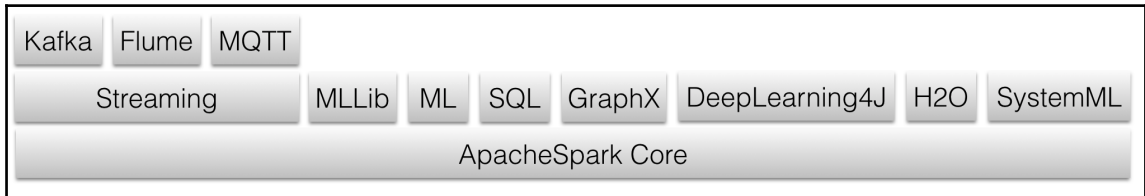
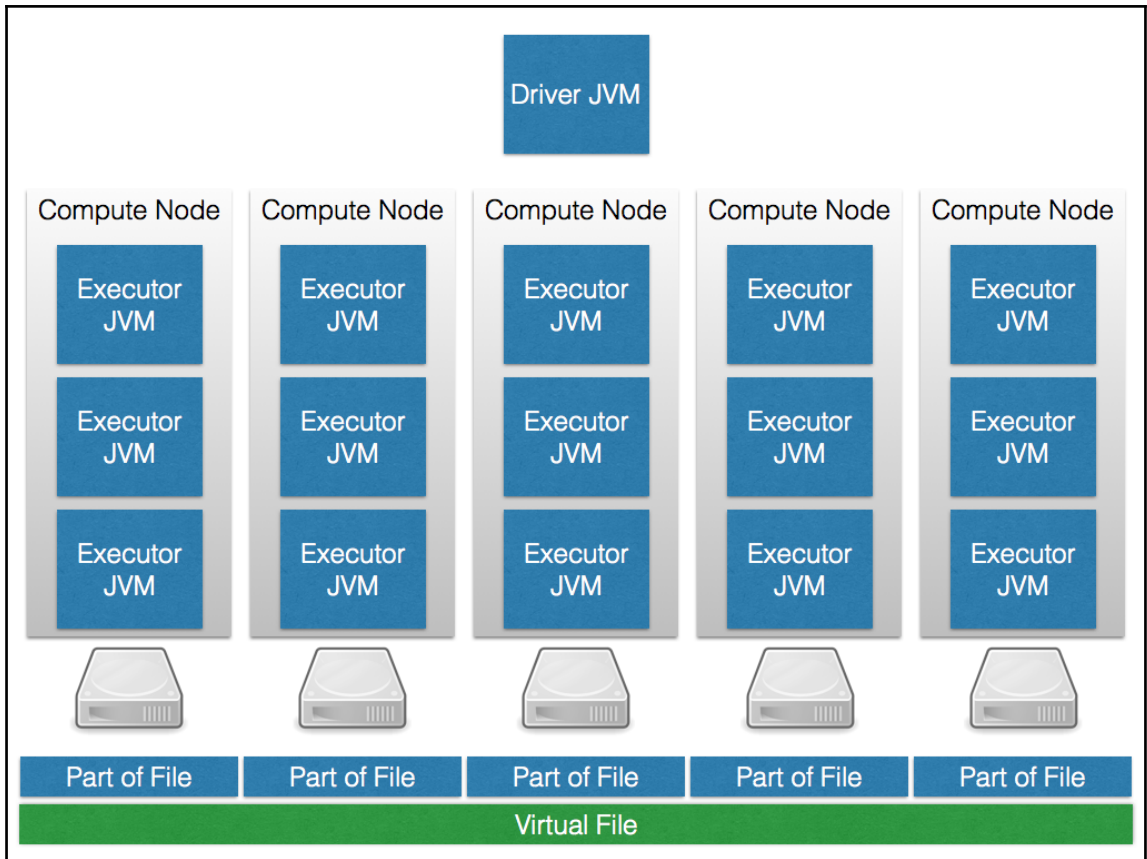
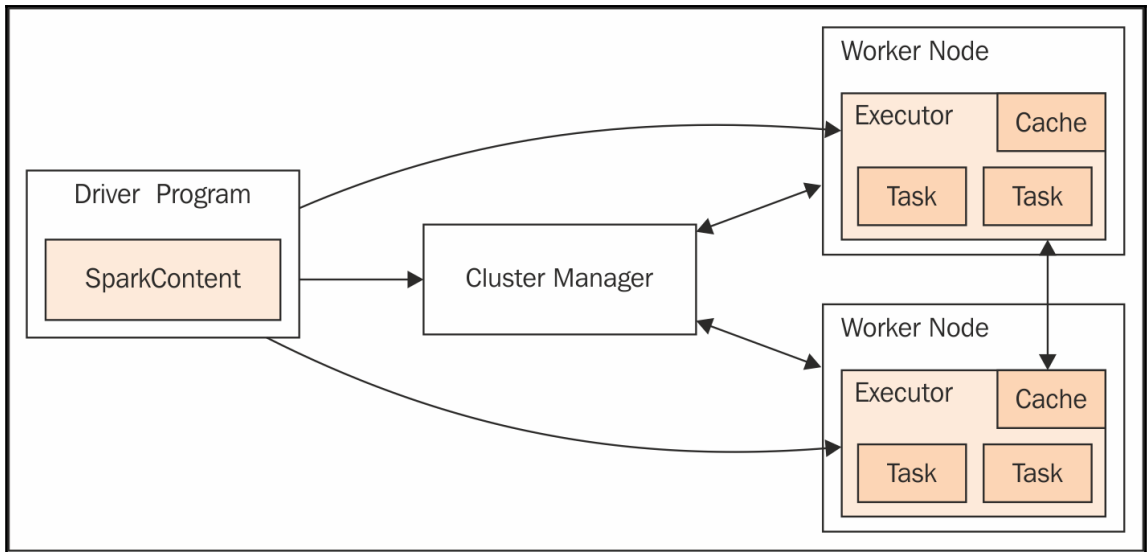


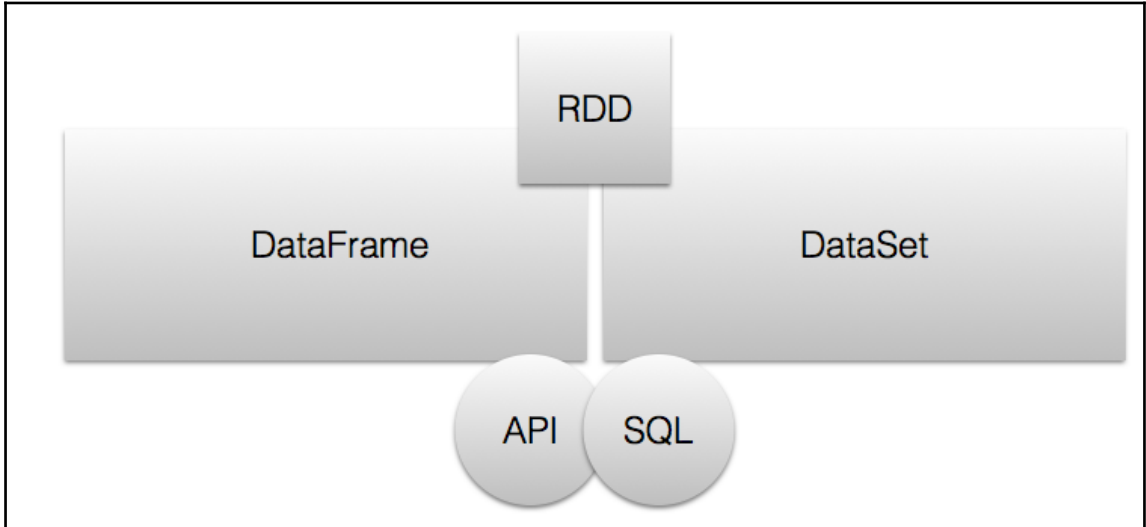
# A First Taste and What's New in Apache Spark V2







# Apache Spark SQL



```
Romeos-MacBook-Pro:~ romeokienzler$ hdfs dfs -ls /tmp/test.json
17/01/09 22:28:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 9 items
-rw-r--r-- 3 romeokienzler supergroup 0 2017-01-09 22:28 /tmp/test.json/_SUCCESS
-rw-r--r-- 3 romeokienzler supergroup 0 2017-01-09 22:28 /tmp/test.json/part-r-00000-cd2d2d53-969b-4ad9-9700-62740dfd1779.json
-rw-r--r-- 3 romeokienzler supergroup 0 2017-01-09 22:28 /tmp/test.json/part-r-00001-cd2d2d53-969b-4ad9-9700-62740dfd1779.json
-rw-r--r-- 3 romeokienzler supergroup 12 2017-01-09 22:28 /tmp/test.json/part-r-00002-cd2d2d53-969b-4ad9-9700-62740dfd1779.json
-rw-r--r-- 3 romeokienzler supergroup 0 2017-01-09 22:28 /tmp/test.json/part-r-00003-cd2d2d53-969b-4ad9-9700-62740dfd1779.json
-rw-r--r-- 3 romeokienzler supergroup 0 2017-01-09 22:28 /tmp/test.json/part-r-00004-cd2d2d53-969b-4ad9-9700-62740dfd1779.json
-rw-r--r-- 3 romeokienzler supergroup 12 2017-01-09 22:28 /tmp/test.json/part-r-00005-cd2d2d53-969b-4ad9-9700-62740dfd1779.json
-rw-r--r-- 3 romeokienzler supergroup 0 2017-01-09 22:28 /tmp/test.json/part-r-00006-cd2d2d53-969b-4ad9-9700-62740dfd1779.json
-rw-r--r-- 3 romeokienzler supergroup 12 2017-01-09 22:28 /tmp/test.json/part-r-00007-cd2d2d53-969b-4ad9-9700-62740dfd1779.json
```

```
Romeos-MacBook-Pro:~ romeokienzler$ hdfs dfs -ls /tmp/test_single_partition.json
17/01/09 22:32:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r-- 3 romeokienzler supergroup 0 2017-01-09 22:31 /tmp/test_single_partition.json/_SUCCESS
-rw-r--r-- 3 romeokienzler supergroup 36 2017-01-09 22:31 /tmp/test_single_partition.json/part-r-00000-f764852b-0ffa-4b58-9cdb-3fdd684c6789.json
```

```
Romeos-MacBook-Pro:~ romeokienzler$ hdfs dfs -cat /tmp/test_single_partition.json/part-r-00000-f764852b-0ffa-4b58-9cdb-3fdd684c6789.json
{"value":1}
{"value":2}
{"value":3}
```

```
Romeos-MacBook-Pro:~ romeokienzler$ hdfs dfs -ls /tmp/test.parquet
