more
Cluster configuration
* Cluster Type
Spark
* Operating System
Linux
* Version
Spark 2.0.1 (HDI 3.5)
* Cluster Tier
STANDARD
PREMIUM
Spark : Fast data analytics and cluster computing using in-memory processing.
Features
* denotes preview feature
Available
+ Secure shell (SSH) access
+ HDInsight applications
+ Custom virtual network
+ Custom Hive metastore
+ Custom Oozie metastore
+ Data Lake Store access
+ ADLS as primary FS (storage)
Not available
+ Apache Ranger* (PREMIUM)
+ Domain joining* (PREMIUM)
+ Remote Desktop access

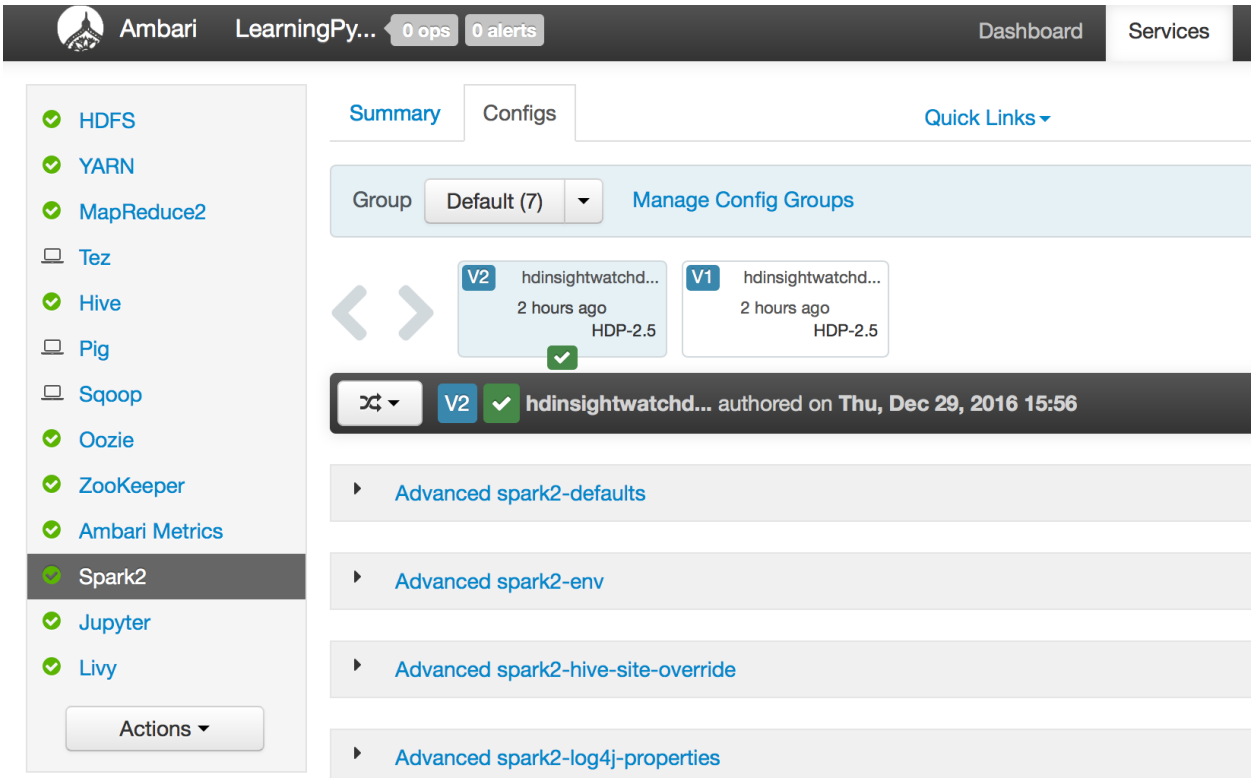

New HDInsight Cluster
* Cluster Name
LearningPySparkTestCluster
.azurehdinsight.net
* Subscription
Free Trial
* Cluster configuration
Spark 2.0 on Linux (HDI 3.5)
Applications
* Credentials
Configured
* Data Source
Configure required settings
* Pricing
Please configure required settings
Data Source
The cluster will use this data source as the primary location for most data access, such as job input and log output.
* Primary storage type
Azure Storage
Data Lake Store
Selection Method
From all subscriptions
* Create a new storage account
learningpyspark
Select existing
* Choose Default Container
storage
* Location
West US
Cluster AAD Identity
Not Configured

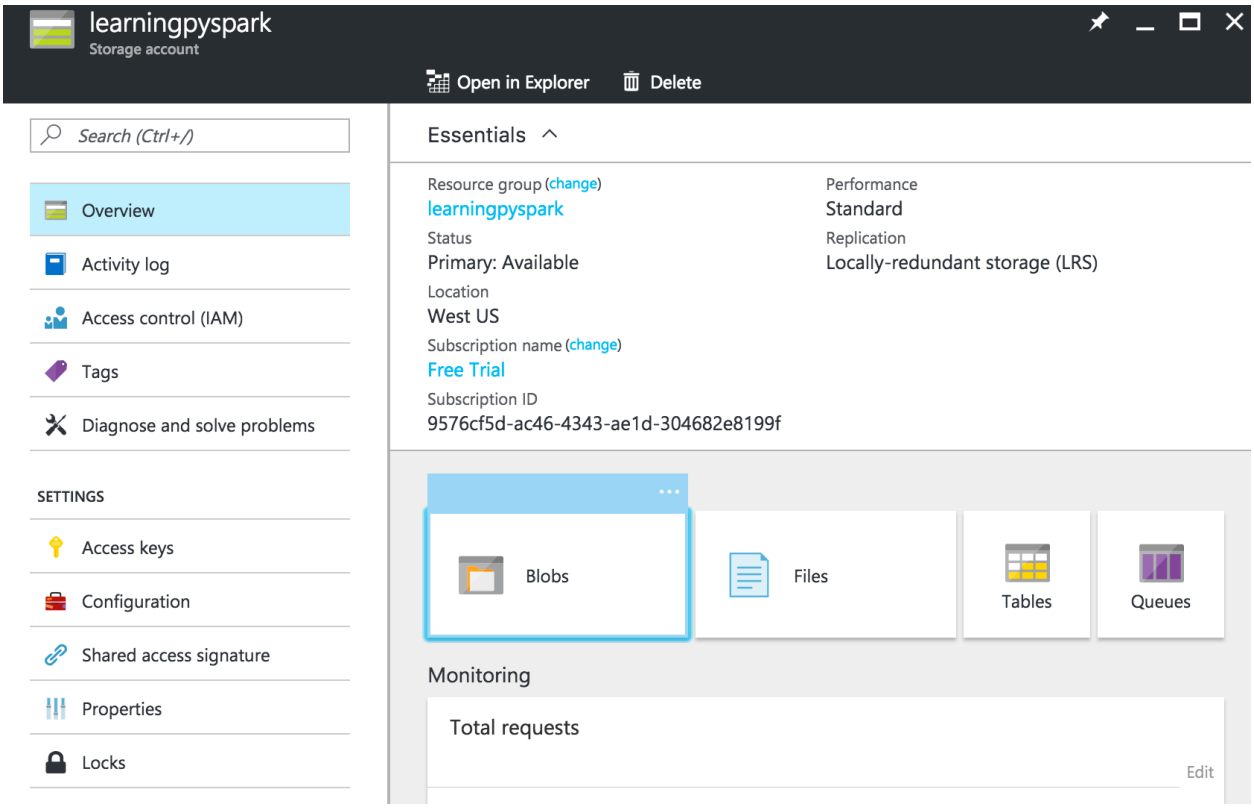

New HDInsight Cluster
* Cluster Name
LearningPySparkTestCluster
.azurehdinsight.net
* Subscription
Free Trial
* Cluster configuration
Spark 2.0 on Linux (HDI 3.5)
Applications
* Credentials
Configured
* Data Source
learningpyspark (West US)
* Pricing
Please configure required settings
Advanced configurations
* Resource Group
Create new
Use existing
Pricing
To learn more, visit our pricing page. Learn more
Number of Worker nodes
2
* Worker node size
D4 v2 (2 nodes, 16 cores)
* Head node size
D12 v2 (2 nodes, 8 cores)
WORKER NODES
1.24 x 2 = 2.49
HEAD NODES
0.76 x 2 = 1.52
TOTAL COST
4.01
USD/HOUR (ESTIMATED)
24 of 60 cores would be used in West US.
This price estimate does not include storage costs, network egress costs, or subscription discounts.
Questions? Contact billing support.
Note: Clusters with more than 32 Worker nodes require a Head node size with at least 8 cores and 14 GB RAM.

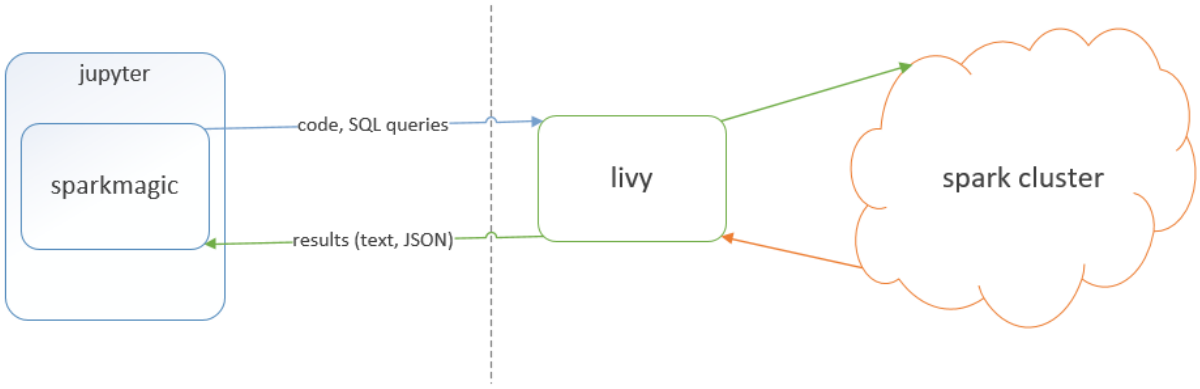

Choose your node size
Browse the available node sizes and their features. Learn more
★ Recommended | View all
D4 V2 Optimized
8 Cores
28 GB RAM
16 Disks
400 GB Local SSD
35% faster CPU
1.24
USD/HOUR (ESTIMATED)
D12 V2 Optimized
4 Cores
28 GB RAM
8 Disks
200 GB Local SSD
35% faster CPU
0.76
USD/HOUR (ESTIMATED)
D13 V2 Optimized
8 Cores
56 GB RAM
16 Disks
400 GB Local SSD
35% faster CPU
1.37
USD/HOUR (ESTIMATED)
D14 V2 Optimized
16 Cores
112 GB RAM
32 Disks
800 GB Local SSD
35% faster CPU


