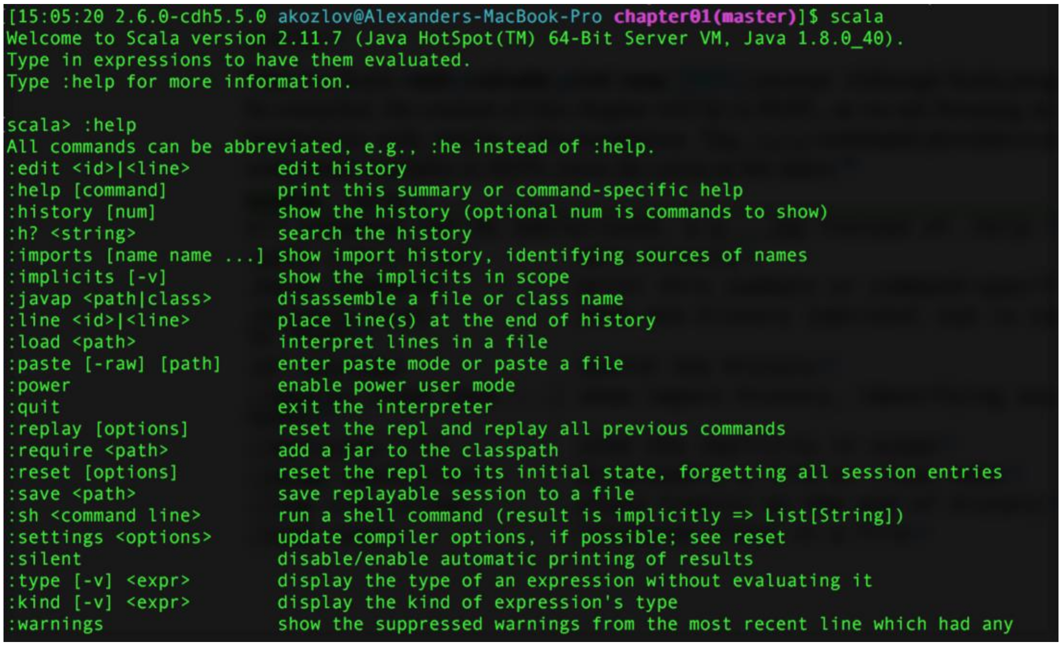
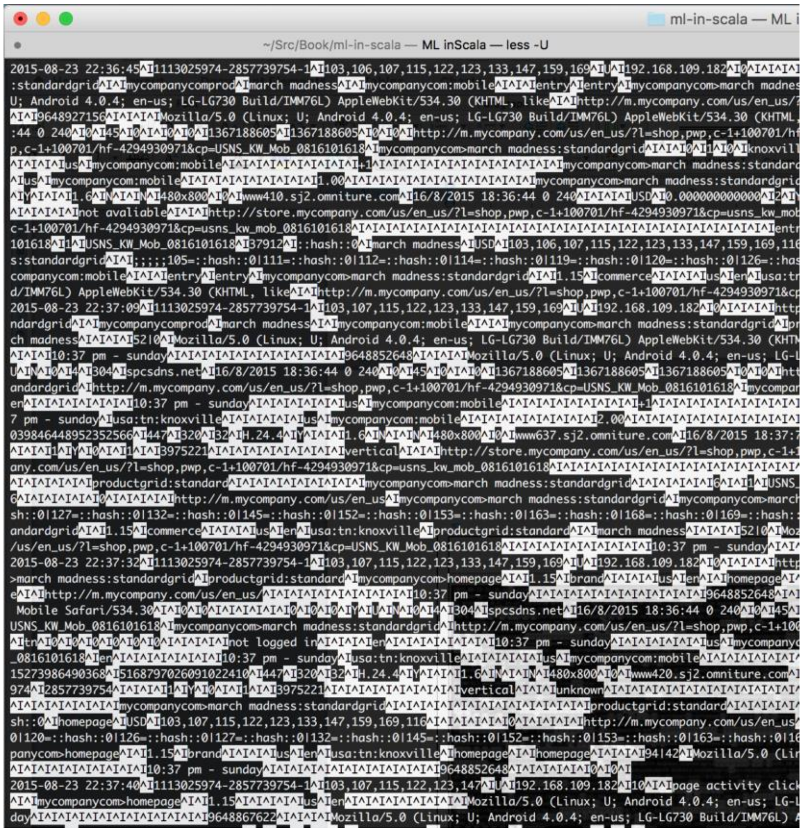
Chapter 1 - Exploratory Data Analysis

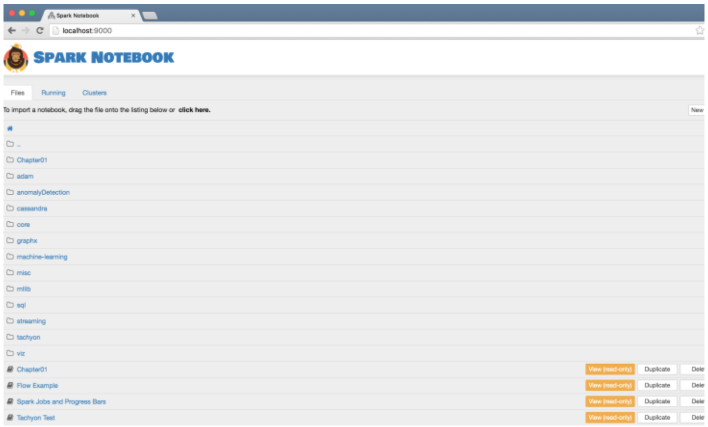


```
[15:05:20 2.6.0-cdh5.5.0 akozlov@Alexanders-MacBook-Pro chapter01(master)]$ scala
Welcome to Scala version 2.11.7 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_40).
Type in expressions to have them evaluated.
Type :help for more information.

scala> :help
All commands can be abbreviated, e.g., :he instead of :help.
:edit <id>|<line>        edit history
:help [command]          print this summary or command-specific help
:history [num]           show the history (optional num is commands to show)
:h? <string>             search the history
:imports [name name ...] show import history, identifying sources of names
:implicits [-v]          show the implicits in scope
:javap <path|class>      disassemble a file or class name
:line <id>|<line>        place line(s) at the end of history
:load <path>             interpret lines in a file
:paste [-raw] [path]     enter paste mode or paste a file
:power                   enable power user mode
:quit                    exit the interpreter
:replay [options]        reset the repl and replay all previous commands
:require <path>          add a jar to the classpath
:reset [options]         reset the repl to its initial state, forgetting all session entries
:save <path>             save replayable session to a file
:sh <command line>       run a shell command (result is implicitly => List[String])
:settings <options>      update compiler options, if possible; see reset
:silent                  disable/enable automatic printing of results
:type [-v] <expr>        display the type of an expression without evaluating it
:kind [-v] <expr>        display the kind of expression's type
:warnings                show the suppressed warnings from the most recent line which had any
```