17/01/09 22:36:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 9 items
-rw-r--r-- 3 romeokienzler supergroup 0 2017-01-09 22:36 /tmp/test.parquet/_SUCCESS
-rw-r--r-- 3 romeokienzler supergroup 252 2017-01-09 22:36 /tmp/test.parquet/part-r-00000-72743c20-e4a9-4cde-a7df-07a56615ea66.snappy.parquet
-rw-r--r-- 3 romeokienzler supergroup 252 2017-01-09 22:36 /tmp/test.parquet/part-r-00001-72743c20-e4a9-4cde-a7df-07a56615ea66.snappy.parquet
-rw-r--r-- 3 romeokienzler supergroup 343 2017-01-09 22:36 /tmp/test.parquet/part-r-00002-72743c20-e4a9-4cde-a7df-07a56615ea66.snappy.parquet
-rw-r--r-- 3 romeokienzler supergroup 252 2017-01-09 22:36 /tmp/test.parquet/part-r-00003-72743c20-e4a9-4cde-a7df-07a56615ea66.snappy.parquet
-rw-r--r-- 3 romeokienzler supergroup 252 2017-01-09 22:36 /tmp/test.parquet/part-r-00004-72743c20-e4a9-4cde-a7df-07a56615ea66.snappy.parquet
-rw-r--r-- 3 romeokienzler supergroup 343 2017-01-09 22:36 /tmp/test.parquet/part-r-00005-72743c20-e4a9-4cde-a7df-07a56615ea66.snappy.parquet
-rw-r--r-- 3 romeokienzler supergroup 252 2017-01-09 22:36 /tmp/test.parquet/part-r-00006-72743c20-e4a9-4cde-a7df-07a56615ea66.snappy.parquet
-rw-r--r-- 3 romeokienzler supergroup 343 2017-01-09 22:36 /tmp/test.parquet/part-r-00007-72743c20-e4a9-4cde-a7df-07a56615ea66.snappy.parquet
```

```
scala> val washing = spark.read.json("hdfs://localhost:9000/tmp/washing.json")
washing: org.apache.spark.sql.DataFrame = [_corrupt_record: string, doc: struct<_id: string, _rev: string ... 9 more fields> ... 3 more fields]

scala> washing.printSchema
root
 |-- _corrupt_record: string (nullable = true)
 |-- doc: struct (nullable = true)
 |   |-- _id: string (nullable = true)
 |   |-- _rev: string (nullable = true)
 |   |-- count: long (nullable = true)
 |   |-- flowrate: long (nullable = true)
 |   |-- fluidlevel: string (nullable = true)
 |   |-- frequency: long (nullable = true)
 |   |-- hardness: long (nullable = true)
 |   |-- speed: long (nullable = true)
 |   |-- temperature: long (nullable = true)
 |   |-- ts: long (nullable = true)
 |   |-- voltage: long (nullable = true)
 |-- id: string (nullable = true)
 |-- key: string (nullable = true)
 |-- value: struct (nullable = true)