Microsoft Azure
Search resources
New
All resources
Resource groups
App Services
SQL databases
Dashboard
New dashboard
Edit dashboard
Share
Fullscreen
Clone
Delete
All resources
ALL SUBSCRIPTIONS
learningpyspark
Storage account
LearningPySparkTestClu:
HDInsight Cluster
Get started
Virtual Machines
Provision Windows and Linux virtual machines in minutes
App Service
Create web and mobile apps for any platform and
LearningPySparkTest...
HDINSIGHT CLUSTER
Running


LearningPySparkTestCluster
HDInsight Cluster
Dashboard
Secure Shell
Scale Cluster
Delete
Search (Ctrl+/)
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
SETTINGS
Locks
Automation script
GETTING STARTED
Quick Start
Tools for HDInsight
CONFIGURATION
Cluster Login
Subscription Cores Usage
Essentials
Resource group (change)
learningPySpark
Status
Running
Location
West US
Subscription name (change)
Free Trial
Subscription ID
9576cf5d-ac46-4343-ae1d-304682e8199f
Cluster type, HDInsight version
Spark 2.0 on Linux (HDI 3.5.1000.0)
URL
https://LearningPySparkTestCluster.azurehd...
Learn more
Documentation
Getting Started
Quickstart
Head Nodes, Worker Nodes
D12 v2 (x2), D4 v2 (x2)
Cluster Dashboards
Ambari Views
Scale Cluster
Usage
Cluster nodes
4 nodes
TYPE
NODE SIZE
CORES
NODES
Head
D12 v2
8
2
Worker
D4 v2
16
2
Applications
Script Actions


Ambari
LearningPy...
0 ops
0 alerts
Dashboard
Services
HDFS
YARN
MapReduce2
Tez
Hive
Pig
Sqoop
Oozie
ZooKeeper
Ambari Metrics
Spark2
Jupyter
Livy
Actions
Summary
Configs
Quick Links
Group
Default (7)
Manage Config Groups
V2 hdinsightwatchd...
2 hours ago
HDP-2.5
V1 hdinsightwatchd...
2 hours ago
HDP-2.5
V2 hdinsightwatchd... authored on Thu, Dec 29, 2016 15:56
Advanced spark2-defaults
Advanced spark2-env
Advanced spark2-hive-site-override
Advanced spark2-log4j-properties


learningpyspark
Storage account
Open in Explorer
Delete
Search (Ctrl+/)
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
SETTINGS
Access keys
Configuration
Shared access signature
Properties
Locks
Essentials
Resource group (change)
learningpyspark
Status
Primary: Available
Location
West US
Subscription name (change)
Free Trial
Subscription ID
9576cf5d-ac46-4343-ae1d-304682e8199f
Performance
Standard
Replication
Locally-redundant storage (LRS)
Blobs
Files
Tables
Queues
Monitoring
Total requests
Edit



Current session configs: `{u'executorCores': 4, u'numExecutors': 2, u'executorMemory': u'2GB', u'name': u'learningPySpark_Example', u'kind': 'pyspark'}`

No active sessions.

Magic	Example	Explanation
info	%%info	Outputs session information for the current Livy endpoint.
cleanup	%%cleanup -f	Deletes all sessions for the current Livy endpoint, including this notebook's session. The force flag is mandatory.
delete	%%delete -f -s 0	Deletes a session by number for the current Livy endpoint. Cannot delete this kernel's session.
logs	%%logs	Outputs the current session's Livy logs.
configure	%%configure -f {"executorMemory": "1000M", "executorCores": 4}	Configure the session creation parameters. The force flag is mandatory if a session has already been created and the session will be dropped and recreated. Look at Livy's POST /sessions Request Body for a list of valid parameters. Parameters must be passed in as a JSON string.
sql	%%sql -o tables -q SHOW TABLES	Executes a SQL query against the variable sqlContext (Spark v1.x) or spark (Spark v2.x). Parameters: • -o VAR_NAME: The result of the query will be available in the %%local Python context as a Pandas dataframe. • -q: The magic will return None instead of the dataframe (no visualization). • -m METHOD: Sample method, either `take` or `sample`. • -n MAXROWS: The maximum number of rows of a SQL query that will be pulled from Livy to Jupyter. If this number is negative, then the number of rows will be unlimited. • -r FRACTION: Fraction used for sampling.
local	%%local a = 1	All the code in subsequent lines will be executed locally. Code must be valid Python code.

`Starting Spark application`

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
5	application_1483055828481_0010	pyspark	idle	Link	Link	✔

`SparkSession available as 'spark'.`

```
Row(BIRTH_PLACE=1, count=44558)
Row(BIRTH_PLACE=2, count=136)
Row(BIRTH_PLACE=3, count=224)
Row(BIRTH_PLACE=4, count=327)
Row(BIRTH_PLACE=5, count=74)
Row(BIRTH_PLACE=6, count=11)
Row(BIRTH_PLACE=7, count=91)
Row(BIRTH_PLACE=9, count=8)
```

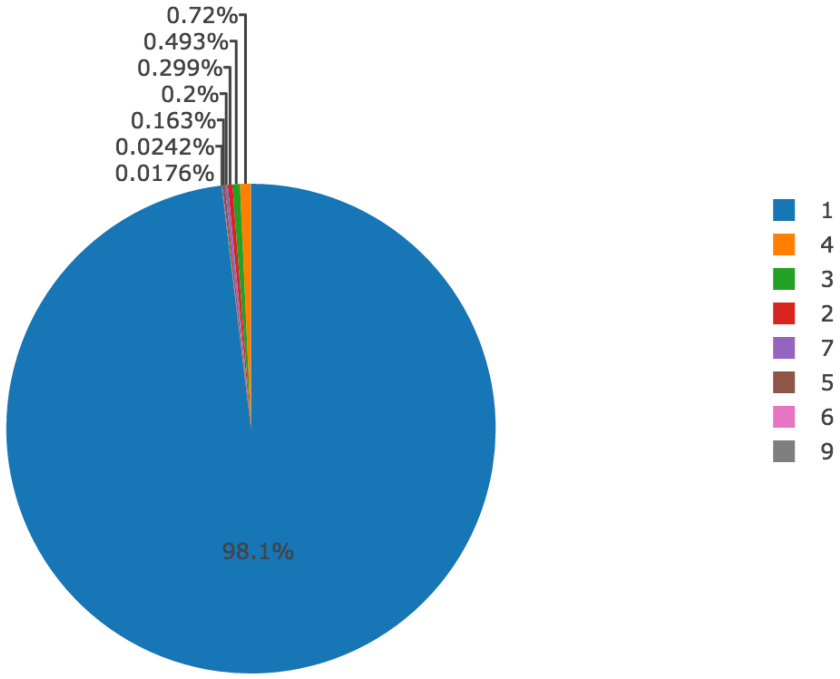
BIRTH_PLACE	Count
1	44558
2	136
3	224
4	327
5	74
6	11
7	91
9	8

Type: Table | Pie | Scatter | Line | Area | Bar

Encoding:

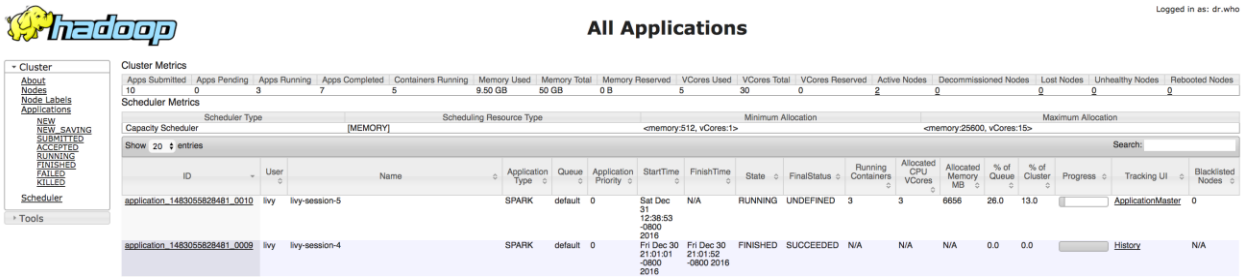
X BIRTH_PLACE

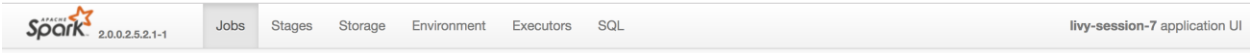
Y Count Func. Max



Type: Table Pie Scatter Line Area Bar

BIRTH_PLACE	Count
1	44558
6	11
3	224
5	74
9	8





Spark Jobs (?)

User: yarn
Total Uptime: 8.0 min
Scheduling Mode: FIFO
Completed Jobs: 9

▸ Event Timeline

Completed Jobs (9)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
8	runJob at PythonRDD.scala:441	2016/12/31 22:54:49	75 ms	1/1 (2 skipped)	4/4 (201 skipped)
7	runJob at PythonRDD.scala:441	2016/12/31 22:54:49	67 ms	1/1 (2 skipped)	4/4 (201 skipped)
6	runJob at PythonRDD.scala:441	2016/12/31 22:54:48	0.9 s	2/2 (1 skipped)	201/201 (1 skipped)
5	toJSON at NativeMethodAccessorImpl.java:-2	2016/12/31 22:54:47	0.8 s	2/2	201/201
4	collect at <stdin>:4	2016/12/31 22:54:43	0.6 s	2/2 (1 skipped)	209/209 (1 skipped)
3	collect at <stdin>:4	2016/12/31 22:54:41	2 s	2/2	201/201
2	csv at NativeMethodAccessorImpl.java:-2	2016/12/31 22:54:35	2 s	1/1	1/1
1	csv at NativeMethodAccessorImpl.java:-2	2016/12/31 22:54:35	0.1 s	1/1	1/1
0	csv at NativeMethodAccessorImpl.java:-2	2016/12/31 22:54:33	1 s	1/1	1/1

Details for Stage 6 (Attempt 0)

Total Time Across All Tasks: 1.0 s
Locality Level Summary: Node local: 8; Process local: 192
Shuffle Read: 504.0 B / 8
Shuffle Write: 504.0 B / 8

▸ DAG Visualization
▸ Show Additional Metrics
▸ Event Timeline

Summary Metrics for 200 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	1 ms	2 ms	3 ms	5 ms	27 ms
GC Time	0 ms	0 ms	0 ms	0 ms	0 ms
Shuffle Read Size / Records	0.0 B / 0	0.0 B / 0	0.0 B / 0	0.0 B / 0	63.0 B / 1
Shuffle Write Size / Records	0.0 B / 0	0.0 B / 0	0.0 B / 0	0.0 B / 0	63.0 B / 1