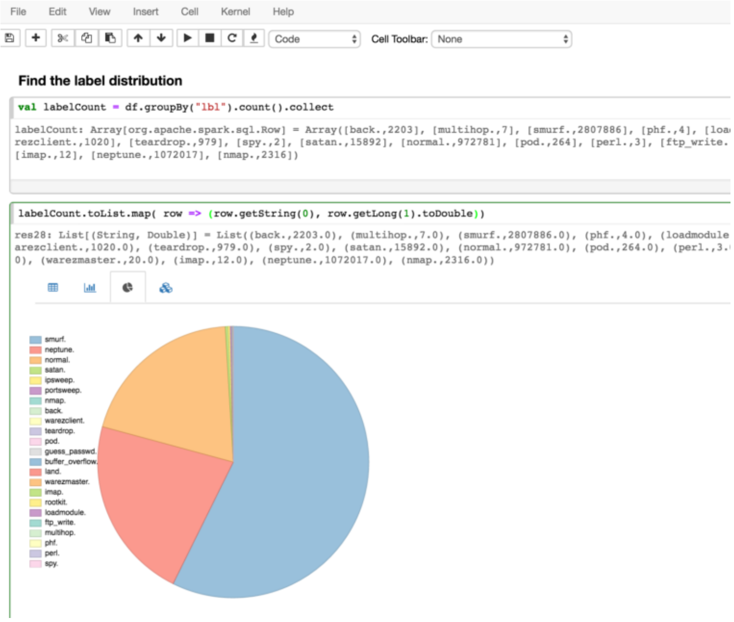


```
2015-08-23 22:36:45^I1113025974-2857739754-1^I103,106,107,115,122,123,133,147,159,169^IU^I192.168.109.182^I0^I^I^I
:standardgrid^I^Imycompanycomprod^Imarch madness^I^Imycompanycom:mobile^I^I^Ientry^Ientry^Imycompanycom>march madnes
U; Android 4.0.4; en-us; LG-LG730 Build/IMM76L) AppleWebKit/534.30 (KHTML, like^I^Ihttp://m.mycompany.com/us/en_us/?
^I^I9648927156^I^I^I^IMozilla/5.0 (Linux; U; Android 4.0.4; en-us; LG-LG730 Build/IMM76L) AppleWebKit/534.30 (KHTML,
:44 0 240^I0^I45^I0^I^I0^I0^I1367188605^I1367188605^I0^I0^Ihttp://m.mycompany.com/us/en_us/?l=shop,pwp,c-1+100701/hf
p,c-1+100701/hf-4294930971&cp=USNS_KW_Mob_0816101618^Imycompanycom>march madness:standardgrid^I^I^I0^I1^I0^Iknoxvill
^I^I^I^Ius^Imycompanycom:mobile^I^I^I^I^I^I^I^I^I^I+1^I^I^I^I^I^I^I^I^I^I^I^I^I^Imycompanycom>march madness:standard
^Ius^Imycompanycom:mobile^I^I^I^I^I^I^I^I^I^I1.00^I^I^I^I^I^I^I^I^I^I^I^I^I^Imycompanycom>march madness:standardgrid
^IY^I^I^I1.6^IN^I^IN^I480x800^I0^Iwww410.sj2.omniture.com^I16/8/2015 18:36:44 0 240^I^I^I^IUSD^I0.000000000000^I2^IY
^I^I^I^I^Inot avaliable^I^I^Ihttp://store.mycompany.com/us/en_us/?l=shop,pwp,c-1+100701/hf-4294930971&cp=usns_kw_mob
c-1+100701/hf-4294930971&cp=usns_kw_mob_0816101618^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^Ientr
101618^I1^IUSNS_KW_Mob_0816101618^I37912^I::hash::0^Imarch madness^IUSD^I103,106,107,115,122,123,133,147,159,169,116
s:standardgrid^I^I;;;;;105=::hash::0|111=::hash::0|112=::hash::0|114=::hash::0|119=::hash::0|120=::hash::0|126=::has
companycom:mobile^I^I^Ientry^Ientry^Imycompanycom>march madness:standardgrid^I^I1.15^Icommerce^I^I^I^Ius^Ien^Iusa:tr
d/IMM76L) AppleWebKit/534.30 (KHTML, like^I^Ihttp://m.mycompany.com/us/en_us/?l=shop,pwp,c-1+100701/hf-4294930971&cp
2015-08-23 22:37:09^I1113025974-2857739754-1^I103,107,115,122,123,133,147,159,169^IU^I192.168.109.182^I0^I^I^I^Ihttp
ndardgrid^I^Imycompanycomprod^Imarch madness^I^Imycompanycom:mobile^I^I^Imycompanycom>march madness:standardgrid^Ipr
ch madness^I^I^I^I52|0^IMozilla/5.0 (Linux; U; Android 4.0.4; en-us; LG-LG730 Build/IMM76L) AppleWebKit/534.30 (KHTM
^I^I^I10:37 pm - sunday^I^I^I^I^I^I^I^I^I^I^I^I^I9648852648^I^I^I^IMozilla/5.0 (Linux; U; Android 4.0.4; en-us; LG-L
U^IN^I0^I4^I304^Ispcsdns.net^I16/8/2015 18:36:44 0 240^I0^I45^I0^I^I0^I1367188605^I1367188605^I1367188605^I0^I0^Ihtt
andardgrid^Ihttp://m.mycompany.com/us/en_us/?l=shop,pwp,c-1+100701/hf-4294930971&cp=USNS_KW_Mob_0816101618^Imycompan
en^I^I^I^I^I^I^I^I10:37 pm - sunday^I^I^I^I^I^I^Ius^Imycompanycom:mobile^I^I^I^I^I^I^I^I^I^I+1^I^I^I^I^I^I^I^I^I^I^I
7 pm - sunday^Iusa:tn:knoxville^I^I^I^I^Ius^Imycompanycom:mobile^I^I^I^I^I^I^I^I^I^I2.00^I^I^I^I^I^I^I^I^I^I^I^I^I
039846448952352566^I447^I320^I32^IH.24.4^IY^I^I^I1.6^IN^I^IN^I480x800^I0^Iwww637.sj2.omniture.com^I16/8/2015 18:37:7
^I^I^I1^IY^I0^I^I1^I^I3975221^I^I^I^I^I^I^I^I^I^Ivertical^I^I^Ihttp://store.mycompany.com/us/en_us/?l=shop,pwp,c-1+1
any.com/us/en_us/?l=shop,pwp,c-1+100701/hf-4294930971&cp=usns_kw_mob_0816101618^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I^I
^I^I^I^I^I^Iproductgrid:standard^I^I^I^I^I^I^I^I^I^Imycompanycom>march madness:standardgrid^I^I^I^I^I6^I^I1^IUSNS_
6^I^I^I^I^I0^I^I^I^I^Ihttp://m.mycompany.com/us/en_us^Imycompanycom>march madness:standardgrid^Imycompanycom>march
sh::0|127=::hash::0|132=::hash::0|145=::hash::0|152=::hash::0|153=::hash::0|163=::hash::0|168=::hash::0|169=::hash::
andardgrid^I^I1.15^Icommerce^I^I^I^Ius^Ien^Iusa:tn:knoxville^Iproductgrid:standard^I^Imarch madness^I^I^I^I52|0^IMoz
/us/en_us/?l=shop,pwp,c-1+100701/hf-4294930971&cp=USNS_KW_Mob_0816101618^I^I^I^I^I^I^I^I^I^I^I10:37 pm - sunday^I^I^
2015-08-23 22:37:32^I1113025974-2857739754-1^I103,107,115,122,123,133,147,159,169^IU^I192.168.109.182^I0^I^I^I^Ihttp
>march madness:standardgrid^Iproductgrid:standard^Imycompanycom>homepage^I^I1.15^Ibrand^I^I^I^Ius^Ien^I^Ihomepage^I^
e^I^Ihttp://m.mycompany.com/us/en_us/^I^I^I^I^I^I^I^I^I^I^I10:37 pm - sunday^I^I^I^I^I^I^I^I^I^I^I^I^I9648852648^I^I
 Mobile Safari/534.30^I^I0^I0^I^I^I^I^I0^I0^I0^IY^IU^IN^I0^I4^I304^Ispcsdns.net^I16/8/2015 18:36:44 0 240^I0^I45^I
USNS_KW_Mob_0816101618^Imycompanycom>march madness:standardgrid^Ihttp://m.mycompany.com/us/en_us/?l=shop,pwp,c-1+100
^Itn^I0^I0^I0^I0^I0^I0^I^I^I^Inot logged in^I^I^Ien^I^I^I^I^I^I^I10:37 pm - sunday^I^I^I^I^I^Ius^Imycompanyc
_0816101618^Ien^I^I^I^I^I^I^I10:37 pm - sunday^Iusa:tn:knoxville^I^I^I^I^Ius^Imycompanycom:mobile^I^I^I^I^I^I^
15273986490368^I5168797026091022410^I447^I320^I32^IH.24.4^IY^I^I^I1.6^IN^I^IN^I480x800^I0^Iwww420.sj2.omniture.com^I
974^I2857739754^I^I^I^I1^IY^I0^I^I1^I^I3975221^I^I^I^I^I^I^I^Ivertical^I^I^Iunknown^I^I^I^I^I^I^I^I^I^I^I^I^I^I
^I^I^I^I^I^I^Imycompanycom>march madness:standardgrid^I^I^I^I^I^I^I^I^I^I^I^I^I^I^Iproductgrid:standard^I^I^I^I^
sh::0^Ihomepage^IUSD^I103,107,115,122,123,133,147,159,169,116^I^I^I^I^I^I0^I^I^I^I^Ihttp://m.mycompany.com/us/en_us/
0|120=::hash::0|126=::hash::0|127=::hash::0|132=::hash::0|145=::hash::0|152=::hash::0|153=::hash::0|163=::hash::0|16
panycom>homepage^I^I1.15^Ibrand^I^I^I^Ius^Ien^Iusa:tn:knoxville^Ihomepage^I^Ihomepage^I^I^I^I94|42^IMozilla/5.0 (Lin
^I^I^I^I^I^I^I^I^I^I10:37 pm - sunday^I^I^I^I^I^I^I^I^I^I^I^I^I9648852648^I^I^I^I^I^I0^I0^I
2015-08-23 22:37:40^I1113025974-2857739754-1^I103,107,115,122,123,147^IU^I192.168.109.182^I10^I^Ipage activity click
^I^Imycompanycom>homepage^I^I1.15^I^I^I^I^Ius^Ien^I^I^I^I^I^I^I^I^IMozilla/5.0 (Linux; U; Android 4.0.4; en-us; LG-L
day^I^I^I^I^I^I^I^I^I^I^I^I^I^I9648867622^I^I^I^IMozilla/5.0 (Linux; U; Android 4.0.4; en-us; LG-LG730 Build/IMM76L) A
```



To import a notebook, drag the file onto the listing below or **click here.**

..
Chapter01
adam
anomalyDetection
cassandra
core
graphx
machine-learning
misc
mllib
sql
streaming
tachyon
viz
Chapter01
Flow Example
Spark Jobs and Progress Bars
Tachyon Test



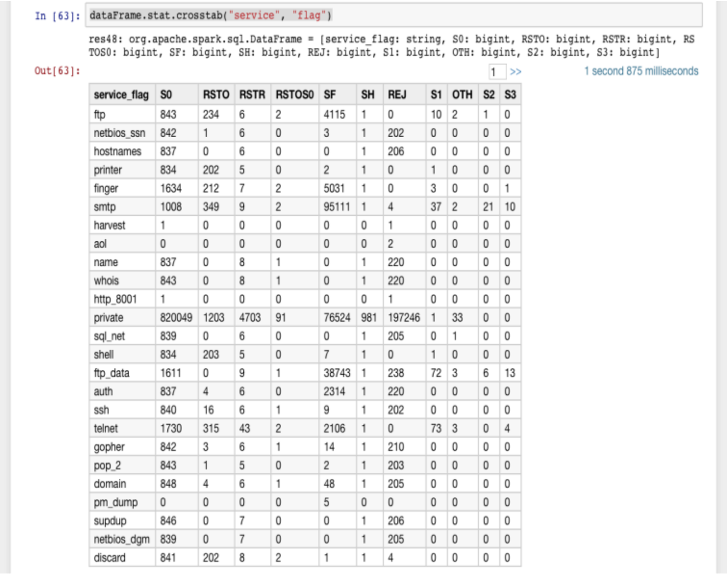
Find the label distribution

```
val labelCount = df.groupBy("lbl").count().collect
```

```
labelCount: Array[org.apache.spark.sql.Row] = Array([back.,2203], [multihop.,7], [smurf.,2807886], [phf.,4], [loa
rezclient.,1020], [teardrop.,979], [spy.,2], [satan.,15892], [normal.,972781], [pod.,264], [perl.,3], [ftp_write.
[imap.,12], [neptune.,1072017], [nmap.,2316])
```

```
labelCount.toList.map( row => (row.getString(0), row.getLong(1).toDouble))
```

```
res28: List[(String, Double)] = List((back.,2203.0), (multihop.,7.0), (smurf.,2807886.0), (phf.,4.0), (loadmodule
arezclient.,1020.0), (teardrop.,979.0), (spy.,2.0), (satan.,15892.0), (normal.,972781.0), (pod.,264.0), (perl.,3.
0), (warezmaster.,20.0), (imap.,12.0), (neptune.,1072017.0), (nmap.,2316.0))
```

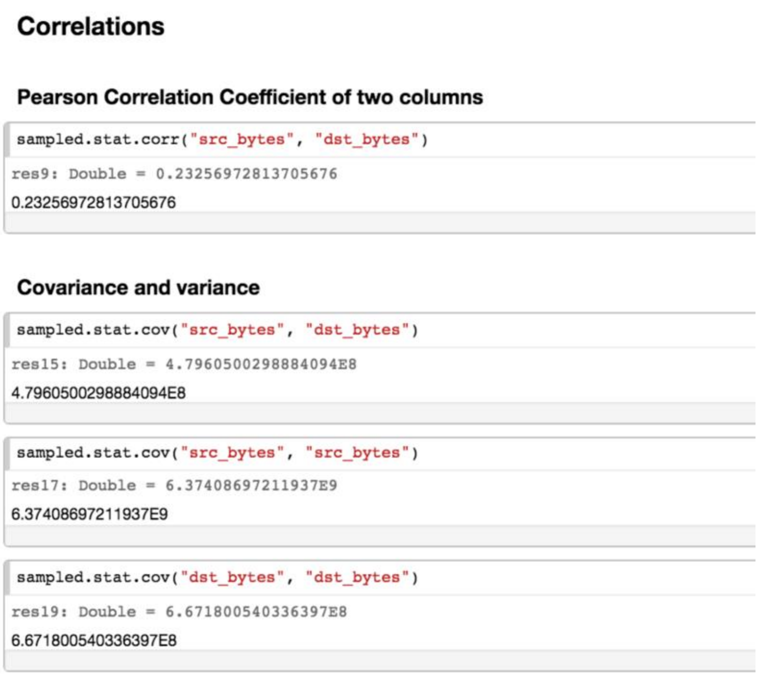


```
In [63]: dataFrame.stat.crosstab("service", "flag")
```

```
res48: org.apache.spark.sql.DataFrame = [service_flag: string, S0: bigint, RSTO: bigint, RSTR: bigint, RS
TOS0: bigint, SF: bigint, SH: bigint, REJ: bigint, S1: bigint, OTH: bigint, S2: bigint, S3: bigint]
```