 |     |-- rev: string (nullable = true)
```

```
scala> val washing_flat = washing.select("doc.*")
washing_flat: org.apache.spark.sql.DataFrame = [_id: string, _rev: string ... 9 more fields]

scala> washing_flat.printSchema
root
 |-- _id: string (nullable = true)
 |-- _rev: string (nullable = true)
 |-- count: long (nullable = true)
 |-- flowrate: long (nullable = true)
 |-- fluidlevel: string (nullable = true)
 |-- frequency: long (nullable = true)
 |-- hardness: long (nullable = true)
 |-- speed: long (nullable = true)
 |-- temperature: long (nullable = true)
 |-- ts: long (nullable = true)
 |-- voltage: long (nullable = true)
```

```
scala> washing_flat.select("temperature","hardness","voltage","speed").show(3)
+-----+-----+-----+-----+
|temperature|hardness|voltage|speed|
+-----+-----+-----+-----+
|          null|      null|      null| 1259|
|          null|      null|      237|  null|
|           99|      105|      null|  null|
+-----+-----+-----+-----+
only showing top 3 rows
```

```
scala> washing_flat.select("voltage","frequency").filter(washing_flat("voltage")>235).show(3)
+-----+-----+
|voltage|frequency|
+-----+-----+
|    237|      72|
|    244|      66|
|    253|      71|
+-----+-----+
only showing top 3 rows
```

```
scala> washing_flat.groupBy("fluidlevel").count().show()
+-----+-----+
|fluidlevel|count|
+-----+-----+
|      null| 8254|
|acceptable|15449|
+-----+-----+
```

```
scala> washing_flat.createOrReplaceTempView("washing_flat")

scala> spark.sql("select count(*) from washing_flat").show
+-----+
|count(1)|
+-----+
|   23703|
+-----+
```

```
Romeos-MacBook-Pro:~ romeokienzler$ hdfs dfs -ls /tmp/washing_flat.csv/
17/01/09 23:16:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 3 items
-rw-r--r--  3 romeokienzler supergroup          0 2017-01-09 23:16 /tmp/washing_flat.csv/_SUCCESS
-rw-r--r--  3 romeokienzler supergroup    1403285 2017-01-09 23:16 /tmp/washing_flat.csv/part-r-00000-60cc84d1-f7bb-4f3d-bf2b-b581fbbe6658.csv
-rw-r--r--  3 romeokienzler supergroup    1114646 2017-01-09 23:16 /tmp/washing_flat.csv/part-r-00001-60cc84d1-f7bb-4f3d-bf2b-b581fbbe6658.csv
```

```