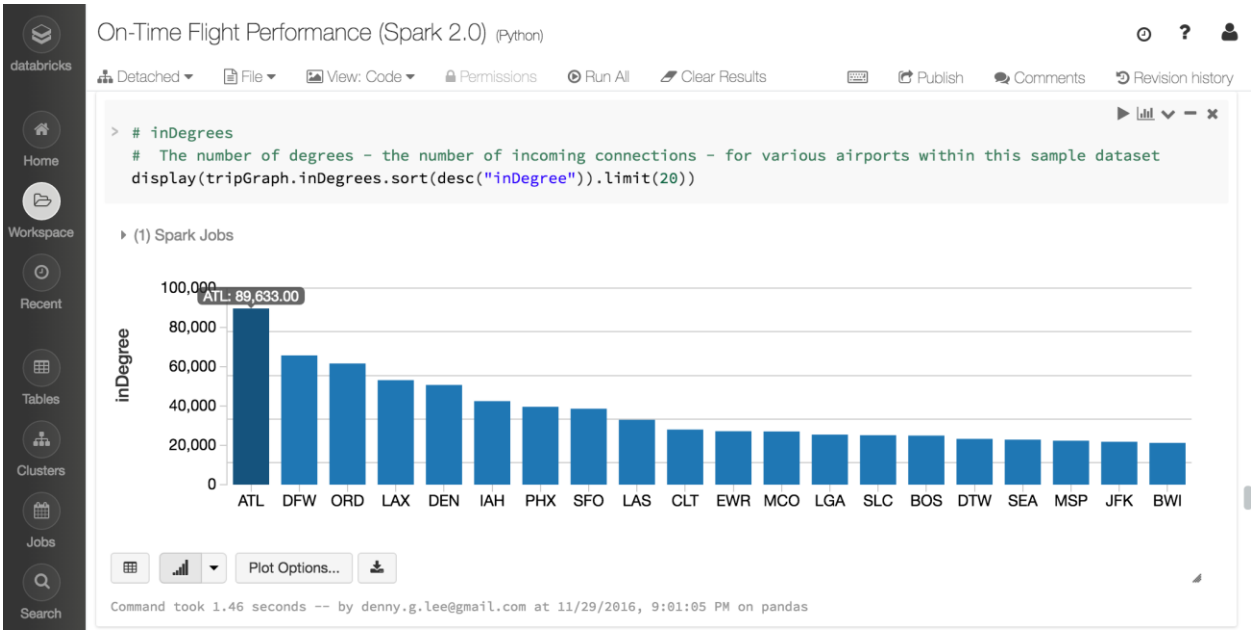
Aggregated Metrics by Executor

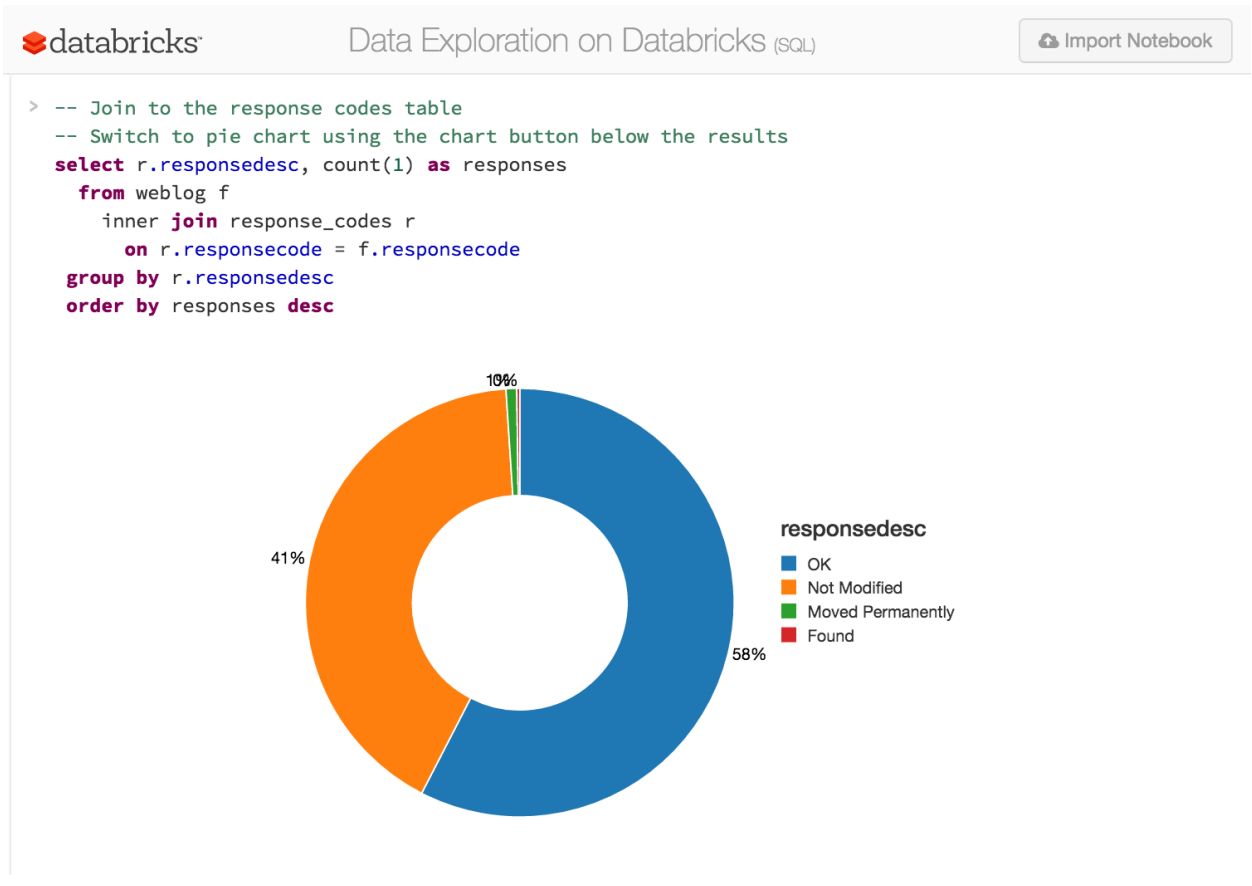
Executor ID ▲	Address	Task Time	Total Tasks	Failed Tasks	Succeeded Tasks	Shuffle Read Size / Records	Shuffle Write Size / Records
1	10.0.0.7:44793	2 s	150	0	150	504.0 B / 8	504.0 B / 8
2	10.0.0.6:46439	2 s	50	0	50	0.0 B / 0	0.0 B / 0

Tasks (200)

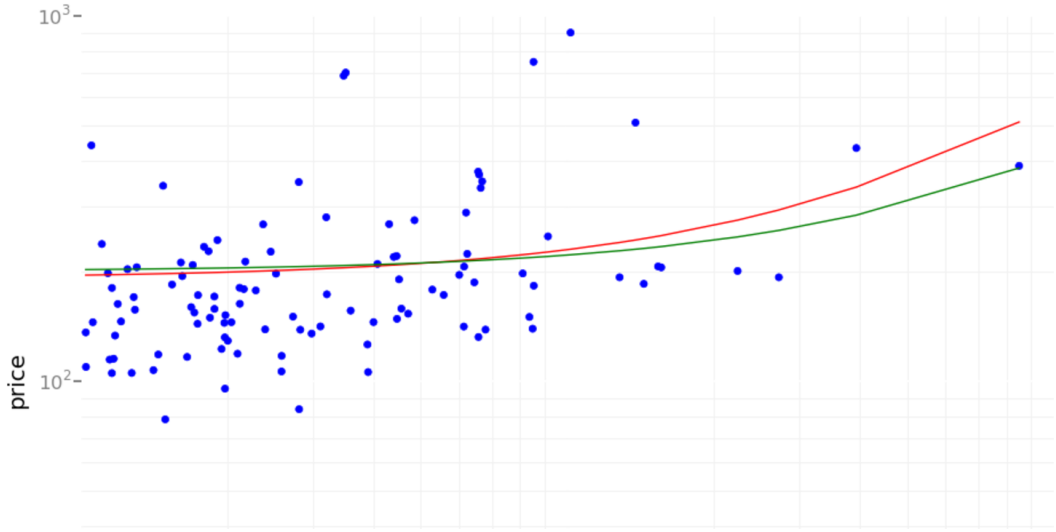
Page: 1 2 >
2 Pages. Jump to 1 . Show 100 items in a page. Go

Index ▲	ID	Attempt	Status	Locality Level	Executor ID / Host	Launch Time	Duration	GC Time	Shuffle Read Size / Records	Write Time	Shuffle Write Size / Records	Errors
0	208	0	SUCCESS	PROCESS_LOCAL	2 / 10.0.0.6	2016/12/31 22:54:43	24 ms		0.0 B / 0		0.0 B / 0	
1	209	0	SUCCESS	PROCESS_LOCAL	2 / 10.0.0.6	2016/12/31 22:54:43	11 ms		0.0 B / 0		0.0 B / 0	
2	210	0	SUCCESS	PROCESS_LOCAL	2 / 10.0.0.6	2016/12/31 22:54:43	15 ms		0.0 B / 0		0.0 B / 0	
3	211	0	SUCCESS	PROCESS_LOCAL	2 / 10.0.0.6	2016/12/31 22:54:43	9 ms		0.0 B / 0		0.0 B / 0	
4	216	0	SUCCESS	PROCESS_LOCAL	2 / 10.0.0.6	2016/12/31 22:54:43	8 ms		0.0 B / 0		0.0 B / 0	
5	217	0	SUCCESS	PROCESS_LOCAL	2 / 10.0.0.6	2016/12/31 22:54:43	7 ms		0.0 B / 0		0.0 B / 0	


databricks
On-Time Flight Performance (Spark 2.0) (Python)
Detached
File
View: Code
Permissions
Run All
Clear Results
Publish
Comments
Revision history
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
# inDegrees
# The number of degrees - the number of incoming connections - for various airports within this sample dataset
display(tripGraph.inDegrees.sort(desc("inDegree")).limit(20))
(1) Spark Jobs
inDegree
100,000
80,000
60,000
40,000
20,000
0
ATL: 89,633.00
ATL DFW ORD LAX DEN IAH PHX SFO LAS CLT EWR MCO LGA SLC BOS DTW SEA MSP JFK BWI
Plot Options...
Command took 1.46 seconds -- by denny.g.lee@gmail.com at 11/29/2016, 9:01:05 PM on pandas


databricks
Data Exploration on Databricks (SQL)
Import Notebook
-- Join to the response codes table
-- Switch to pie chart using the chart button below the results
select r.responsedesc, count(1) as responses
from weblog f
inner join response_codes r
on r.responsecode = f.responsecode
group by r.responsedesc
order by responses desc
41%
58%
responsedesc
OK
Not Modified
Moved Permanently
Found