Out[63]:

service_flag	S0	RSTO	RSTR	RSTOS0	SF	SH	REJ	S1	OTH	S2	S3
ftp	843	234	6	2	4115	1	0	10	2	1	0
netbios_ssn	842	1	6	0	3	1	202	0	0	0	0
hostnames	837	0	6	0	0	1	206	0	0	0	0
printer	834	202	5	0	2	1	0	1	0	0	0
finger	1634	212	7	2	5031	1	0	3	0	0	1
smtp	1008	349	9	2	95111	1	4	37	2	21	10
harvest	1	0	0	0	0	0	1	0	0	0	0
aol	0	0	0	0	0	0	2	0	0	0	0
name	837	0	8	1	0	1	220	0	0	0	0
whois	843	0	8	1	0	1	220	0	0	0	0
http_8001	1	0	0	0	0	0	1	0	0	0	0
private	820049	1203	4703	91	76524	981	197246	1	33	0	0
sql_net	839	0	6	0	0	1	205	0	1	0	0
shell	834	203	5	0	7	1	0	1	0	0	0
ftp_data	1611	0	9	1	38743	1	238	72	3	6	13
auth	837	4	6	0	2314	1	220	0	0	0	0
ssh	840	16	6	1	9	1	202	0	0	0	0
telnet	1730	315	43	2	2106	1	0	73	3	0	4
gopher	842	3	6	1	14	1	210	0	0	0	0
pop_2	843	1	5	0	2	1	203	0	0	0	0
domain	848	4	6	1	48	1	205	0	0	0	0
pm_dump	0	0	0	0	5	0	0	0	0	0	0
supdup	846	0	7	0	0	1	206	0	0	0	0
netbios_dgm	839	0	7	0	0	1	205	0	0	0	0
discard	841	202	8	2	1	1	4	0	0	0	0



Correlations

Pearson Correlation Coefficient of two columns

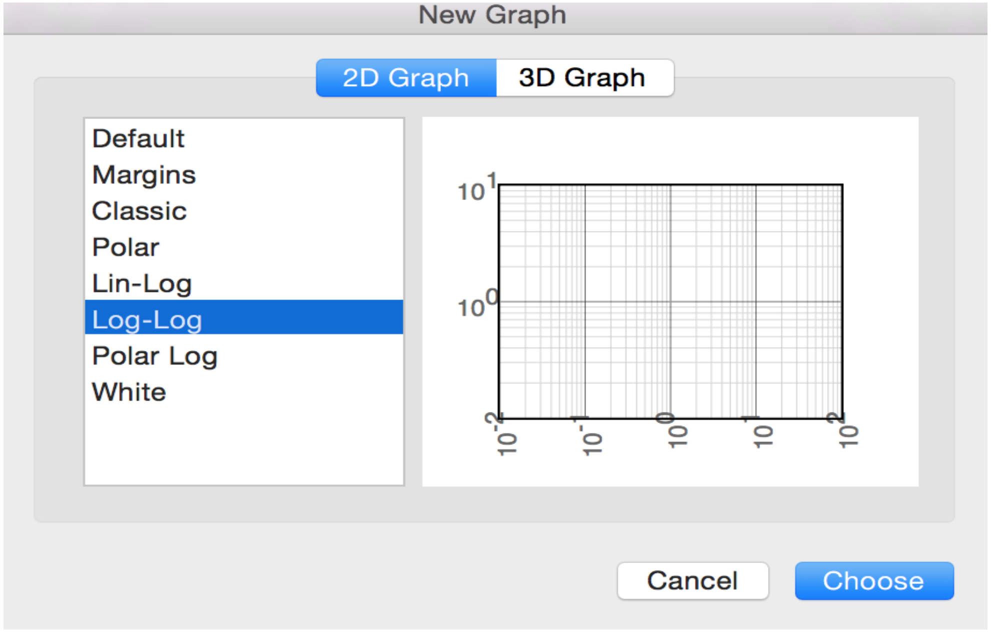
```
sampled.stat.corr("src_bytes", "dst_bytes")
res9: Double = 0.23256972813705676
0.23256972813705676
```

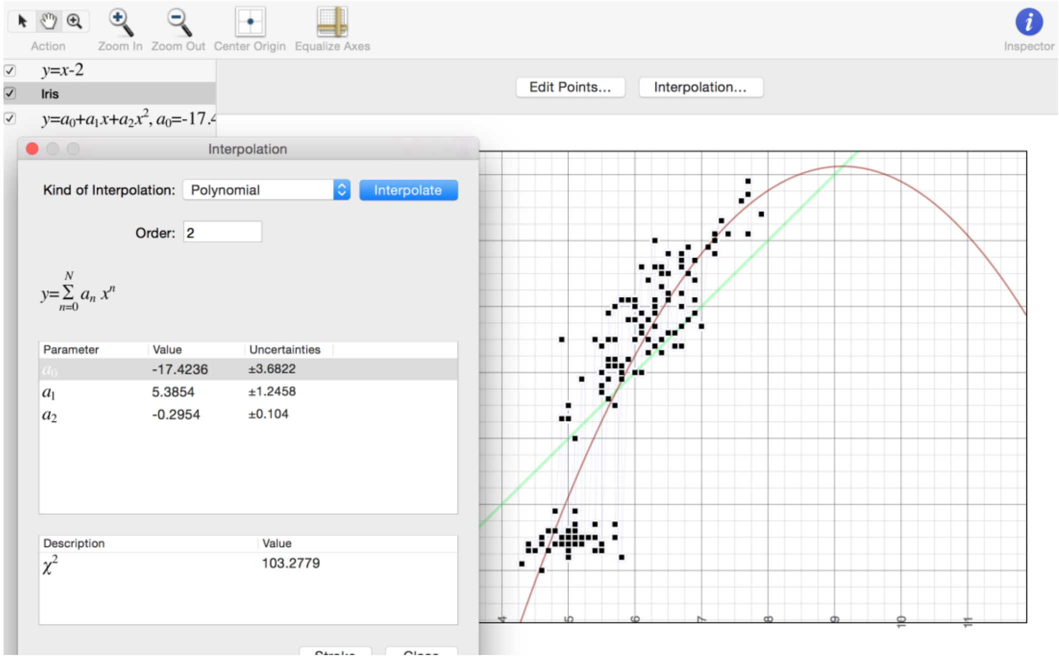
Covariance and variance

```
sampled.stat.cov("src_bytes", "dst_bytes")
res15: Double = 4.7960500298884094E8
4.7960500298884094E8
```

```
sampled.stat.cov("src_bytes", "src_bytes")
res17: Double = 6.37408697211937E9
6.37408697211937E9
```

```
sampled.stat.cov("dst_bytes", "dst_bytes")
res19: Double = 6.671800540336397E8
6.671800540336397E8
```





☑ $y=x-2$

☑ Iris

☑ $y=a_0+a_1x+a_2x^2, a_0=-17.4$

Kind of Interpolation: Polynomial

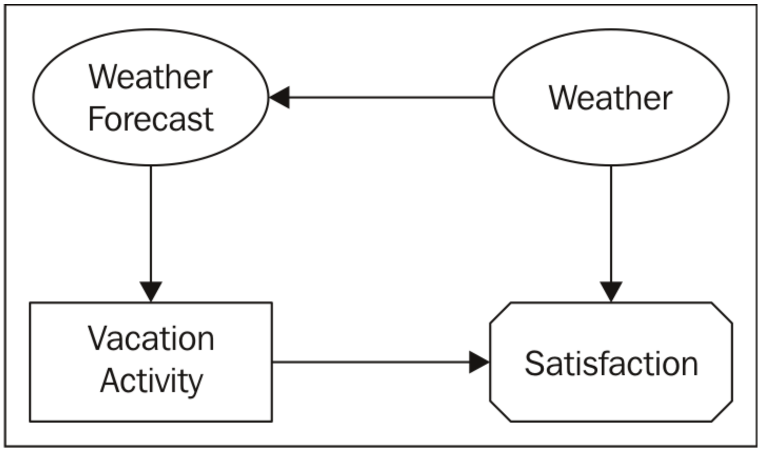
Order: 2

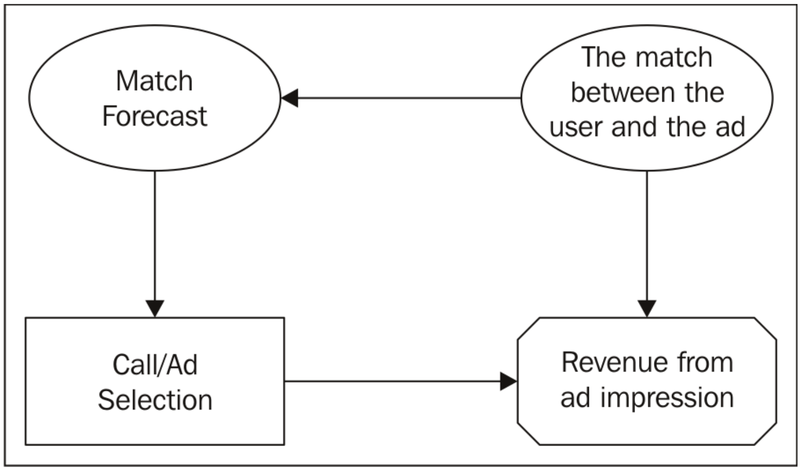
$$y=\sum_{n=0}^{N} a_n\, x^n$$

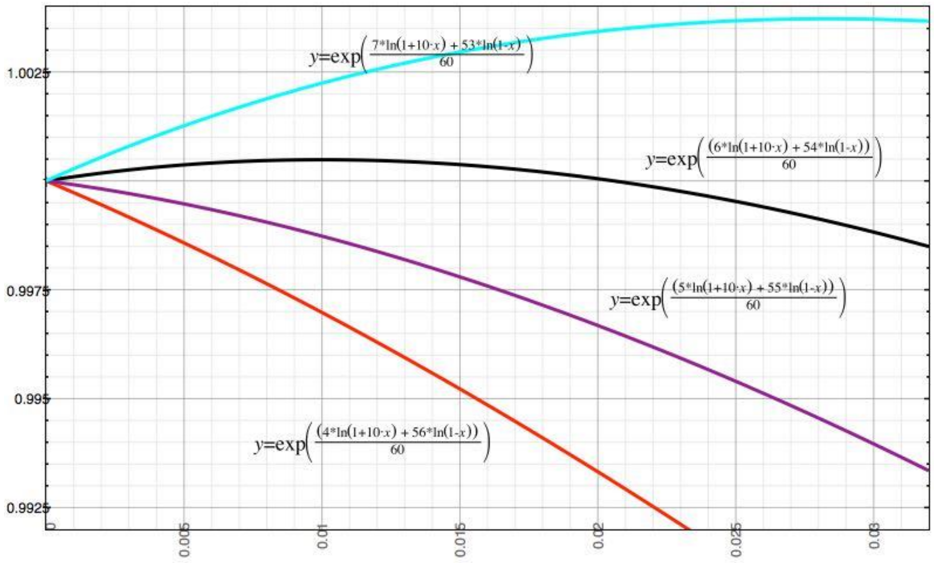
Parameter	Value	Uncertainties
a_0	-17.4236	±3.6822
a_1	5.3854	±1.2458
a_2	-0.2954	±0.104