Romeos-MacBook-Pro:~ romeokienzler$ hdfs dfs -tail /tmp/washing_flat.csv/part-r-00000-60cc84d1-f7bb-4f3d-bf2b-b581fbbe6658.csv
17/01/09 23:18:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23a78e72483c5f0130b42fc5323a8,12537,11,acceptable,,74,,94,1480435944176,
8fc6ac8feea61a782bf984129de154fe,1-fc1fc74a72d0322b55b501c1350b0da7,4186,,60,,1480435947899,230
8fc6ac8feea61a782bf984129de215af,1-14880e64d3963970bdc36bdc84486e3,4187,,78,,1480435950901,228
8fc6ac8feea61a782bf984129de3eb1f,1-3033ac666a37e9b320f80b28a6311e3f,12552,11,acceptable,,70,,91,1480435959210,
8fc6ac8feea61a782bf984129de6ea16,1-961335b72130cfb0b352e377d58749b4,4194,,71,,1480435971918,243
8fc6ac8feea61a782bf984129de7206e,1-c782d2a514ef7a795c46b6804f33ffb3,12566,11,acceptable,,74,,100,1480435973233,
8fc6ac8feea61a782bf984129de8fe4c,1-67eddbfdcc30af8fb37a14f1cf1b0a,12573,11,acceptable,,78,,99,1480435980246,
8fc6ac8feea61a782bf984129dead56b,1-a7c99c2314e0787029ac45c58381b5a4,12579,11,acceptable,,74,,81,1480435986256,
8fc6ac8feea61a782bf984129deb0817,1-5649ad7f0b66dea9c55bed9abf36825b,12580,11,acceptable,,80,,83,1480435987258,
8fc6ac8feea61a782bf984129dec976c,1-3433987da79cf6b044545a708aa7a5ff,2521,,,,,1051,,1480435991874,
```

```
scala> val csvDF = spark.read.csv("hdfs://localhost:9000/tmp/washing_flat.csv")
csvDF: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 9 more fields]

scala> csvDF.printSchema
root
 |-- _c0: string (nullable = true)
 |-- _c1: string (nullable = true)
 |-- _c2: string (nullable = true)
 |-- _c3: string (nullable = true)
 |-- _c4: string (nullable = true)
 |-- _c5: string (nullable = true)
 |-- _c6: string (nullable = true)
 |-- _c7: string (nullable = true)
 |-- _c8: string (nullable = true)
 |-- _c9: string (nullable = true)
 |-- _c10: string (nullable = true)
```

```
scala> val rawRDD = sc.textFile("hdfs://localhost:9000/tmp/washing_flat.csv")
rawRDD: org.apache.spark.rdd.RDD[String] = hdfs://localhost:9000/tmp/washing_flat.csv MapPartitionsRDD[49] at textFile at <console>:30

scala> import org.apache.spark.sql.types._
import org.apache.spark.sql.types._

scala> import org.apache.spark.sql._
import org.apache.spark.sql._

scala> val rowRDD = rawRDD.
  |   map(_.split(",")).