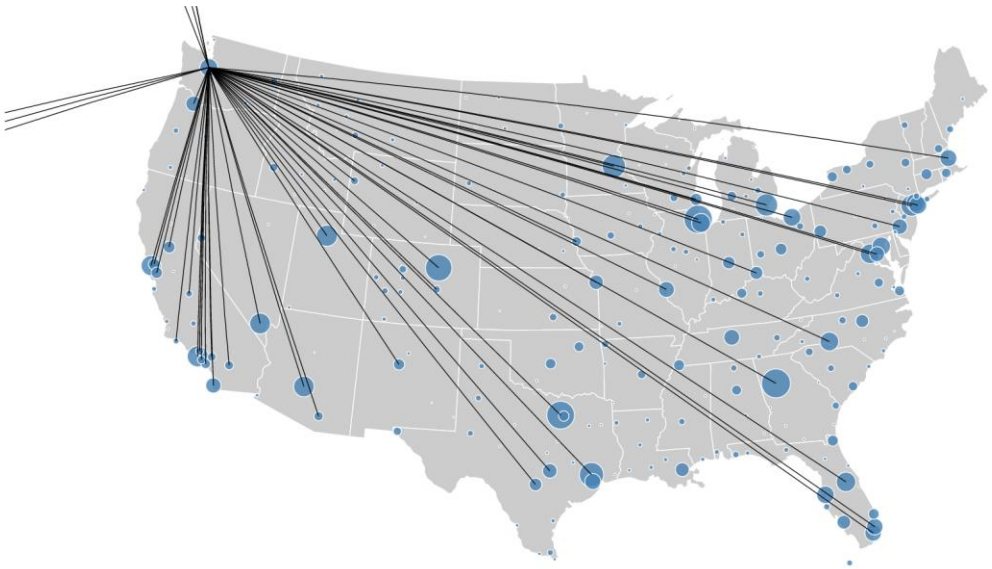
```
> p = ggplot(pydf, aes('pop','price')) + \
      geom_point(color='blue') + \
      geom_line(pydf, aes('pop','predA'), color='red') + \
      geom_line(pydf, aes('pop','predB'), color='green') + \
      scale_x_log10() + scale_y_log10()
  display(p)
```



databricks

On-Time Flight Performance (Python)

Import Notebook



databricks

On-Time Flight Performance (Spark 2.0) (Python)

Detached | File | View: Code | Permissions | Run All | Clear Results | Publish | Comments | Revision history

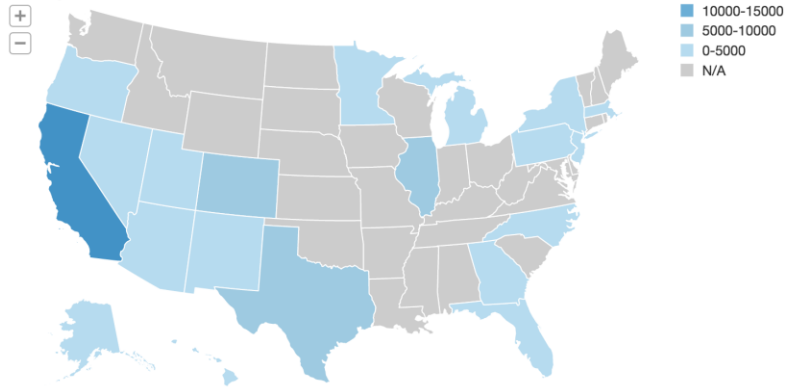
Home
Workspace
Recent
Tables
Clusters
Jobs
Search

What destinations tend to have significant delays departing from SEA

```
> # States with the longest cumulative delays (with individual delays > 100 minutes) (origin: Seattle)
  display(tripGraph.edges.filter("src = 'SEA' and delay > 100"))
```

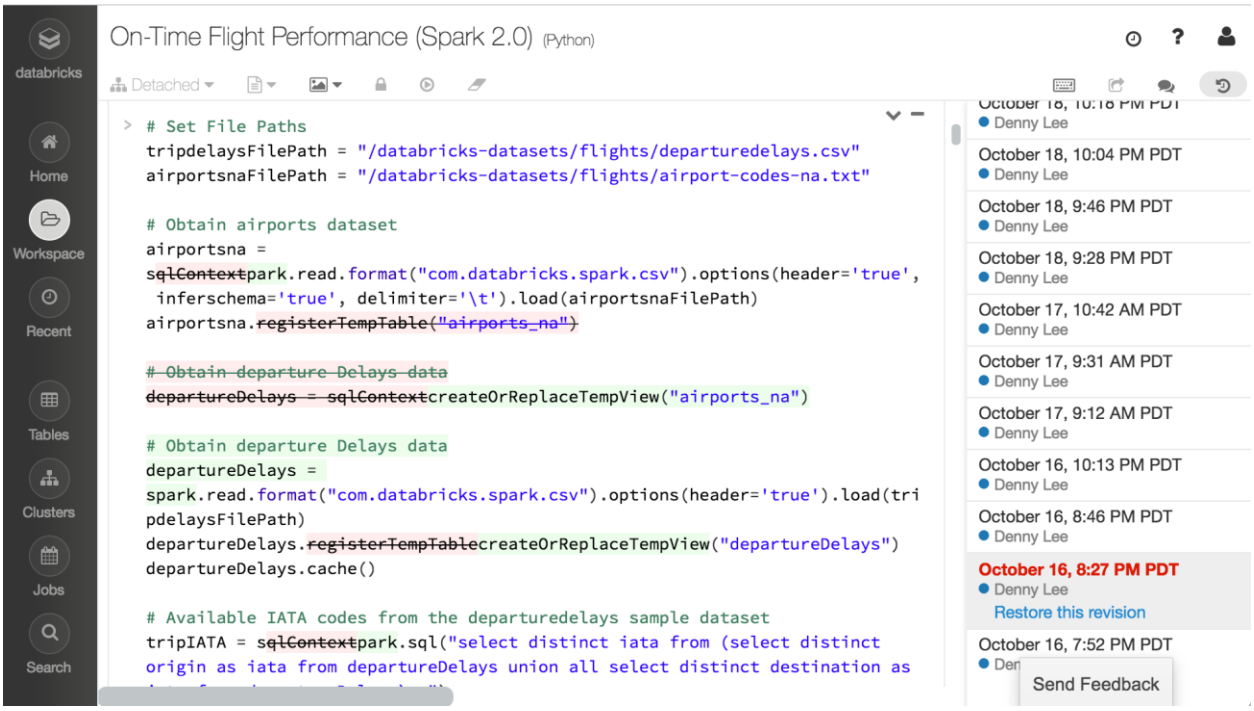
▸ (2) Spark Jobs

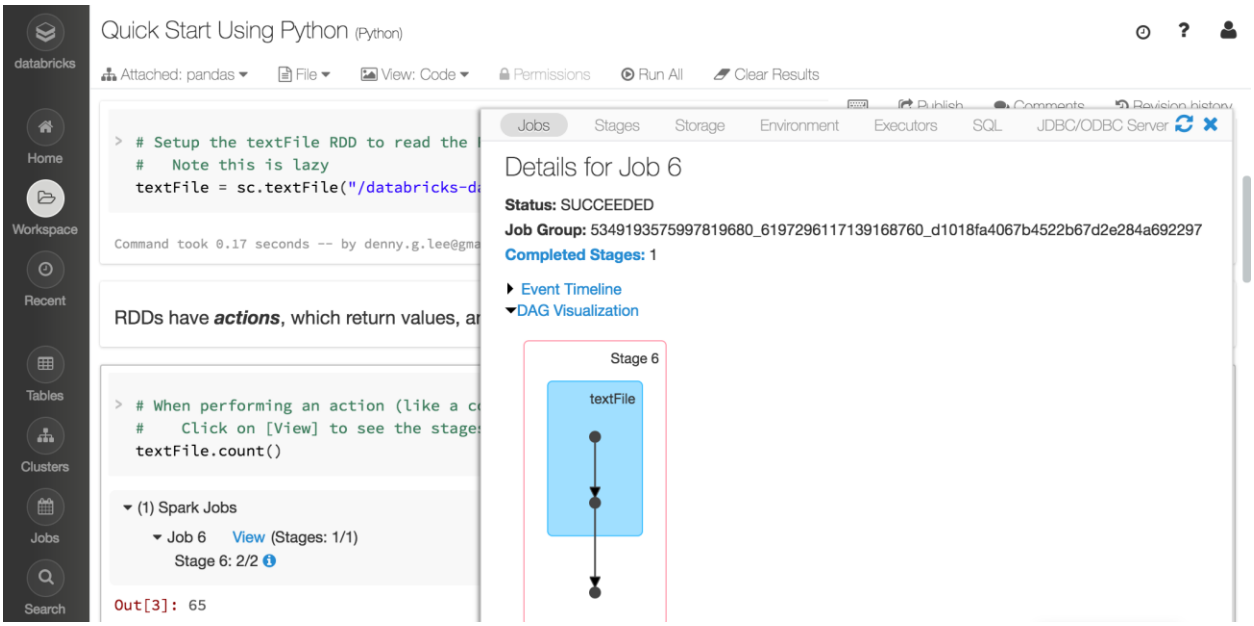
Following states were not found:

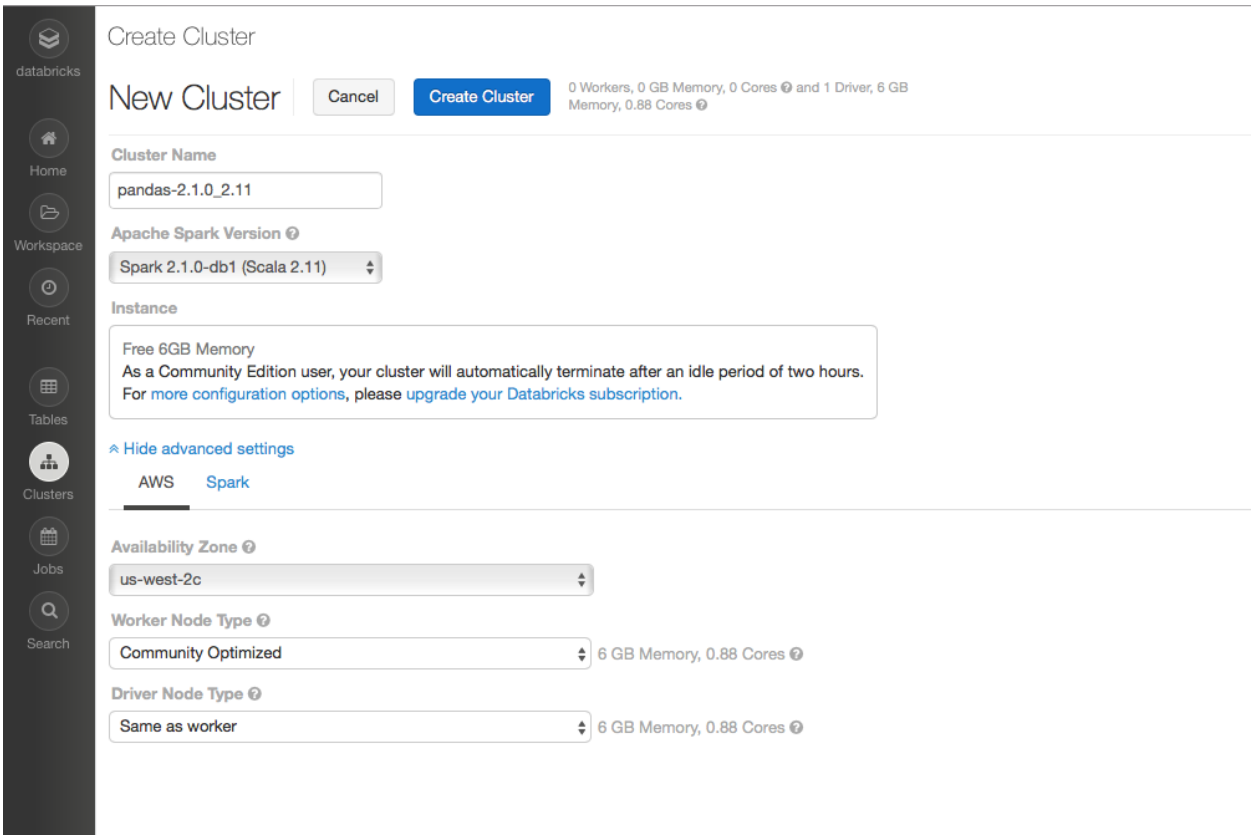


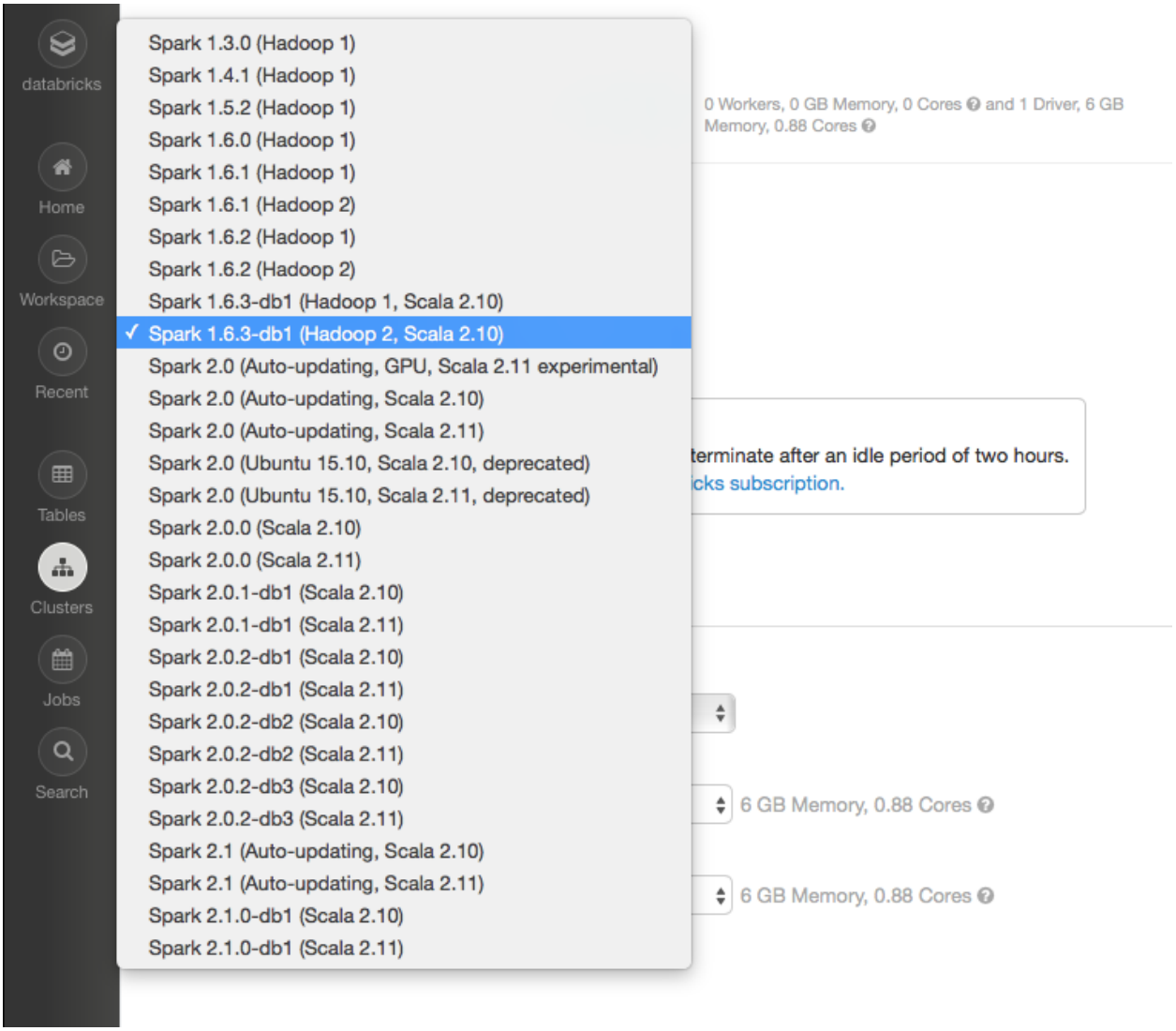
Denny Lee
1/5/2017, 9:35:10 AM

Could you increase the delay to 500?


databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
On-Time Flight Performance (Spark 2.0) (Python)
Detached
# Set File Paths
tripdelaysFilePath = "/databricks-datasets/flights/departuredelays.csv"
airportsnaFilePath = "/databricks-datasets/flights/airport-codes-na.txt"
# Obtain airports dataset
airportsna =
sqlContextpark.read.format("com.databricks.spark.csv").options(header='true',
inferschema='true', delimiter='\t').load(airportsnaFilePath)
airportsna.registerTempTable("airports_na")
# Obtain departure Delays data
departureDelays = sqlContextcreateOrReplaceTempView("airports_na")
# Obtain departure Delays data
departureDelays =
spark.read.format("com.databricks.spark.csv").options(header='true').load(tri
pdelaysFilePath)
departureDelays.registerTempTablecreateOrReplaceTempView("departureDelays")
departureDelays.cache()
# Available IATA codes from the departuredelays sample dataset
tripIATA = sqlContextpark.sql("select distinct iata from (select distinct
origin as iata from departureDelays union all select distinct destination as
Denny Lee
October 18, 10:04 PM PDT
Denny Lee
October 18, 9:46 PM PDT
Denny Lee
October 18, 9:28 PM PDT
Denny Lee
October 17, 10:42 AM PDT
Denny Lee
October 17, 9:31 AM PDT
Denny Lee
October 17, 9:12 AM PDT
Denny Lee
October 16, 10:13 PM PDT
Denny Lee
October 16, 8:46 PM PDT
Denny Lee
October 16, 8:27 PM PDT
Denny Lee
Restore this revision
October 16, 7:52 PM PDT
Send Feedback


databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
Quick Start Using Python (Python)
Attached: pandas
File
View: Code
Permissions
Run All
Clear Results
# Setup the textFile RDD to read the
# Note this is lazy
textFile = sc.textFile("/databricks-d
Command took 0.17 seconds -- by denny.g.lee@gma
RDDs have ***actions***, which return values, a
# When performing an action (like a c
# Click on [View] to see the stage
textFile.count()
(1) Spark Jobs
Job 6 View (Stages: 1/1)
Stage 6: 2/2
Out[3]: 65
Jobs
Stages
Storage
Environment
Executors
SQL
JDBC/ODBC Server
Details for Job 6
Status: SUCCEEDED
Job Group: 5349193575997819680_6197296117139168760_d1018fa4067b4522b67d2e284a692297
Completed Stages: 1
Event Timeline
DAG Visualization
Stage 6
textFile


databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
Create Cluster
New Cluster
Cancel
Create Cluster
0 Workers, 0 GB Memory, 0 Cores and 1 Driver, 6 GB Memory, 0.88 Cores
Cluster Name
pandas-2.1.0_2.11
Apache Spark Version
Spark 2.1.0-db1 (Scala 2.11)
Instance
Free 6GB Memory
As a Community Edition user, your cluster will automatically terminate after an idle period of two hours.
For more configuration options, please upgrade your Databricks subscription.
Hide advanced settings
AWS
Spark
Availability Zone
us-west-2c
Worker Node Type
Community Optimized
6 GB Memory, 0.88 Cores
Driver Node Type
Same as worker
6 GB Memory, 0.88 Cores


databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
Spark 1.3.0 (Hadoop 1)
Spark 1.4.1 (Hadoop 1)
Spark 1.5.2 (Hadoop 1)
Spark 1.6.0 (Hadoop 1)
Spark 1.6.1 (Hadoop 1)
Spark 1.6.1 (Hadoop 2)
Spark 1.6.2 (Hadoop 1)
Spark 1.6.2 (Hadoop 2)
Spark 1.6.3-db1 (Hadoop 1, Scala 2.10)
✓ Spark 1.6.3-db1 (Hadoop 2, Scala 2.10)
Spark 2.0 (Auto-updating, GPU, Scala 2.11 experimental)
Spark 2.0 (Auto-updating, Scala 2.10)
Spark 2.0 (Auto-updating, Scala 2.11)
Spark 2.0 (Ubuntu 15.10, Scala 2.10, deprecated)
Spark 2.0 (Ubuntu 15.10, Scala 2.11, deprecated)
Spark 2.0.0 (Scala 2.10)
Spark 2.0.0 (Scala 2.11)
Spark 2.0.1-db1 (Scala 2.10)
Spark 2.0.1-db1 (Scala 2.11)
Spark 2.0.2-db1 (Scala 2.10)
Spark 2.0.2-db1 (Scala 2.11)
Spark 2.0.2-db2 (Scala 2.10)
Spark 2.0.2-db2 (Scala 2.11)
Spark 2.0.2-db3 (Scala 2.10)
Spark 2.0.2-db3 (Scala 2.11)
Spark 2.1 (Auto-updating, Scala 2.10)
Spark 2.1 (Auto-updating, Scala 2.11)
Spark 2.1.0-db1 (Scala 2.10)
Spark 2.1.0-db1 (Scala 2.11)
0 Workers, 0 GB Memory, 0 Cores and 1 Driver, 6 GB Memory, 0.88 Cores
terminate after an idle period of two hours.
icks subscription.
6 GB Memory, 0.88 Cores
6 GB Memory, 0.88 Cores

Overview

To access all the code examples in this stage, please import the Quick Start using Python or Quick Start using Scala notebooks.