Description	Value
χ^2	103.2779

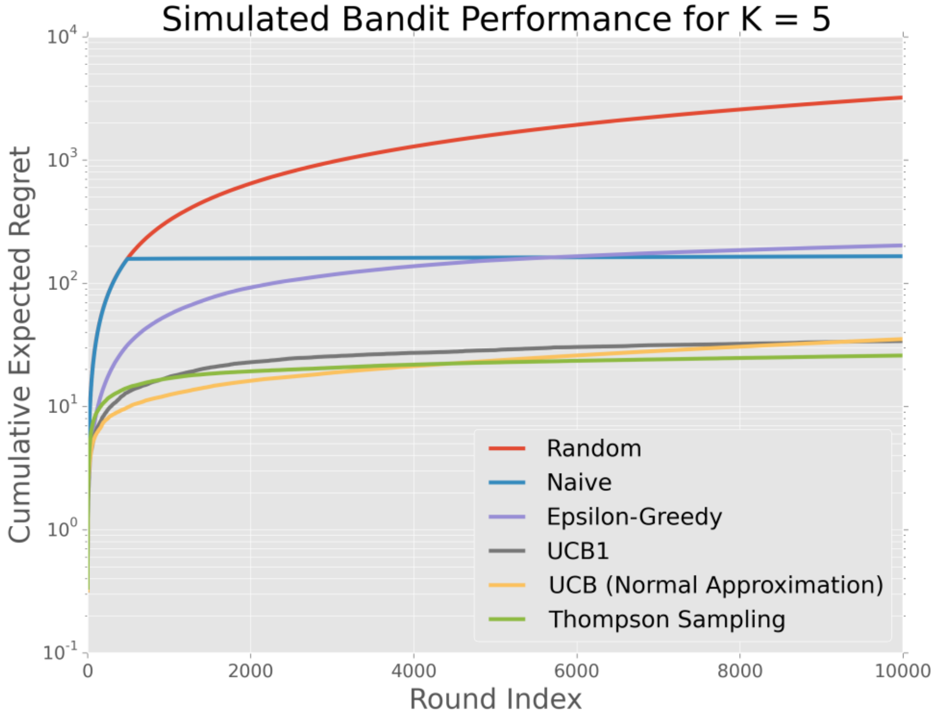
Chapter 2 - Data Pipelines and Modeling



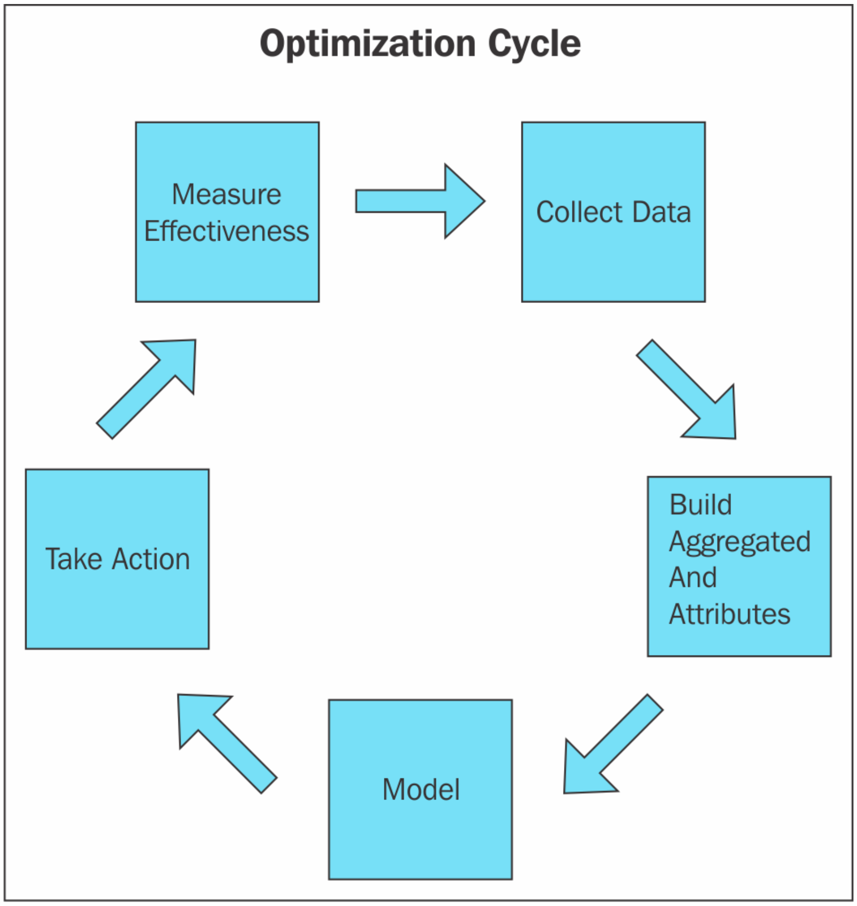

Match Forecast
The match between the user and the ad
Call/Ad Selection
Revenue from ad impression



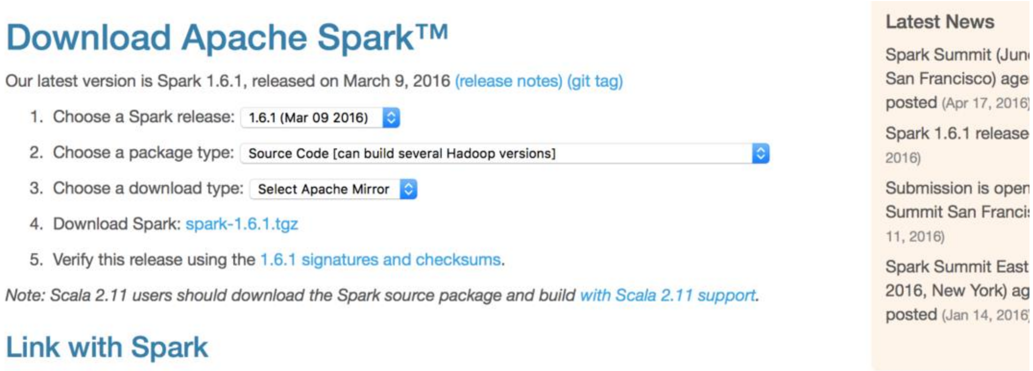
y=exp((7*ln(1+10·x) + 53*ln(1-x))/60)
y=exp(((6*ln(1+10·x) + 54*ln(1-x)))/60)
y=exp(((5*ln(1+10·x) + 55*ln(1-x)))/60)
y=exp(((4*ln(1+10·x) + 56*ln(1-x)))/60)
1.0025
0.9975
0.995
0.9925
0
0.005
0.01
0.015
0.02
0.025
0.03



Simulated Bandit Performance for K = 5
Cumulative Expected Regret
Round Index
10^4
10^3
10^2
10^1
10^0
10^{-1}
0
2000
4000
6000
8000
10000
Random
Naive
Epsilon-Greedy
UCB1
UCB (Normal Approximation)
Thompson Sampling


Optimization Cycle
Measure Effectiveness
Collect Data
Build Aggregated And Attributes
Model
Take Action

Chapter 3 - Working with Spark and MLlib



Download Apache Spark™

Our latest version is Spark 1.6.1, released on March 9, 2016 (release notes) (git tag)

1. Choose a Spark release: 1.6.1 (Mar 09 2016)
2. Choose a package type: Source Code [can build several Hadoop versions]
3. Choose a download type: Select Apache Mirror
4. Download Spark: spark-1.6.1.tgz
5. Verify this release using the 1.6.1 signatures and checksums.

Note: Scala 2.11 users should download the Spark source package and build with Scala 2.11 support.