  |   map(p => Row(
  |     |   p(0),
  |     |   p(1),
  |     |   p(2).trim.toLong,
  |     |   p(3).trim.toLong,
  |     |   p(4),
  |     |   p(5).trim.toLong,
  |     |   p(6).trim.toLong,
  |     |   p(7).trim.toLong,
  |     |   p(8).trim.toLong,
  |     |   p(9).trim.toLong,
  |     |   p(10).trim.toLong
  |   )
  | )
rowRDD: org.apache.spark.rdd.RDD[org.apache.spark.sql.Row] = MapPartitionsRDD[51] at map at <console>:46
```

```
scala> val washing_flat_df = spark.createDataFrame(rowRDD, schema)
washing_flat_df: org.apache.spark.sql.DataFrame = [_id: string, _rev: string ... 9 more fields]
```

```
scala> washing_flat.printSchema
root
 |-- _id: string (nullable = true)
 |-- _rev: string (nullable = true)
 |-- count: long (nullable = true)
 |-- flowrate: long (nullable = true)
 |-- fluidlevel: string (nullable = true)
 |-- frequency: long (nullable = true)
 |-- hardness: long (nullable = true)
 |-- speed: long (nullable = true)
 |-- temperature: long (nullable = true)
 |-- ts: long (nullable = true)
 |-- voltage: long (nullable = true)
```

```
scala> result.show
17/01/09 23:45:08 WARN WindowExec: No Partition Defined for Window operation! Moving all data to a single partition, this can cause serious performance degradation.
17/01/09 23:45:08 WARN Utils: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.debug.maxToStringFields' in SparkEnv.conf.
-----
|min_temperature|max_temperature|min_voltage|max_voltage|min_flowrate|max_flowrate|min_frequency|max_frequency|min_hardness|max_hardness|min_speed|max_speed|
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 84| 100| 221| 227| 11| 11| 78| 80| 71| 79| 1021| 1021|
| 84| 100| 221| 234| 11| 11| 68| 80| 71| 79| 1021| 1021|
| 84| 100| 221| 234| 11| 11| 68| 80| 71| 78| 1021| 1021|
| 84| 100| 221| 234| 11| 11| 68| 80| 71| 78| 1013| 1021|
| 84| 100| 221| 234| 11| 11| 68| 80| 71| 80| 1013| 1021|
| 84| 100| 221| 234| 11| 11| 68| 80| 71| 80| 1013| 1021|
| 84| 100| 221| 235| 11| 11| 68| 80| 71| 80| 1013| 1013|
| 84| 99| 221| 235| 11| 11| 68| 80| 71| 80| 1013| 1013|