This module allows you to quickly start using Apache Spark. We will be using Databricks so you can focus on the programming examples instead of spinning up and maintaining clusters and notebook infrastructure. As this is a quick start, we will be discussing the various concepts briefly so you can complete your end-to-end examples. In the "Additional Resources" section and other modules of this guide, you will have an opportunity to go deeper with the topic of your choice.

databricks Quick Start Using Python (Python) Import Notebook

Quick Start Using Python

- Using a Databricks notebook to showcase RDD operations using Python
- Reference http://spark.apache.org/docs/latest/quick-start.html

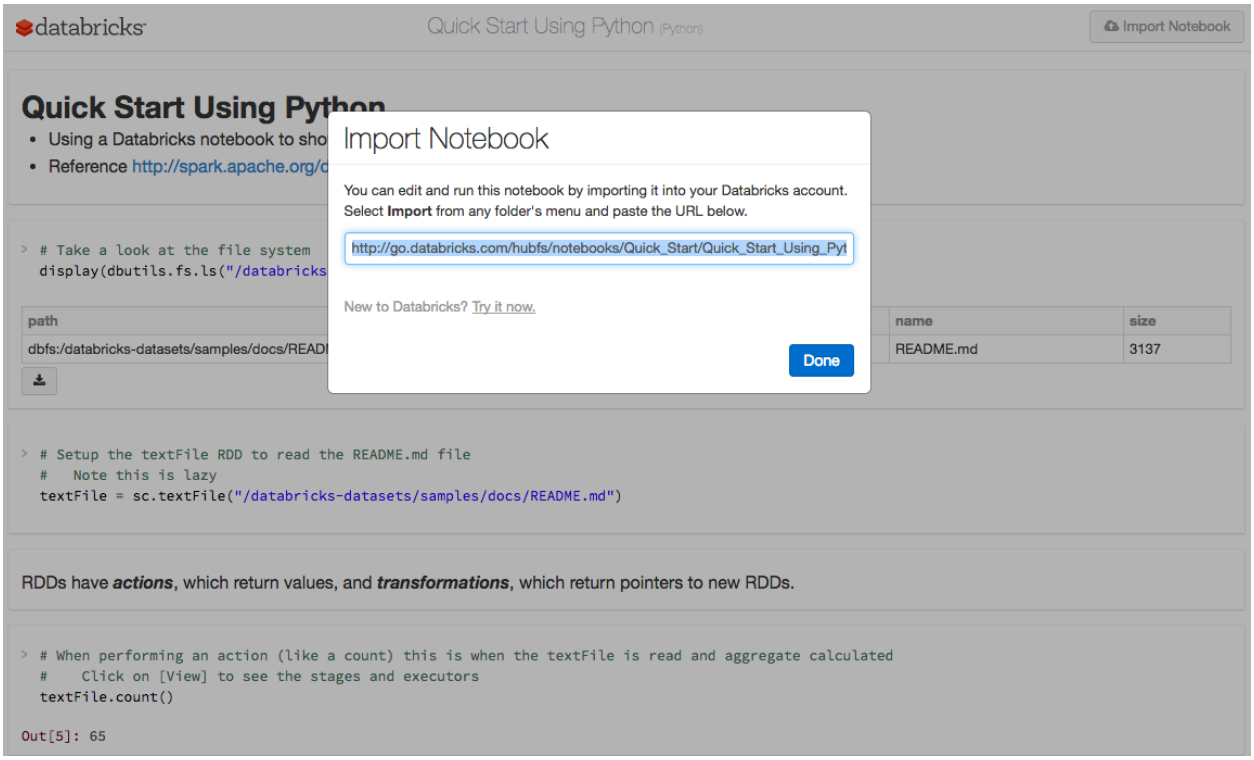
```
> # Take a look at the file system
  display(dbutils.fs.ls("/databricks-datasets/samples/docs/"))
```

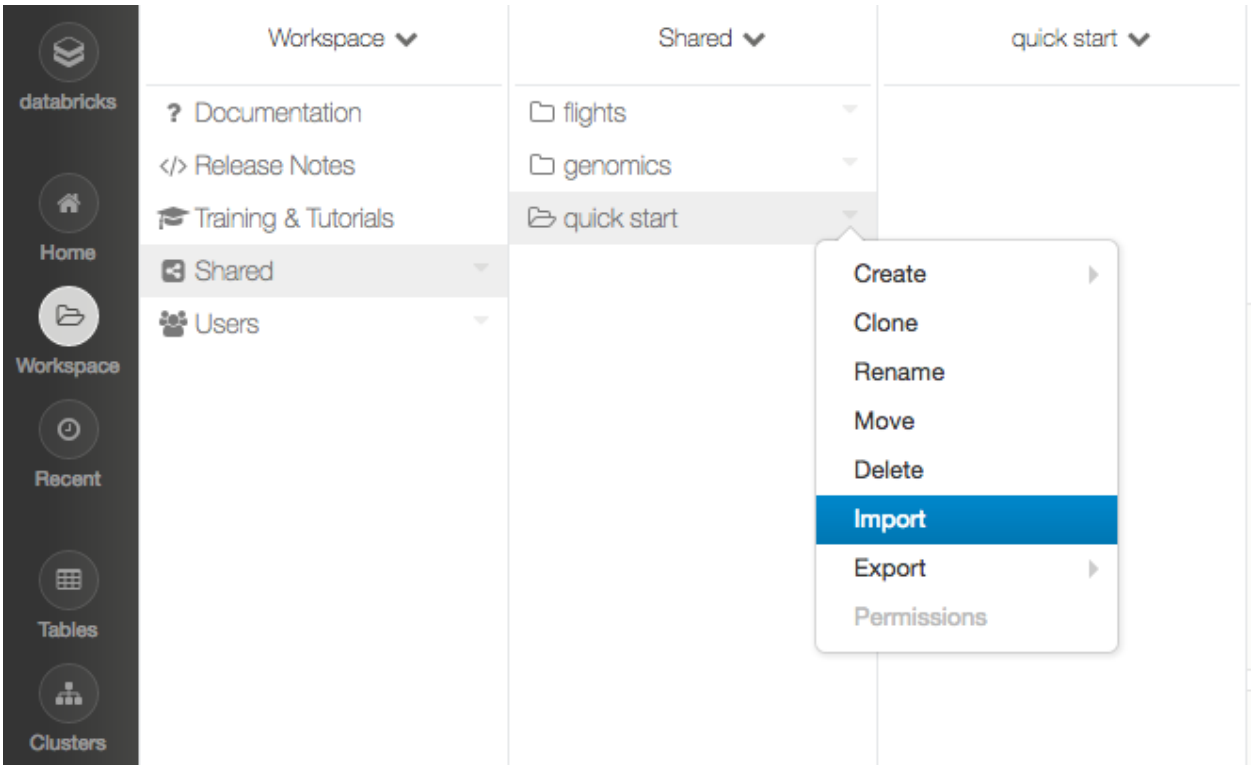
path	name	size
dbfs:/databricks-datasets/samples/docs/README.md	README.md	3137

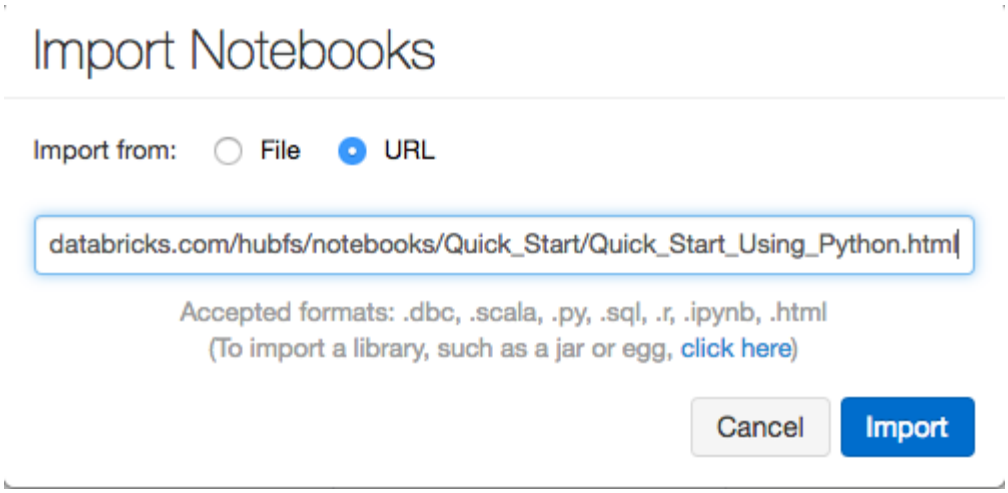
```
> # Setup the textFile RDD to read the README.md file
  #   Note this is lazy
  textFile = sc.textFile("/databricks-datasets/samples/docs/README.md")
```