Link with Spark

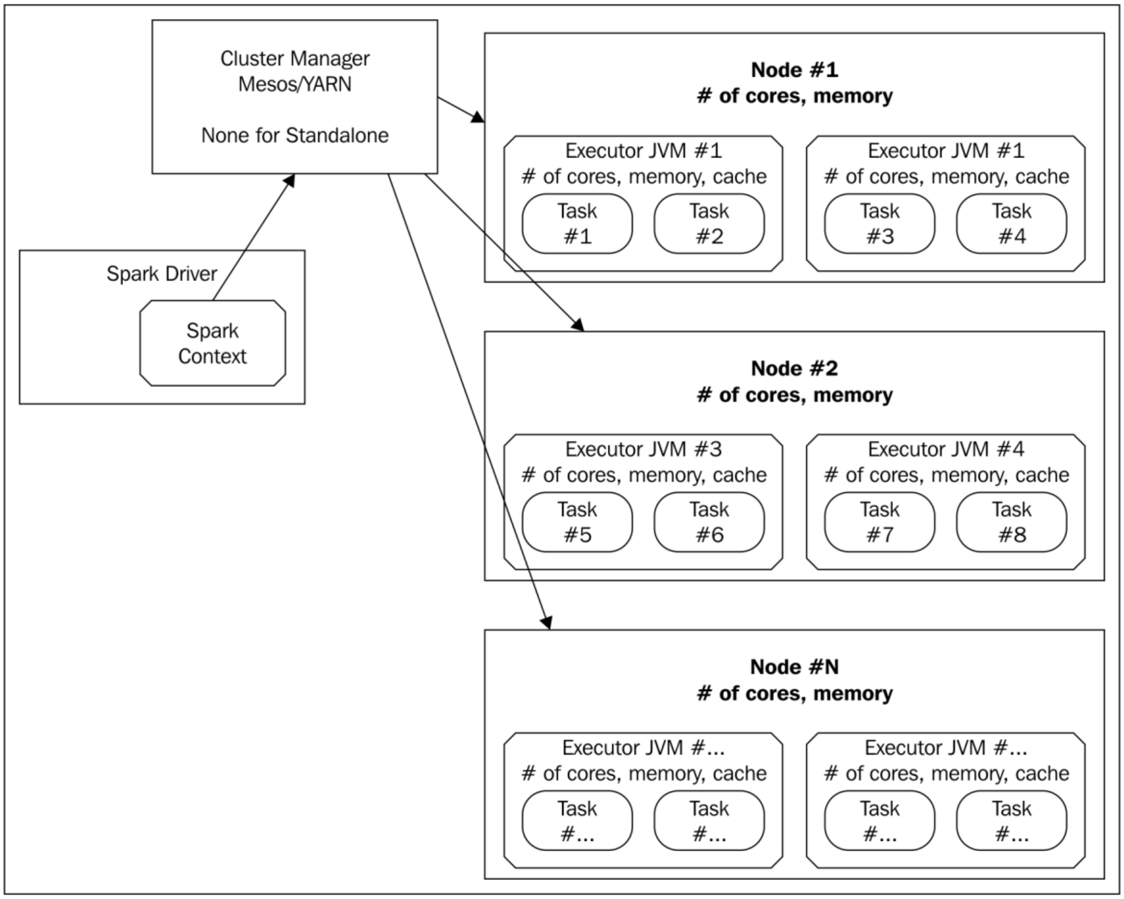
Latest News

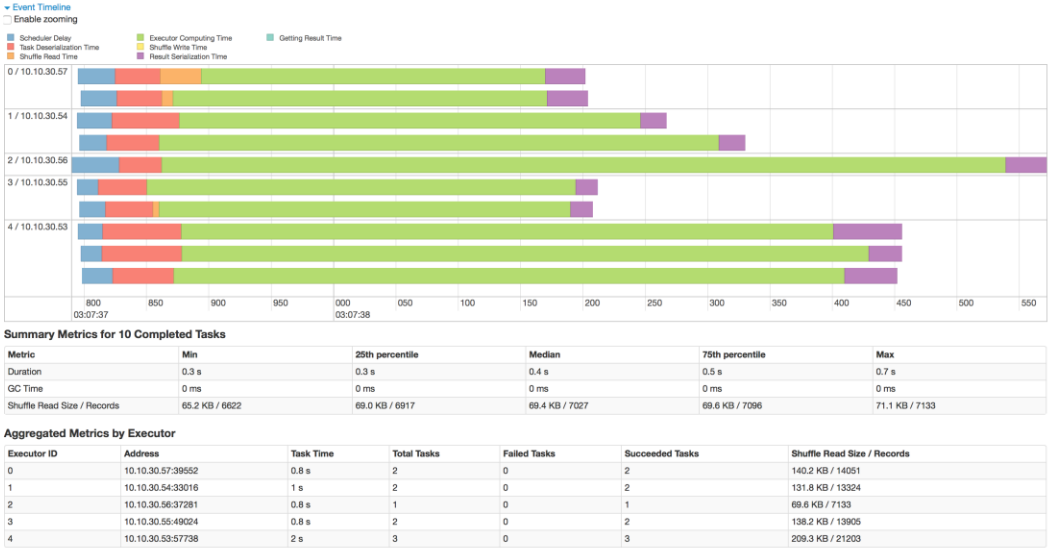
Spark Summit (Jun
San Francisco) age
posted (Apr 17, 2016

Spark 1.6.1 release
2016)

Submission is open
Summit San Franci
11, 2016)

Spark Summit East
2016, New York) ag
posted (Jan 14, 2016



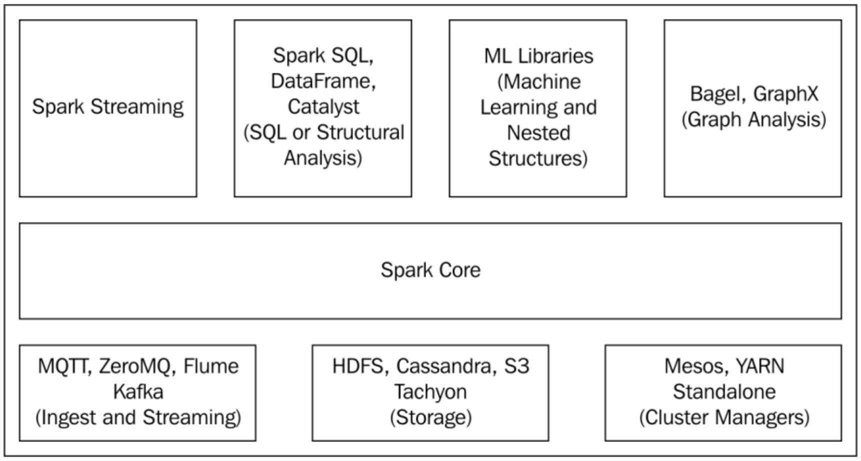


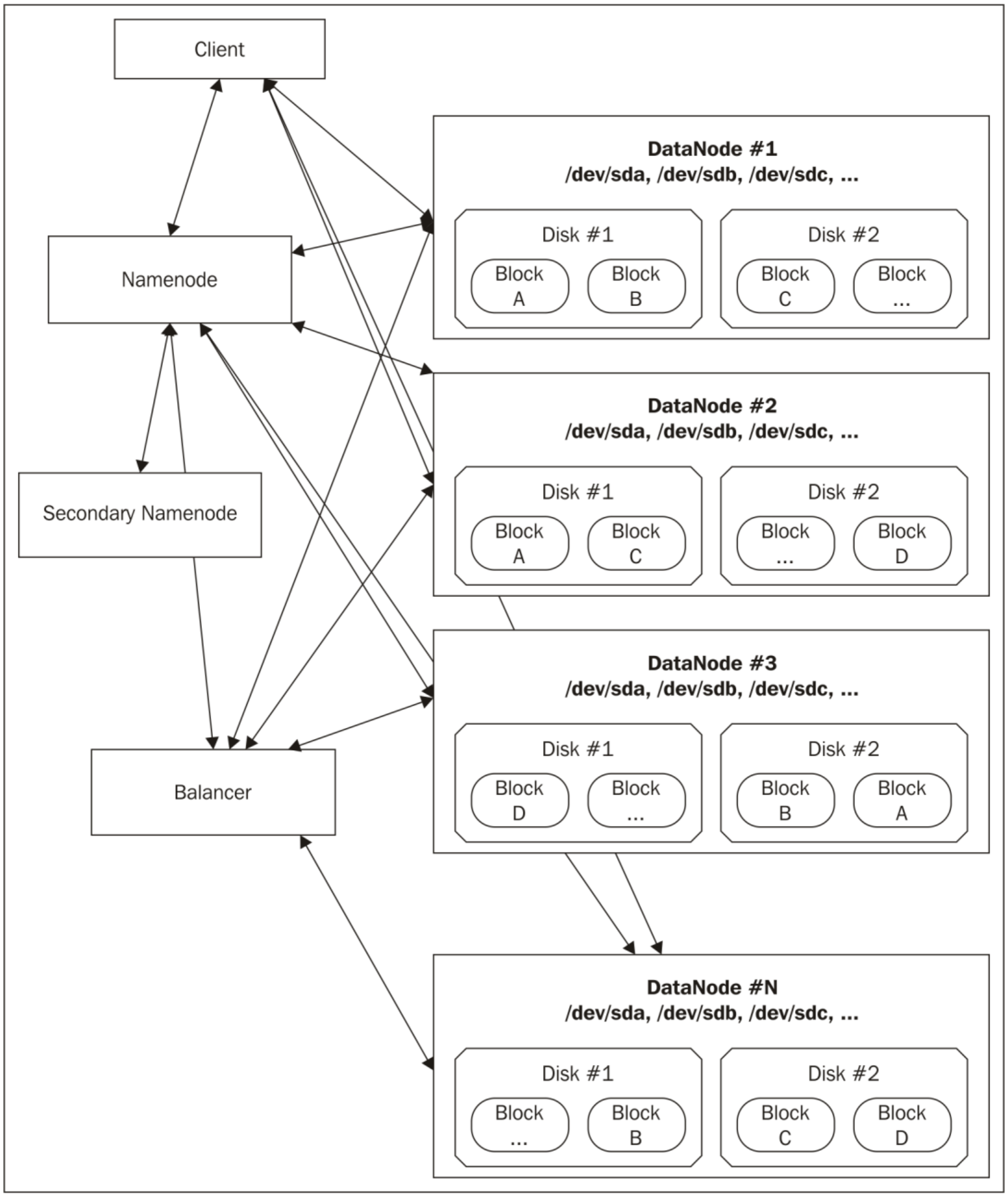
Summary Metrics for 10 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	0.3 s	0.3 s	0.4 s	0.5 s	0.7 s
GC Time	0 ms	0 ms	0 ms	0 ms	0 ms
Shuffle Read Size / Records	65.2 KB / 6622	69.0 KB / 6917	69.4 KB / 7027	69.6 KB / 7096	71.1 KB / 7133

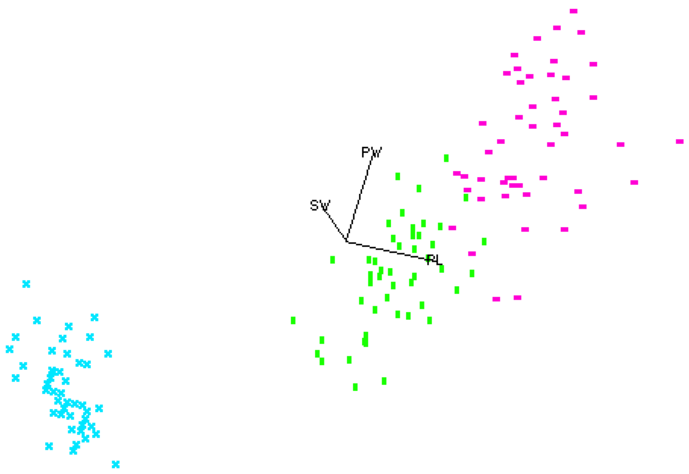
Aggregated Metrics by Executor

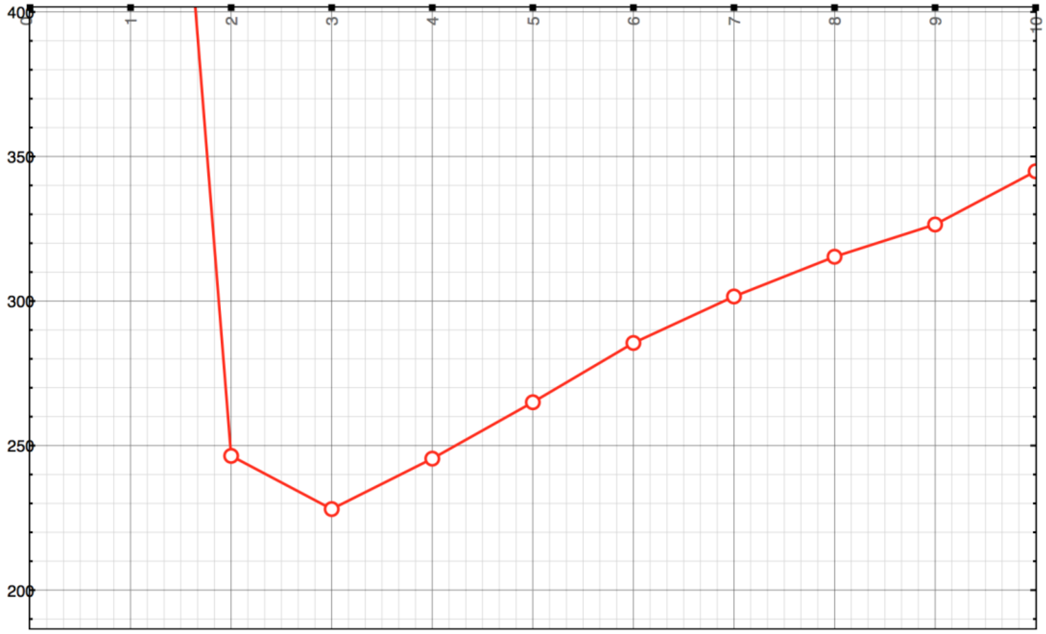
Executor ID	Address	Task Time	Total Tasks	Failed Tasks	Succeeded Tasks	Shuffle Read Size / Records
0	10.10.30.57:39552	0.8 s	2	0	2	140.2 KB / 14051
1	10.10.30.54:33016	1 s	2	0	2	131.8 KB / 13324
2	10.10.30.56:37281	0.8 s	1	0	1	69.6 KB / 7133
3	10.10.30.55:49024	0.8 s	2	0	2	138.2 KB / 13905
4	10.10.30.53:57738	2 s	3	0	3	209.3 KB / 21203