| 84| 99| 234| 235| 11| 11| 68| 69| 71| 80| 1013| 1013|
| 81| 99| 234| 235| 11| 11| 68| 69| 71| 80| 1013| 1013|
| 81| 99| 234| 235| 11| 11| 68| 69| 71| 80| 1013| 1020|
| 81| 99| 228| 235| 11| 11| 60| 69| 71| 80| 1013| 1020|
| 81| 100| 228| 235| 11| 11| 60| 69| 71| 80| 1013| 1020|
| 81| 100| 228| 235| 11| 11| 60| 69| 71| 80| 1013| 1020|
| 81| 100| 228| 235| 11| 11| 60| 73| 71| 80| 1020| 1020|
| 81| 100| 228| 235| 11| 11| 60| 73| 71| 80| 1020| 1020|
| 81| 100| 220| 228| 11| 11| 60| 73| 71| 80| 1020| 1020|
| 81| 100| 220| 228| 11| 11| 60| 73| 72| 80| 1020| 1038|
-----
only showing top 20 rows
```



```
[{"id":"1","name":"testName1","familyName":"familyName1","countryCode":"US","age":33},
{"id":"2","name":"testName2","familyName":"familyName2","countryCode":"DE","age":43},
{"id":"3","name":"testName3","familyName":"familyName3","countryCode":"US","age":53},
{"id":"4","name":"testName4","familyName":"familyName4","countryCode":"CH","age":63},
{"id":"5","name":"testName5","familyName":"familyName5","countryCode":"US","age":73},
{"id":"6","name":"testName6","familyName":"familyName6","countryCode":"DE","age":23},
{"id":"7","name":"testName7","familyName":"familyName7","countryCode":"US","age":36},
{"id":"8","name":"testName8","familyName":"familyName8","countryCode":"CH","age":38}]
```

```
[{"id":"1","clientId":"1","balance":1500},{ "id":"2","clientId":"2","balance":500},
{"id":"3","clientId":"1","balance":1500},{ "id":"4","clientId":"3","balance":500},
{"id":"5","clientId":"1","balance":1500},{ "id":"6","clientId":"4","balance":500},
{"id":"7","clientId":"1","balance":1500},{ "id":"8","clientId":"5","balance":500},
{"id":"9","clientId":"1","balance":1500},{ "id":"10","clientId":"6","balance":500},
{"id":"11","clientId":"1","balance":1500},{ "id":"12","clientId":"7","balance":500},
{"id":"13","clientId":"1","balance":1500},{ "id":"14","clientId":"8","balance":500},
{"id":"15","clientId":"1","balance":1500},{ "id":"16","clientId":"9","balance":500}]
```

```
scala> spark.sql("select * from client").show
```

```
+---+-----+-----+---+-----+