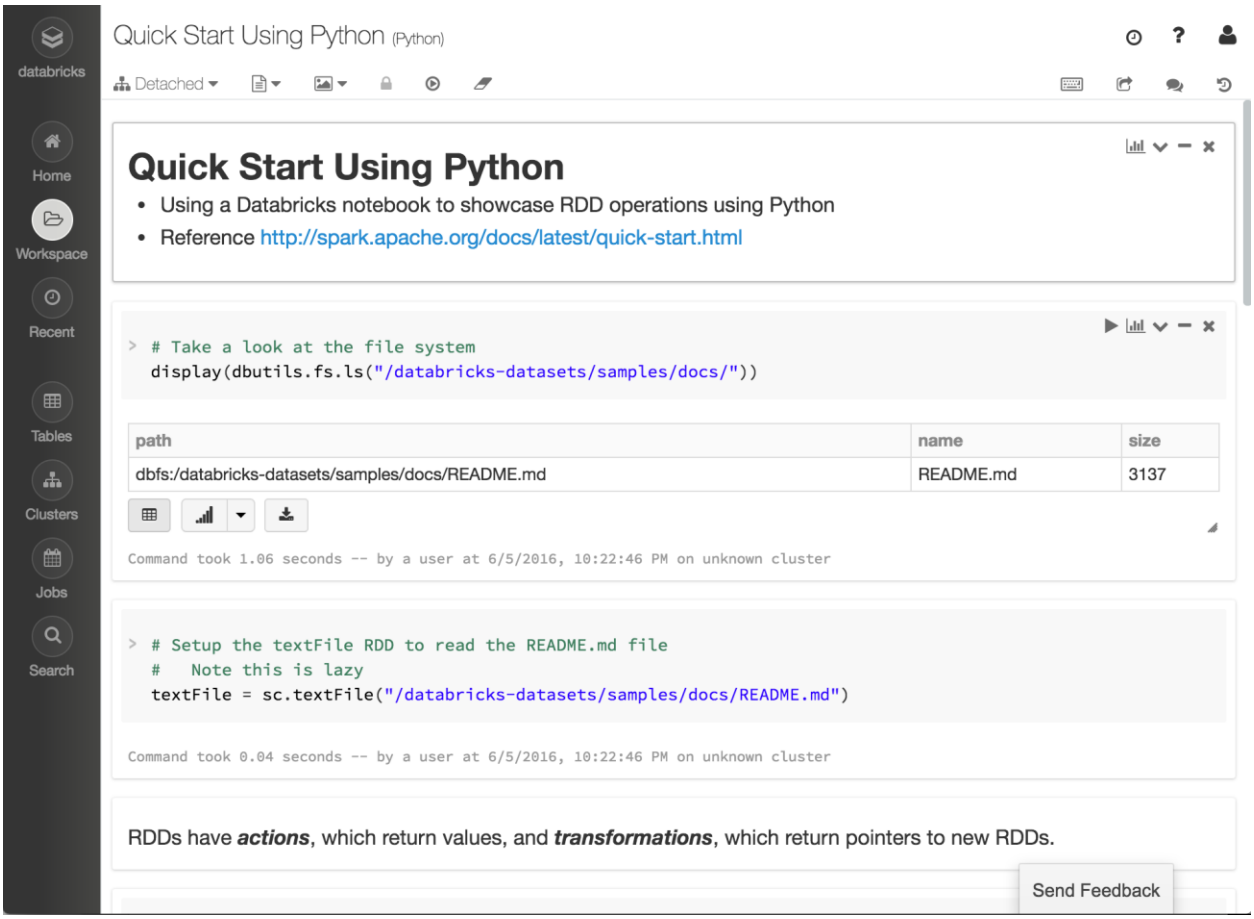
RDDs have ***actions***, which return values, and ***transformations***, which return pointers to new RDDs.

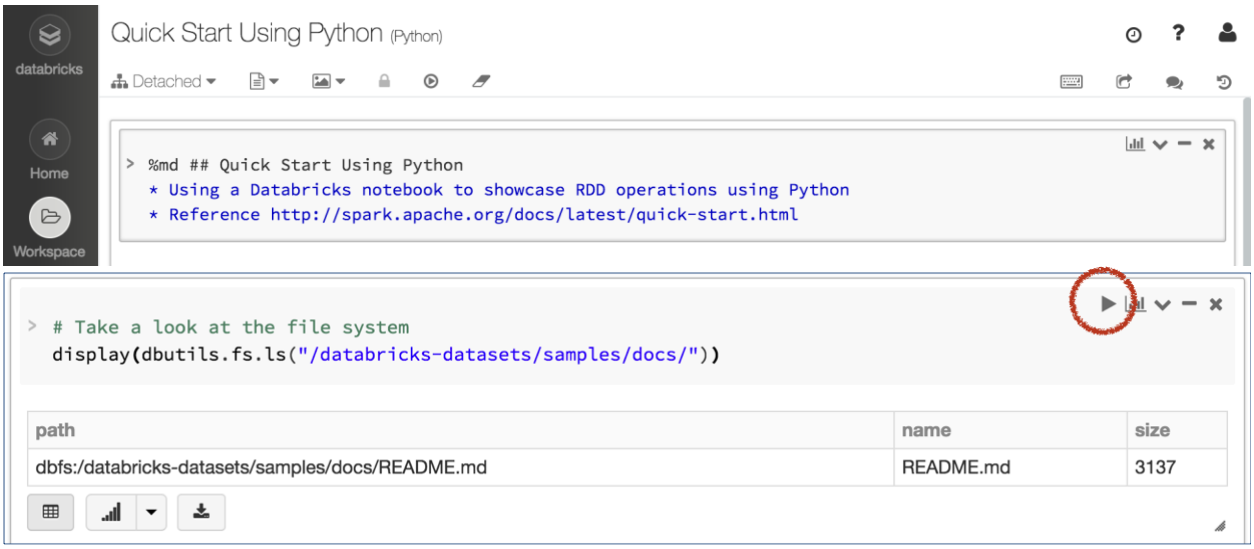
```
> # When performing an action (like a count) this is when the textFile is read and aggregate calculated
  #    Click on [View] to see the stages and executors
  textFile.count()

Out[5]: 65
```

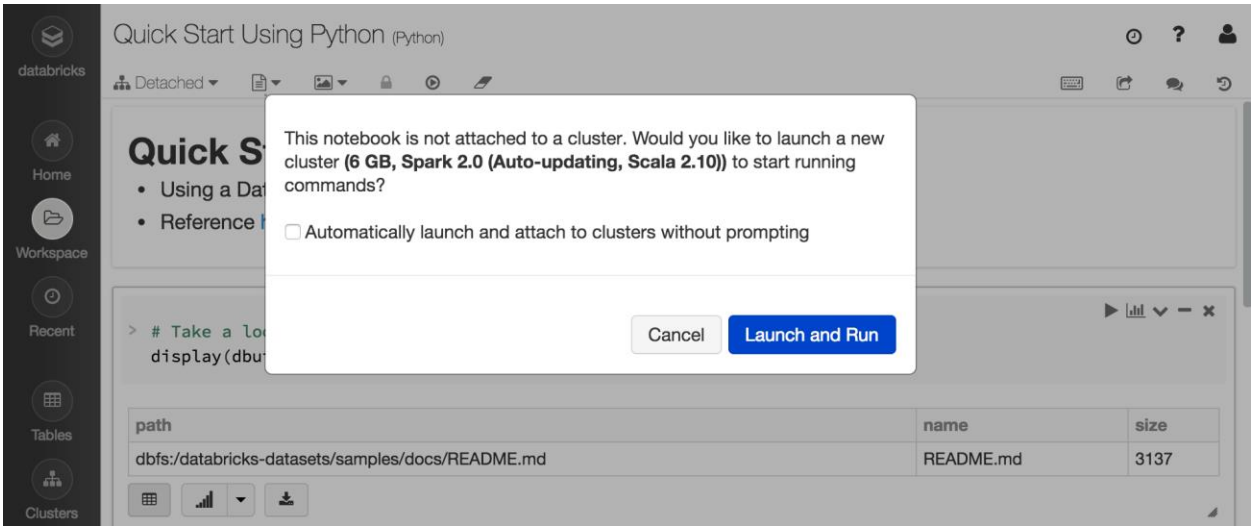

databricks
Quick Start Using Python (Python)
Import Notebook
Quick Start Using Python
Using a Databricks notebook to sho
Reference http://spark.apache.org/d
Import Notebook
You can edit and run this notebook by importing it into your Databricks account.
Select Import from any folder's menu and paste the URL below.
http://go.databricks.com/hubfs/notebooks/Quick_Start/Quick_Start_Using_Pyt
New to Databricks? Try it now.
Done
Take a look at the file system
display(dbutils.fs.ls("/databricks
path
name
size
dbfs:/databricks-datasets/samples/docs/READ
README.md
3137
# Setup the textFile RDD to read the README.md file
# Note this is lazy
textFile = sc.textFile("/databricks-datasets/samples/docs/README.md")
RDDs have actions, which return values, and transformations, which return pointers to new RDDs.
# When performing an action (like a count) this is when the textFile is read and aggregate calculated
# Click on [View] to see the stages and executors
textFile.count()
Out[5]: 65

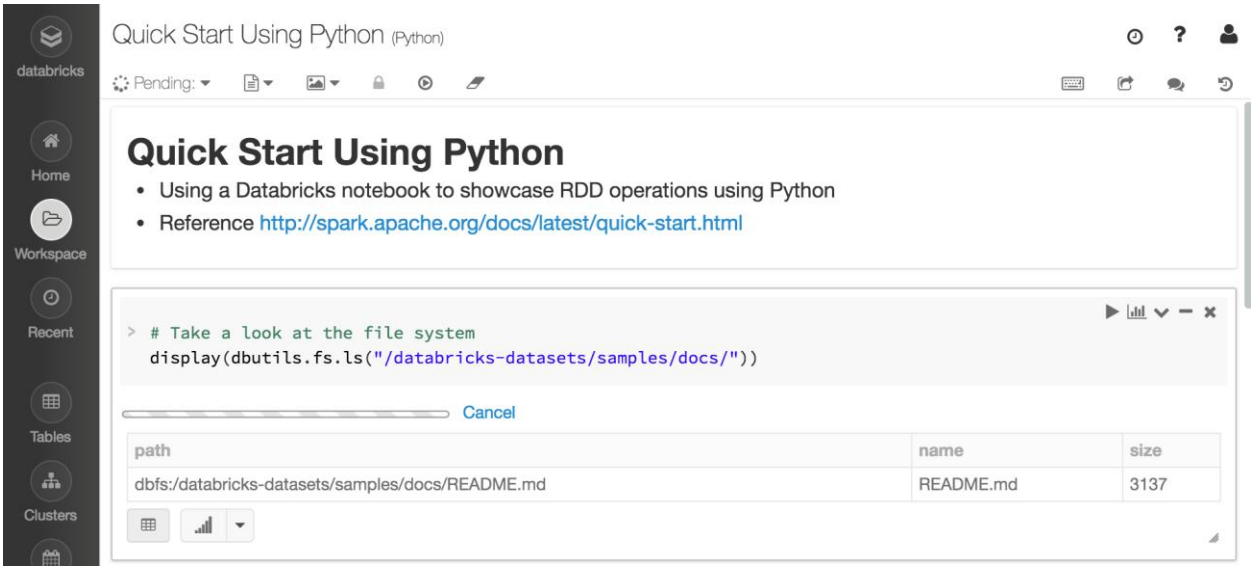

databricks
Home
Workspace
Recent
Tables
Clusters
Workspace
Shared
quick start
Documentation
Release Notes
Training & Tutorials
Shared
Users
flights
genomics
quick start
Create
Clone
Rename
Move
Delete
Import
Export
Permissions