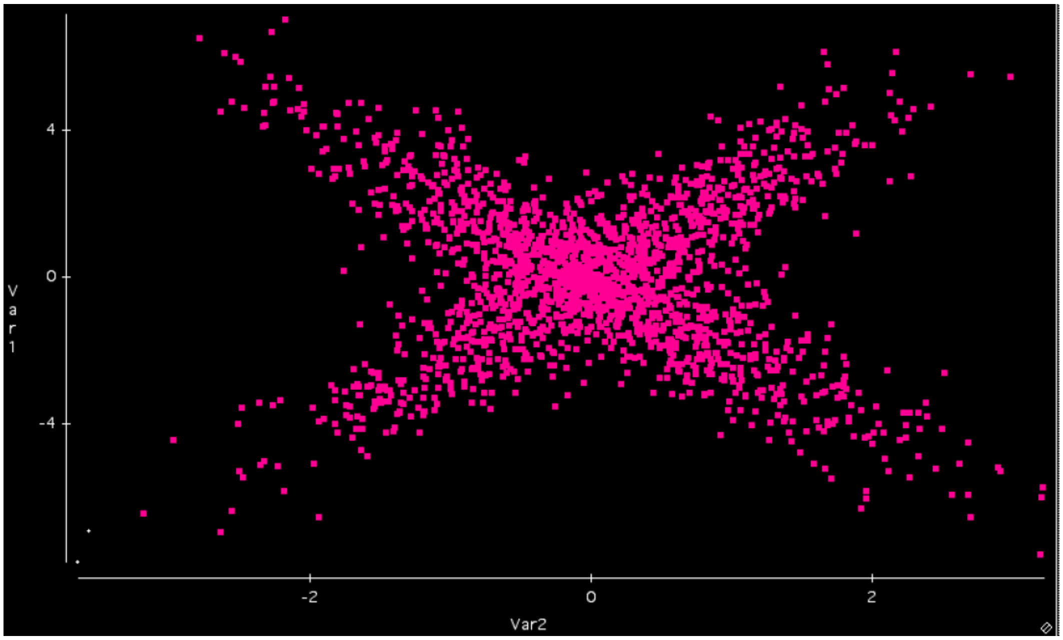

Spark Streaming
Spark SQL, DataFrame, Catalyst (SQL or Structural Analysis)
ML Libraries (Machine Learning and Nested Structures)
Bagel, GraphX (Graph Analysis)
Spark Core
MQTT, ZeroMQ, Flume Kafka (Ingest and Streaming)
HDFS, Cassandra, S3 Tachyon (Storage)
Mesos, YARN Standalone (Cluster Managers)


Client
Namenode
Secondary Namenode
Balancer
DataNode #1
/dev/sda, /dev/sdb, /dev/sdc, ...
Disk #1
Block A
Block B
Disk #2
Block C
Block ...
DataNode #2
/dev/sda, /dev/sdb, /dev/sdc, ...
Disk #1
Block A
Block C
Disk #2
Block ...
Block D
DataNode #3
/dev/sda, /dev/sdb, /dev/sdc, ...
Disk #1
Block D
Block ...
Disk #2
Block B
Block A
DataNode #N
/dev/sda, /dev/sdb, /dev/sdc, ...
Disk #1
Block ...
Block B
Disk #2
Block C
Block D

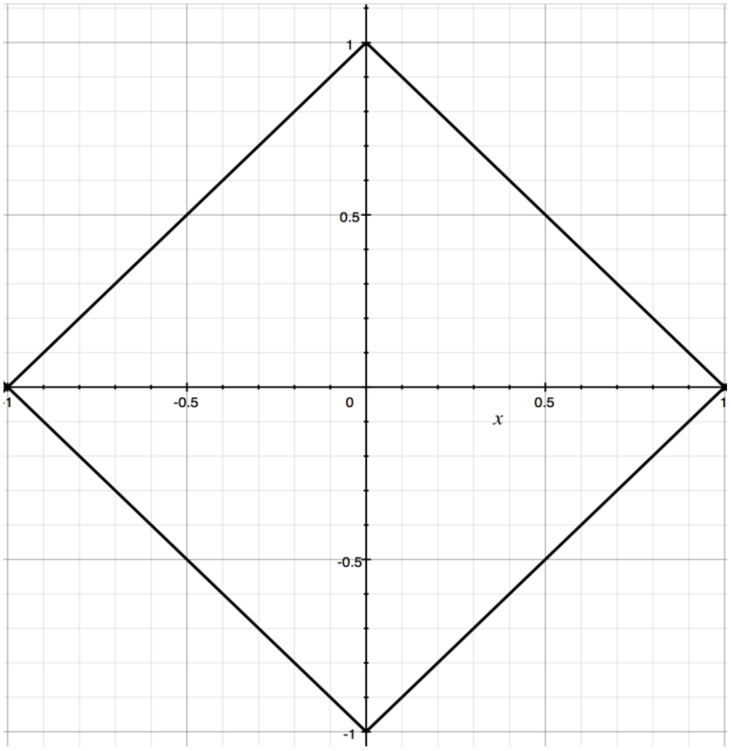
Chapter 4 - Supervised and Unsupervised Learning