|age|countryCode| familyName| id|      name|
+---+-----+-----+---+-----+
| 33|           US|familyName1|  1|testName1|
| 43|           DE|familyName2|  2|testName2|
| 53|           US|familyName3|  3|testName3|
| 63|           CH|familyName4|  4|testName4|
| 73|           US|familyName5|  5|testName5|
| 23|           DE|familyName6|  6|testName6|
| 36|           US|familyName7|  7|testName7|
| 38|           CH|familyName8|  8|testName8|
+---+-----+-----+---+-----+
```

```
scala> spark.sql("select * from account").show
```

```
+-----+-----+----+
|balance|clientId| id|
+-----+-----+----+
|  1500 |      1 |  1|
|   500 |      2 |  2|
|  1500 |      1 |  3|
|   500 |      3 |  4|
|  1500 |      1 |  5|
|   500 |      4 |  6|
|  1500 |      1 |  7|
|   500 |      5 |  8|
|  1500 |      1 |  9|
|   500 |      6 | 10|
|  1500 |      1 | 11|
|   500 |      7 | 12|
|  1500 |      1 | 13|
|   500 |      8 | 14|
|  1500 |      1 | 15|
|   500 |      9 | 16|
+-----+-----+----+
```

```
scala> spark.sql("select * from account inner join client on account.clientid = client.id").show
```

balance	clientId	id	age	countryCode	familyName	id	name
1500	1	1	33	US	familyName1	1	testName1
500	2	2	43	DE	familyName2	2	testName2
1500	1	3	33	US	familyName1	1	testName1
500	3	4	53	US	familyName3	3	testName3
1500	1	5	33	US	familyName1	1	testName1
500	4	6	63	CH	familyName4	4	testName4
1500	1	7	33	US	familyName1	1	testName1
500	5	8	73	US	familyName5	5	testName5
1500	1	9	33	US	familyName1	1	testName1
500	6	10	23	DE	familyName6	6	testName6
1500	1	11	33	US	familyName1	1	testName1
500	7	12	36	US	familyName7	7	testName7
1500	1	13	33	US	familyName1	1	testName1
500	8	14	38	CH	familyName8	8	testName8
1500	1	15	33	US	familyName1	1	testName1

```
scala> spark.sql("select sum(balance),clientId from account inner join client on account.clientid = client.id group by clientId").show
```

sum(balance)	clientId
500	7
500	3
500	8
500	5
500	6
12000	1
500	4
500	2

```
scala> ds.show
```

```
+---+-----+-----+---+-----+
|age|countryCode|familyName|id|name|
+---+-----+-----+---+-----+