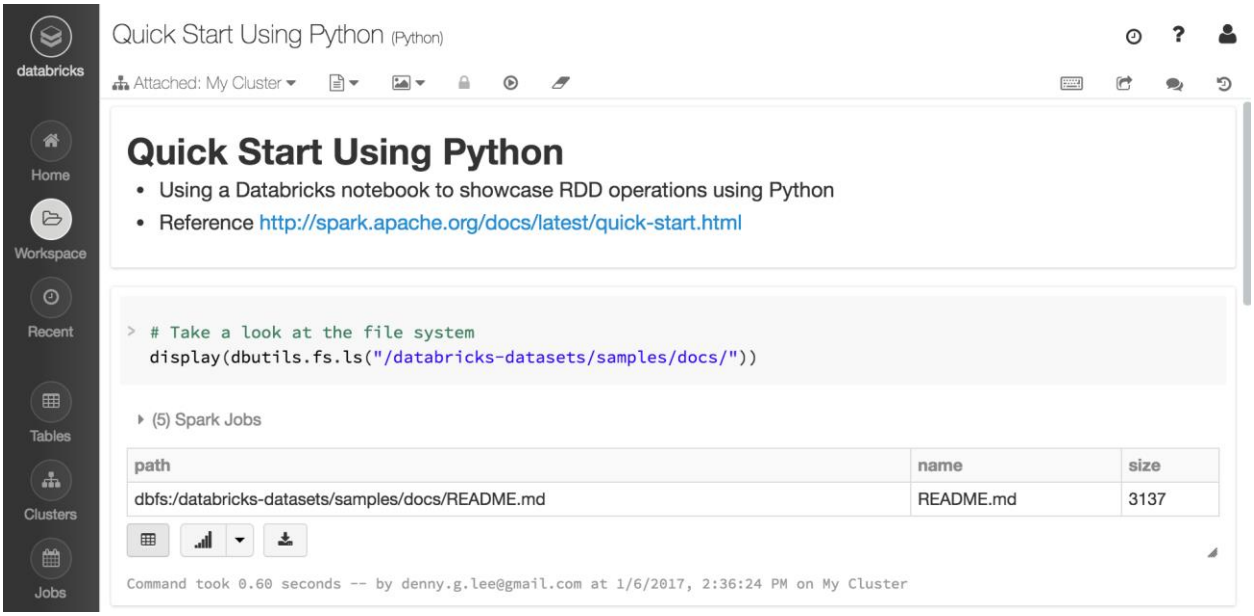

Import Notebooks
Import from:
File
URL
databricks.com/hubfs/notebooks/Quick_Start/Quick_Start_Using_Python.html
Accepted formats: .dbc, .scala, .py, .sql, .r, .ipynb, .html
(To import a library, such as a jar or egg, click here)
Cancel
Import

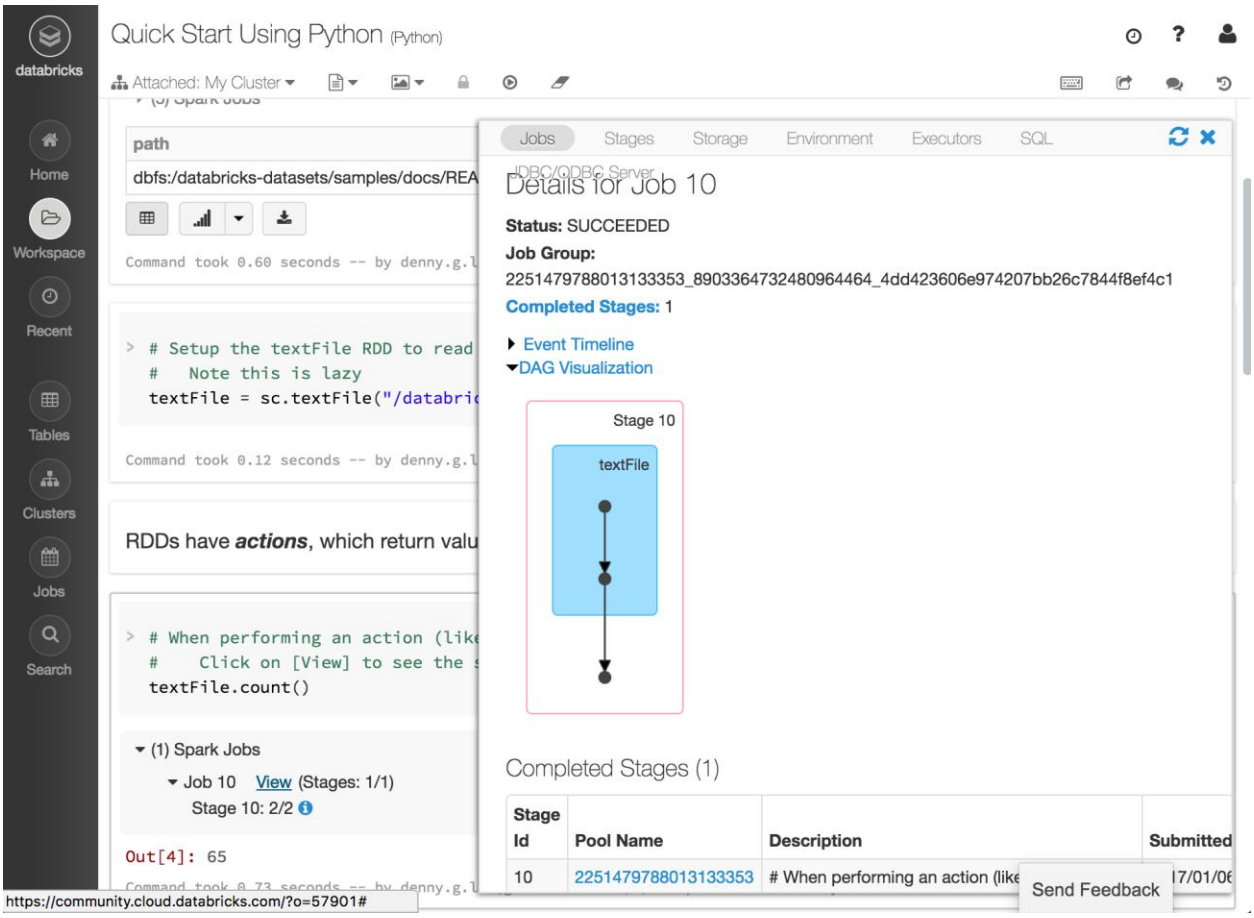

Quick Start Using Python (Python)
databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
Detached
Quick Start Using Python
Using a Databricks notebook to showcase RDD operations using Python
Reference http://spark.apache.org/docs/latest/quick-start.html
# Take a look at the file system
display(dbutils.fs.ls("/databricks-datasets/samples/docs/"))
path
name
size
dbfs:/databricks-datasets/samples/docs/README.md
README.md
3137
Command took 1.06 seconds -- by a user at 6/5/2016, 10:22:46 PM on unknown cluster
# Setup the textFile RDD to read the README.md file
# Note this is lazy
textFile = sc.textFile("/databricks-datasets/samples/docs/README.md")
Command took 0.04 seconds -- by a user at 6/5/2016, 10:22:46 PM on unknown cluster
RDDs have actions, which return values, and transformations, which return pointers to new RDDs.
Send Feedback


databricks
Quick Start Using Python (Python)
Detached
Home
Workspace
> %md ## Quick Start Using Python
* Using a Databricks notebook to showcase RDD operations using Python
* Reference http://spark.apache.org/docs/latest/quick-start.html
> # Take a look at the file system
display(dbutils.fs.ls("/databricks-datasets/samples/docs/"))
path
name
size
dbfs:/databricks-datasets/samples/docs/README.md
README.md
3137


databricks
Quick Start Using Python (Python)
Detached
Home
Workspace
Recent
Tables
Clusters
This notebook is not attached to a cluster. Would you like to launch a new cluster (6 GB, Spark 2.0 (Auto-updating, Scala 2.10)) to start running commands?
Automatically launch and attach to clusters without prompting
Cancel
Launch and Run
path
name
size
dbfs:/databricks-datasets/samples/docs/README.md
README.md
3137


databricks
Quick Start Using Python (Python)
Pending:
Home
Workspace
Recent
Tables
Clusters
Quick Start Using Python
Using a Databricks notebook to showcase RDD operations using Python
Reference http://spark.apache.org/docs/latest/quick-start.html
# Take a look at the file system
display(dbutils.fs.ls("/databricks-datasets/samples/docs/"))
Cancel
path
name
size
dbfs:/databricks-datasets/samples/docs/README.md
README.md
3137


databricks
Quick Start Using Python (Python)
Attached: My Cluster
Home
Workspace
Recent
Tables
Clusters
Jobs
Quick Start Using Python
Using a Databricks notebook to showcase RDD operations using Python
Reference http://spark.apache.org/docs/latest/quick-start.html
# Take a look at the file system
display(dbutils.fs.ls("/databricks-datasets/samples/docs/"))
(5) Spark Jobs
path
name
size
dbfs:/databricks-datasets/samples/docs/README.md
README.md
3137
Command took 0.60 seconds -- by denny.g.lee@gmail.com at 1/6/2017, 2:36:24 PM on My Cluster


databricks
Quick Start Using Python (Python)
Attached: My Cluster
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
path
dbfs:/databricks-datasets/samples/docs/REA
Command took 0.60 seconds -- by denny.g.l
# Setup the textFile RDD to read
# Note this is lazy
textFile = sc.textFile("/databri
Command took 0.12 seconds -- by denny.g.l
RDDs have ***actions***, which return valu
# When performing an action (like
# Click on [View] to see the s
textFile.count()
(1) Spark Jobs
Job 10 View (Stages: 1/1)
Stage 10: 2/2
Out[4]: 65
Command took 0.73 seconds -- by denny.g.l
https://community.cloud.databricks.com/?o=57901#
Jobs
Stages
Storage
Environment
Executors
SQL
JDBC/ODBC Server
Details for Job 10
Status: SUCCEEDED
Job Group:
2251479788013133353_8903364732480964464_4dd423606e974207bb26c7844f8ef4c1
Completed Stages: 1
Event Timeline
DAG Visualization
Stage 10
textFile
Completed Stages (1)
Stage Id
Pool Name
Description
Submitted
10
2251479788013133353
When performing an action (like
17/01/06
Send Feedback