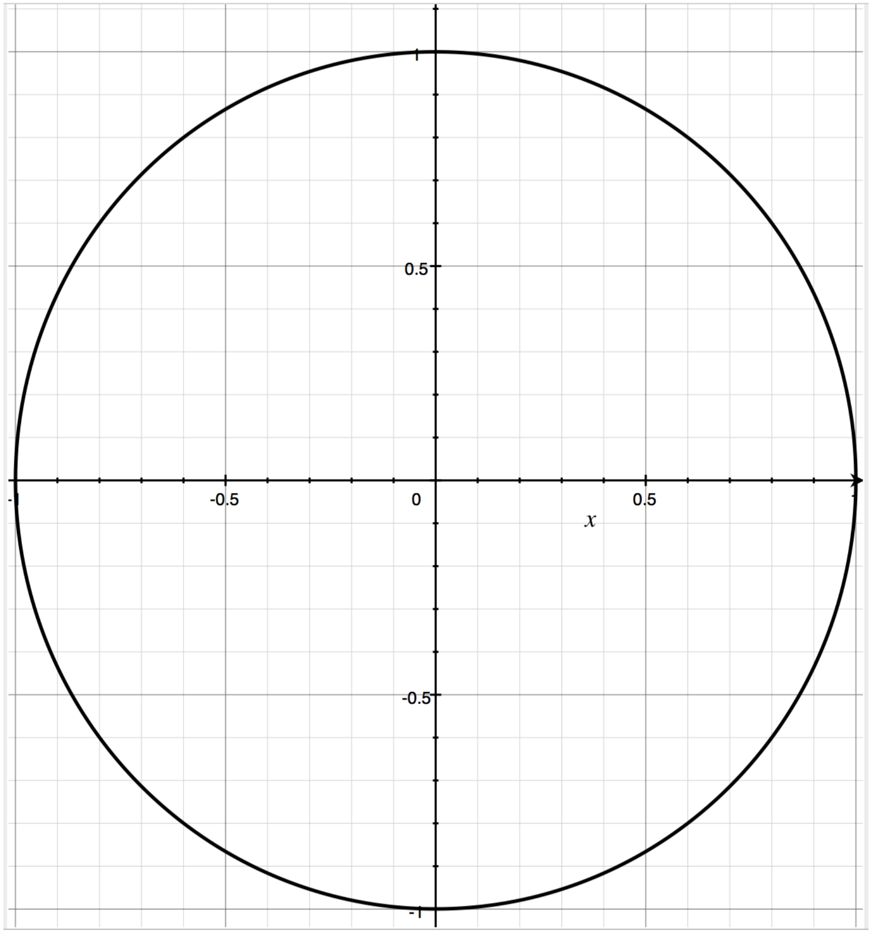


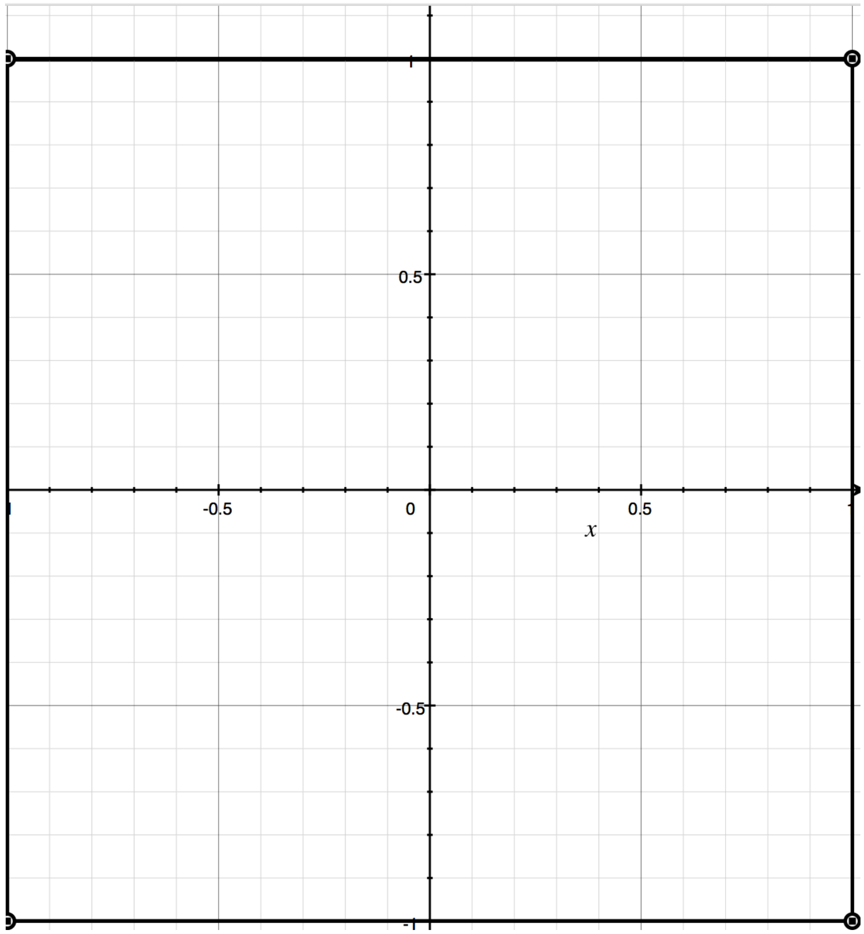

400
350
300
250
200
0
1
2
3
4
5
6
7
8
9
10

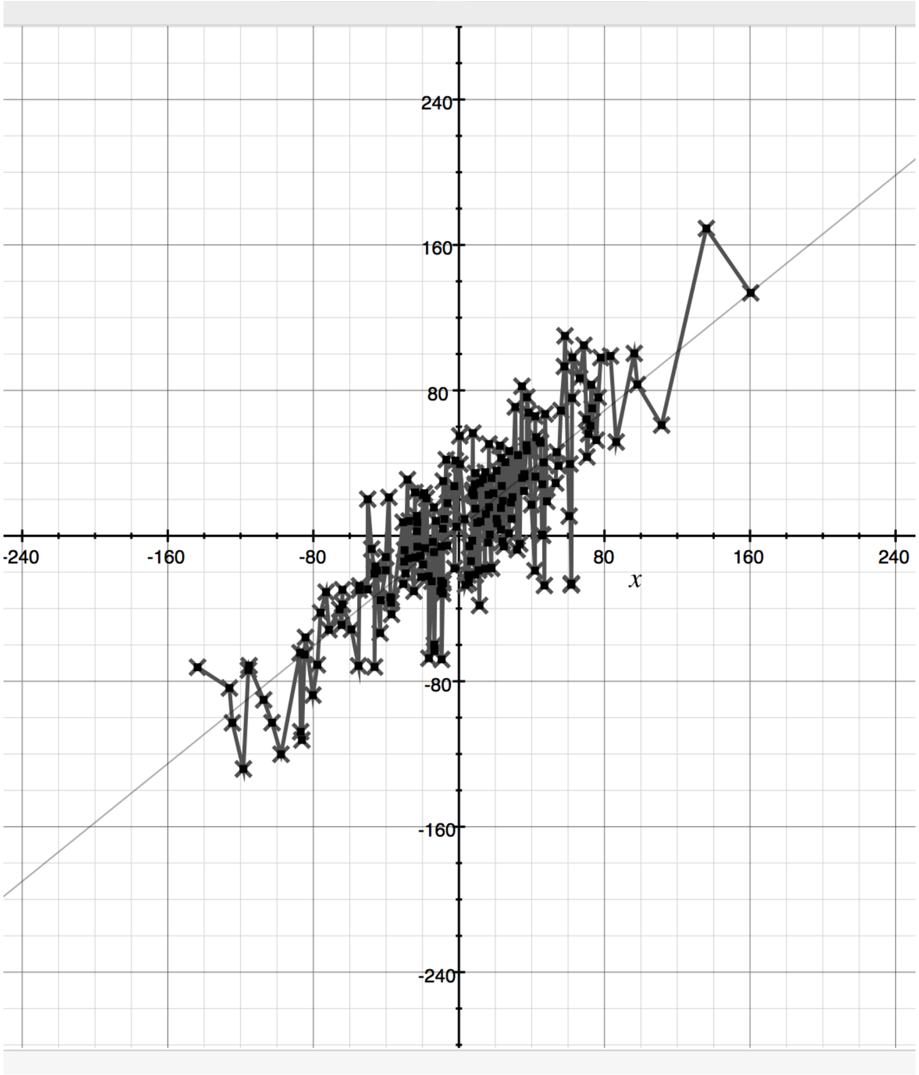

4
0
-4
Var1
-2
0
2
Var2

Chapter 5 - Regression and Classification

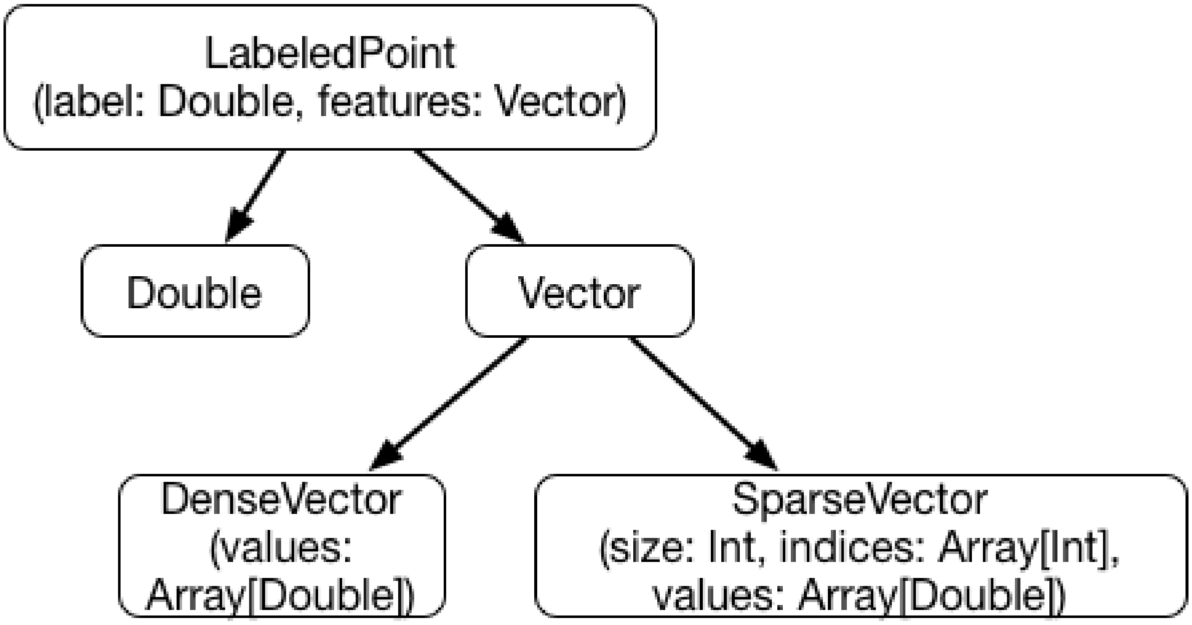



1
0.5
-1
-0.5
0
0.5
x
-0.5
-1

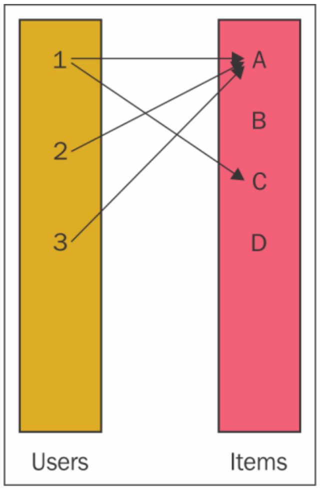

1
0.5
-0.5
-1
-0.5
0
0.5
x


240
160
80
-80
-160
-240
-240
-160
-80
80
160
240
x

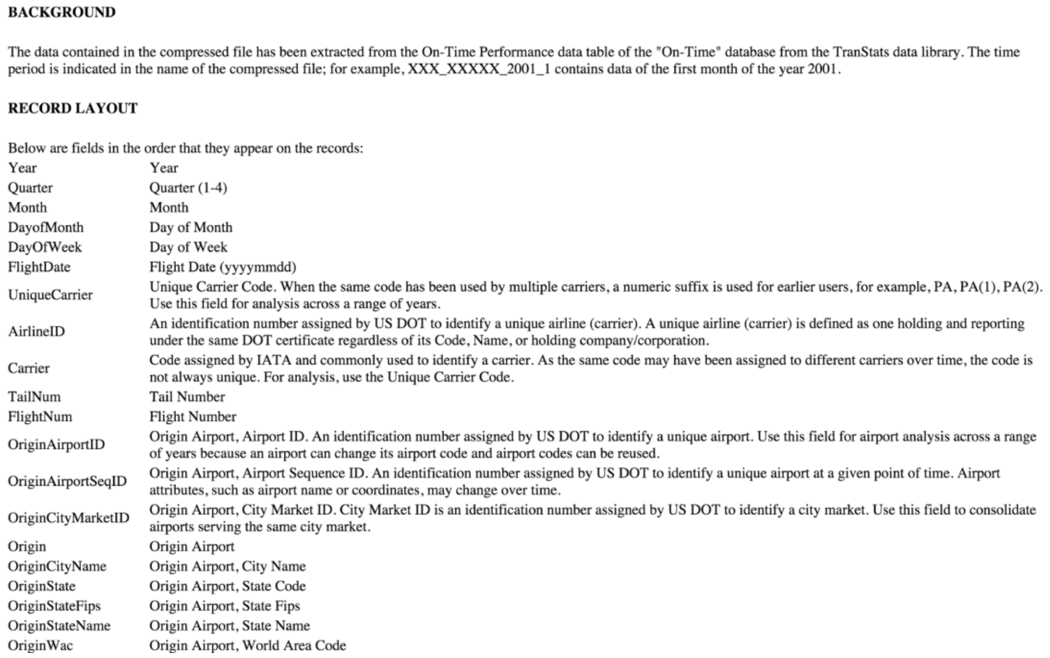
Chapter 6 - Working with Unstructured Data



Chapter 7 - Working with Graph Algorithms



Chapter 8 - Integrating Scala with R and Python



BACKGROUND

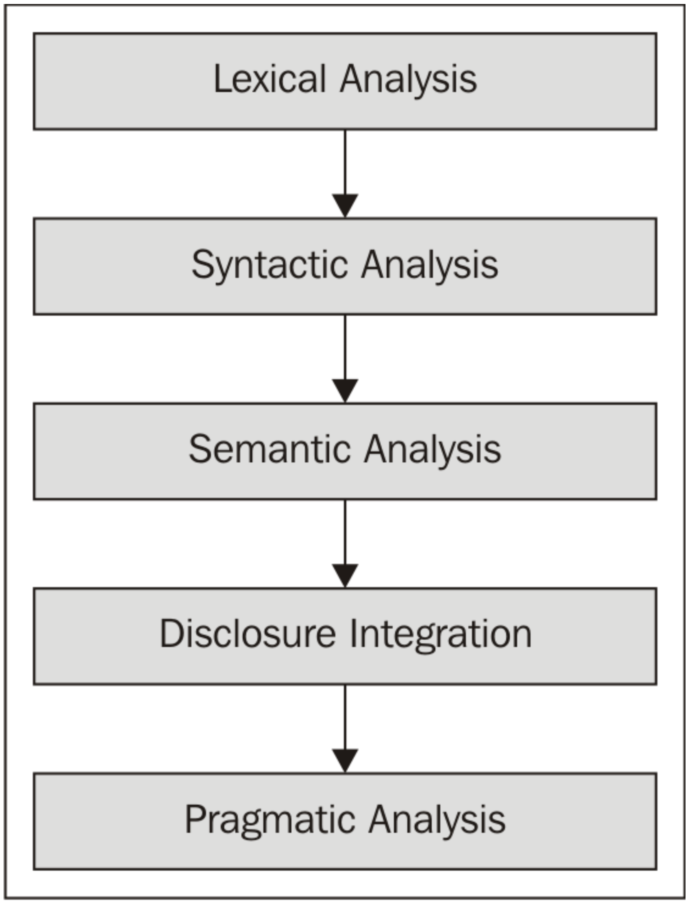
The data contained in the compressed file has been extracted from the On-Time Performance data table of the "On-Time" database from the TranStats data library. The time period is indicated in the name of the compressed file; for example, XXX_XXXXX_2001_1 contains data of the first month of the year 2001.

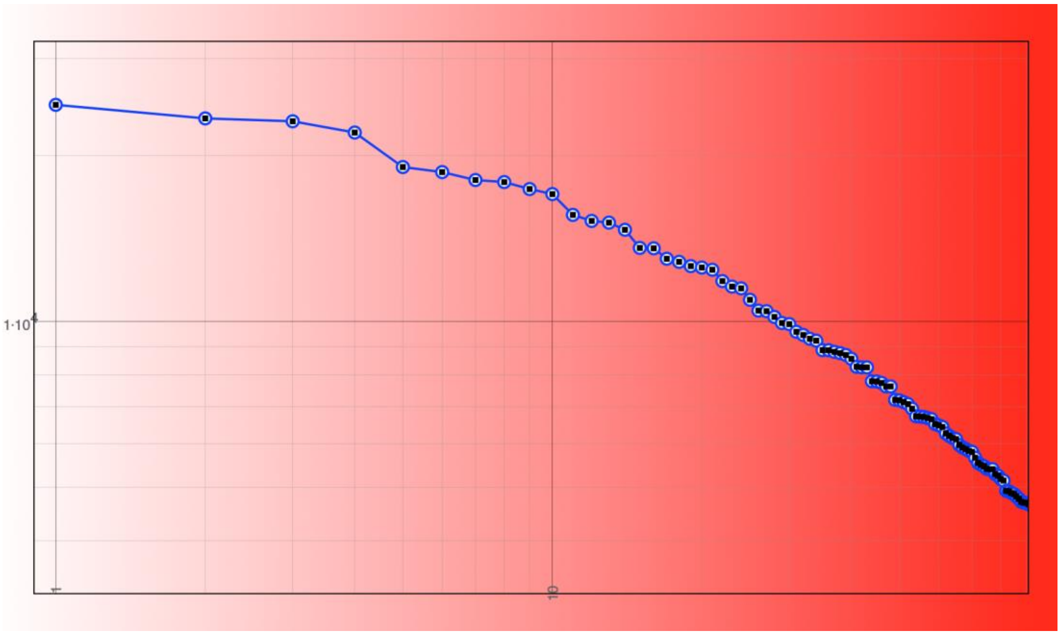
RECORD LAYOUT

Below are fields in the order that they appear on the records:

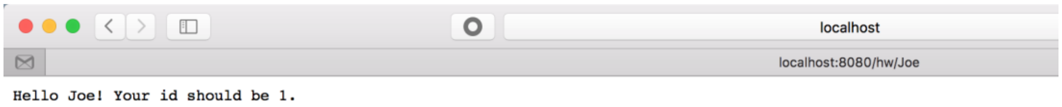
Year	Year
Quarter	Quarter (1-4)
Month	Month
DayofMonth	Day of Month
DayOfWeek	Day of Week
FlightDate	Flight Date (yyyymmdd)
UniqueCarrier	Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years.
AirlineID	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation.
Carrier	Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique. For analysis, use the Unique Carrier Code.
TailNum	Tail Number
FlightNum	Flight Number
OriginAirportID	Origin Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused.
OriginAirportSeqID	Origin Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time.
OriginCityMarketID	Origin Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market.
Origin	Origin Airport
OriginCityName	Origin Airport, City Name
OriginState	Origin Airport, State Code
OriginStateFips	Origin Airport, State Fips
OriginStateName	Origin Airport, State Name
OriginWac	Origin Airport, World Area Code

Chapter 9 - NLP in Scala

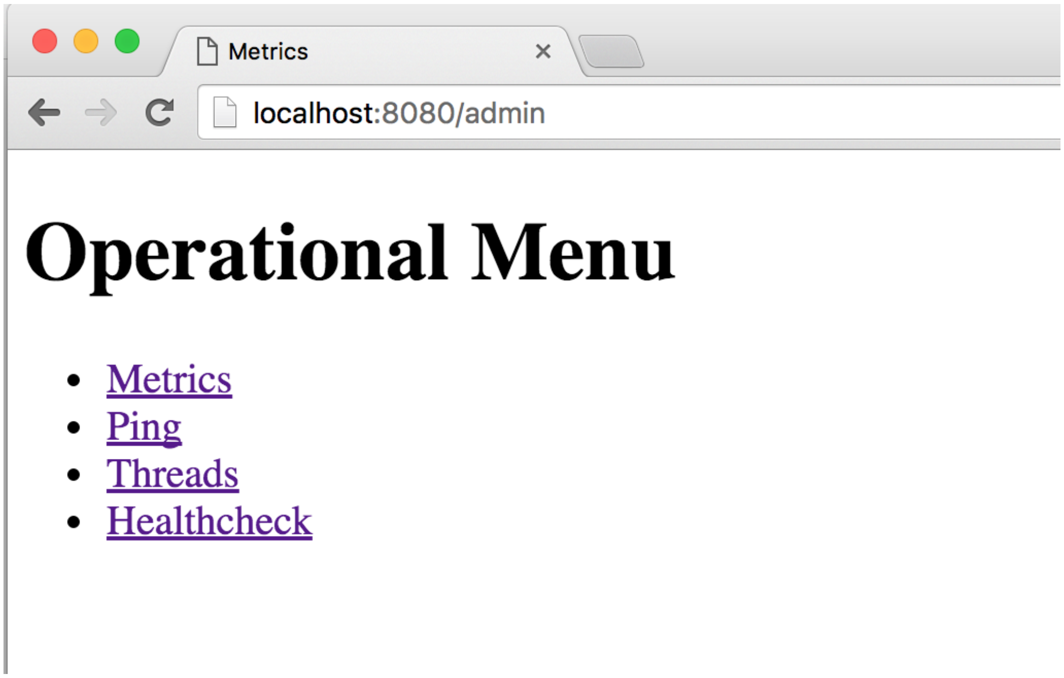



1·10⁴
1
10

Chapter 10 - Advanced Model Monitoring

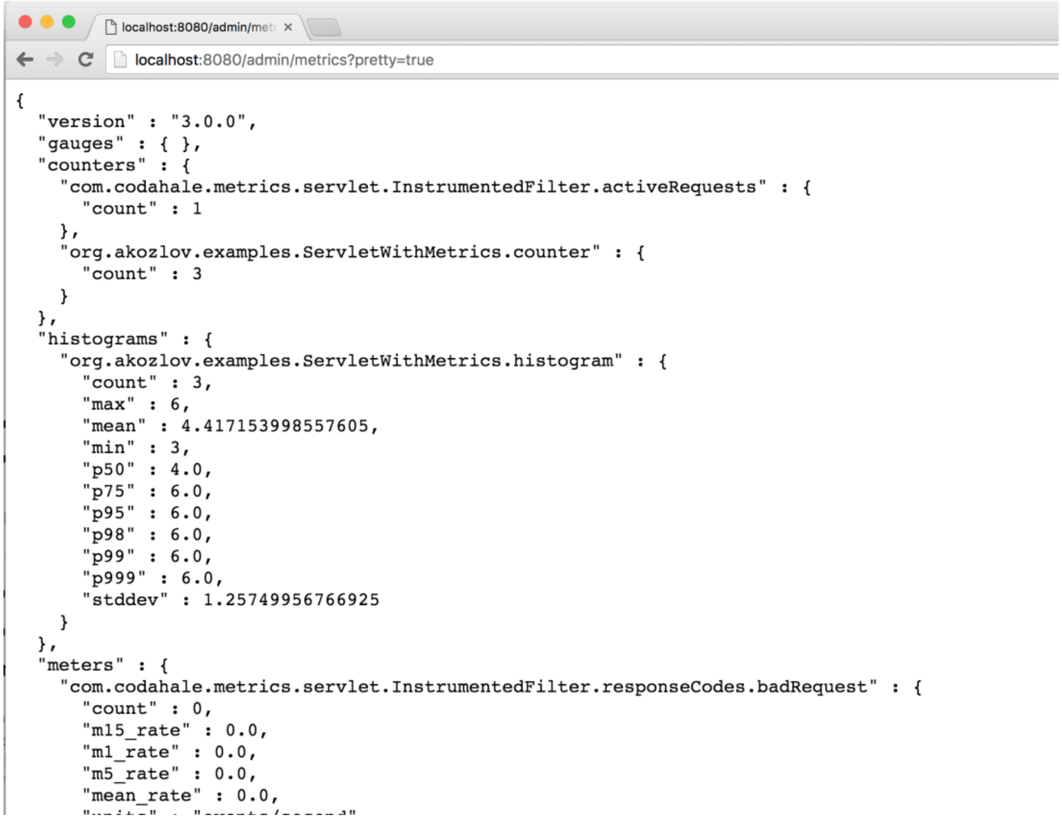


```
Hello Joe! Your id should be 1.
```

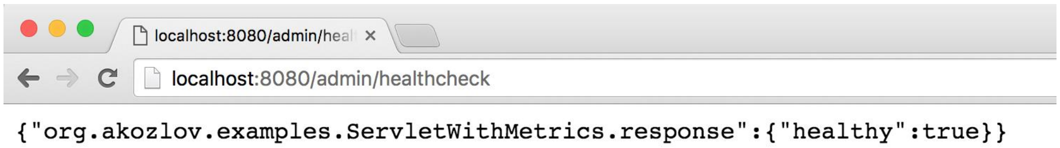


Operational Menu

- Metrics
- Ping
- Threads
- Healthcheck


localhost:8080/admin/metr
localhost:8080/admin/metrics?pretty=true

```
{
  "version" : "3.0.0",
  "gauges" : { },
  "counters" : {
    "com.codahale.metrics.servlet.InstrumentedFilter.activeRequests" : {
      "count" : 1
    },
    "org.akozlov.examples.ServletWithMetrics.counter" : {
      "count" : 3
    }
  },
  "histograms" : {
    "org.akozlov.examples.ServletWithMetrics.histogram" : {
      "count" : 3,
      "max" : 6,
      "mean" : 4.417153998557605,
      "min" : 3,
      "p50" : 4.0,
      "p75" : 6.0,
      "p95" : 6.0,
      "p98" : 6.0,
      "p99" : 6.0,
      "p999" : 6.0,
      "stddev" : 1.25749956766925
    }
  },
  "meters" : {
    "com.codahale.metrics.servlet.InstrumentedFilter.responseCodes.badRequest" : {
      "count" : 0,
      "m15_rate" : 0.0,
      "m1_rate" : 0.0,
      "m5_rate" : 0.0,
      "mean_rate" : 0.0,
      "units" : "events/second"
```



```
{"org.akozlov.examples.ServletWithMetrics.response":{"healthy":true}}
```