| 33|      US|familyName1| 1|testName1|
| 43|      DE|familyName2| 2|testName2|
| 53|      US|familyName3| 3|testName3|
| 63|      CH|familyName4| 4|testName4|
| 73|      US|familyName5| 5|testName5|
| 23|      DE|familyName6| 6|testName6|
| 36|      US|familyName7| 7|testName7|
| 38|      CH|familyName8| 8|testName8|
+---+-----+-----+---+-----+
```

```
scala> ds.printSchema
```

```
root
```

```
|-- age: long (nullable = true)
|-- countryCode: string (nullable = true)
|-- familyName: string (nullable = true)
|-- id: string (nullable = true)
|-- name: string (nullable = true)
```

```
scala> dsNew.show
+----+-----+
|  _2|avg(_1)|
+----+-----+
|  DE|   33.0|
|  US|   48.75|
|  CH|   50.5|
+----+-----+
```

```
scala> ds.createOrReplaceTempView("ds")

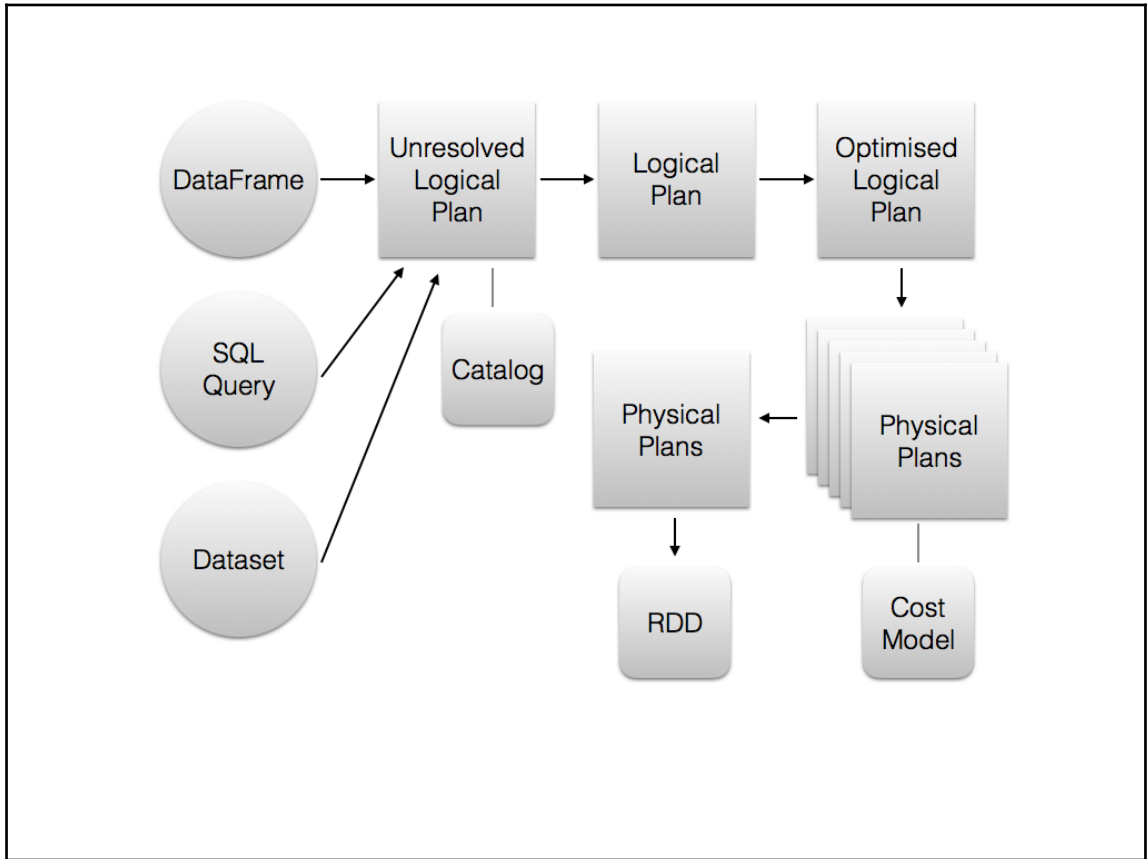
scala> spark.sql("select * from ds").show
+----+-----+-----+----+-----+
|age|countryCode|familyName|id|name|
+----+-----+-----+----+-----+
| 33|           US|familyName1| 1|testName1|
| 43|           DE|familyName2| 2|testName2|
| 53|           US|familyName3| 3|testName3|
| 63|           CH|familyName4| 4|testName4|
| 73|           US|familyName5| 5|testName5|
| 23|           DE|familyName6| 6|testName6|
| 36|           US|familyName7| 7|testName7|
| 38|           CH|familyName8| 8|testName8|
+----+-----+-----+----+-----+
```

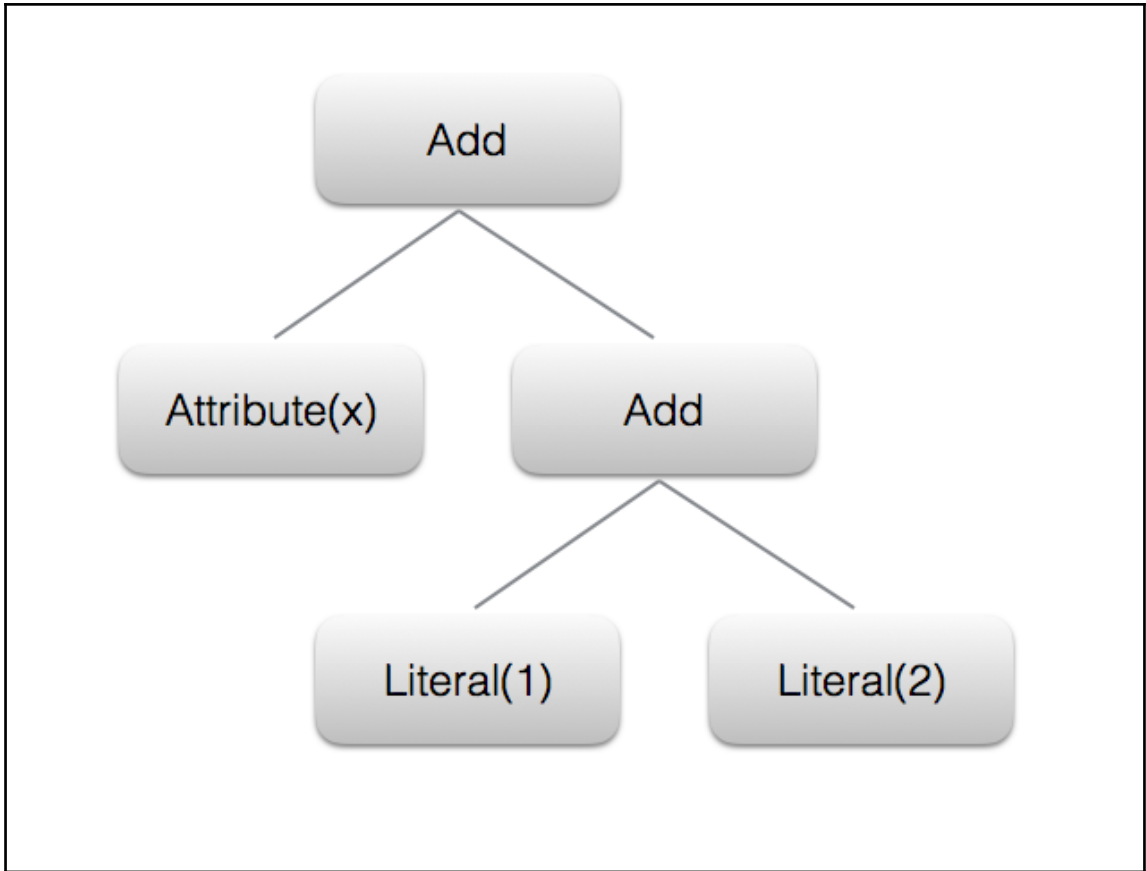
```
scala> spark.udf.register("toAgeRange",AgeRange.asString _)
res12: org.apache.spark.sql.expressions.UserDefinedFunction = UserDefinedFunction(<function1>,StringType,Some(List(IntegerType)))
```

```
scala> spark.sql("select *,toAgeRange(age) as ageRange from client").show
```

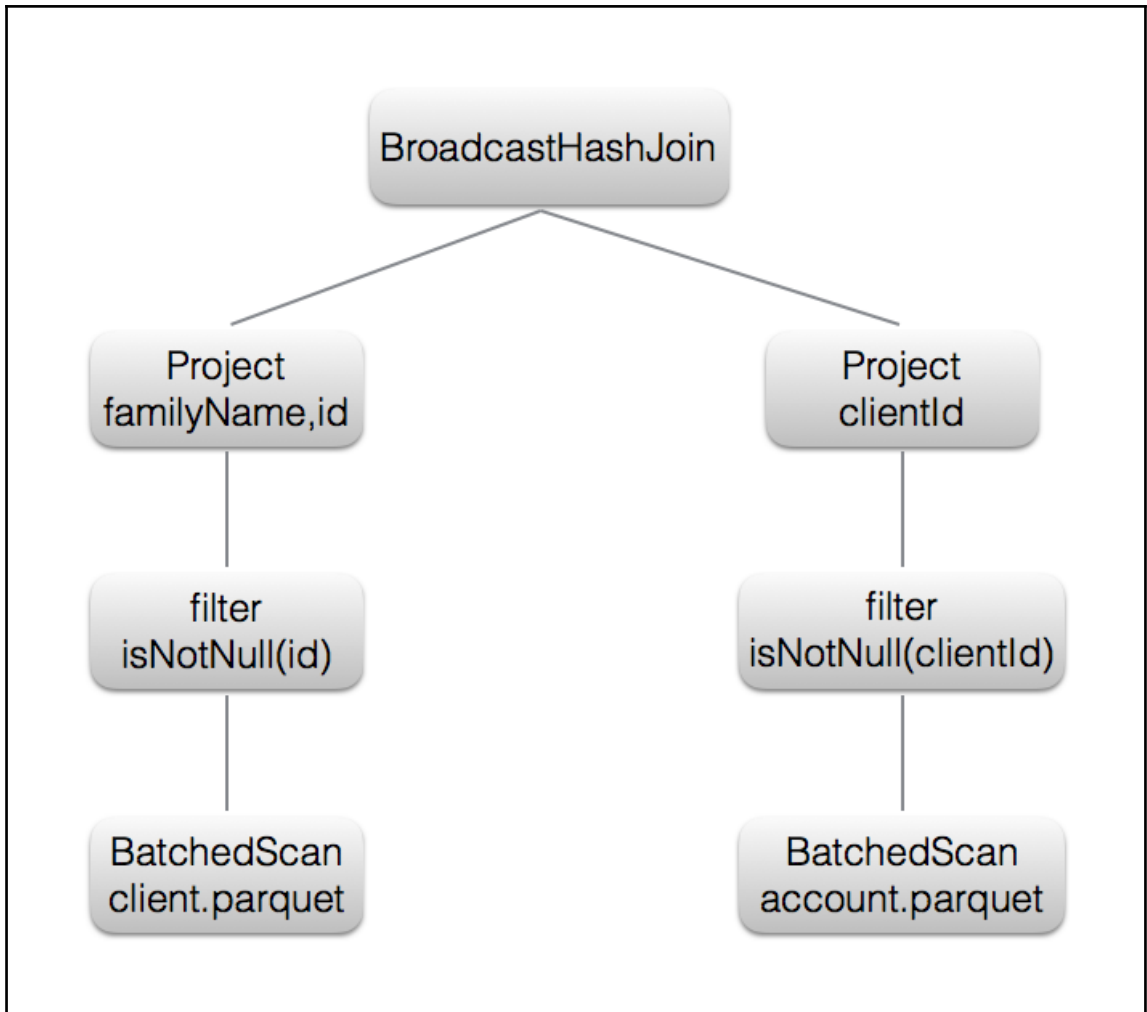
age	countryCode	familyName	id	name	ageRange
33	US	familyName1	1	testName1	Thirty
43	DE	familyName2	2	testName2	Fourty
53	US	familyName3	3	testName3	Fifty
63	CH	familyName4	4	testName4	Sixty
73	US	familyName5	5	testName5	Seventy
23	DE	familyName6	6	testName6	Twenty
36	US	familyName7	7	testName7	Thirty
38	CH	familyName8	8	testName8	Thirty























# The Catalyst Optimizer

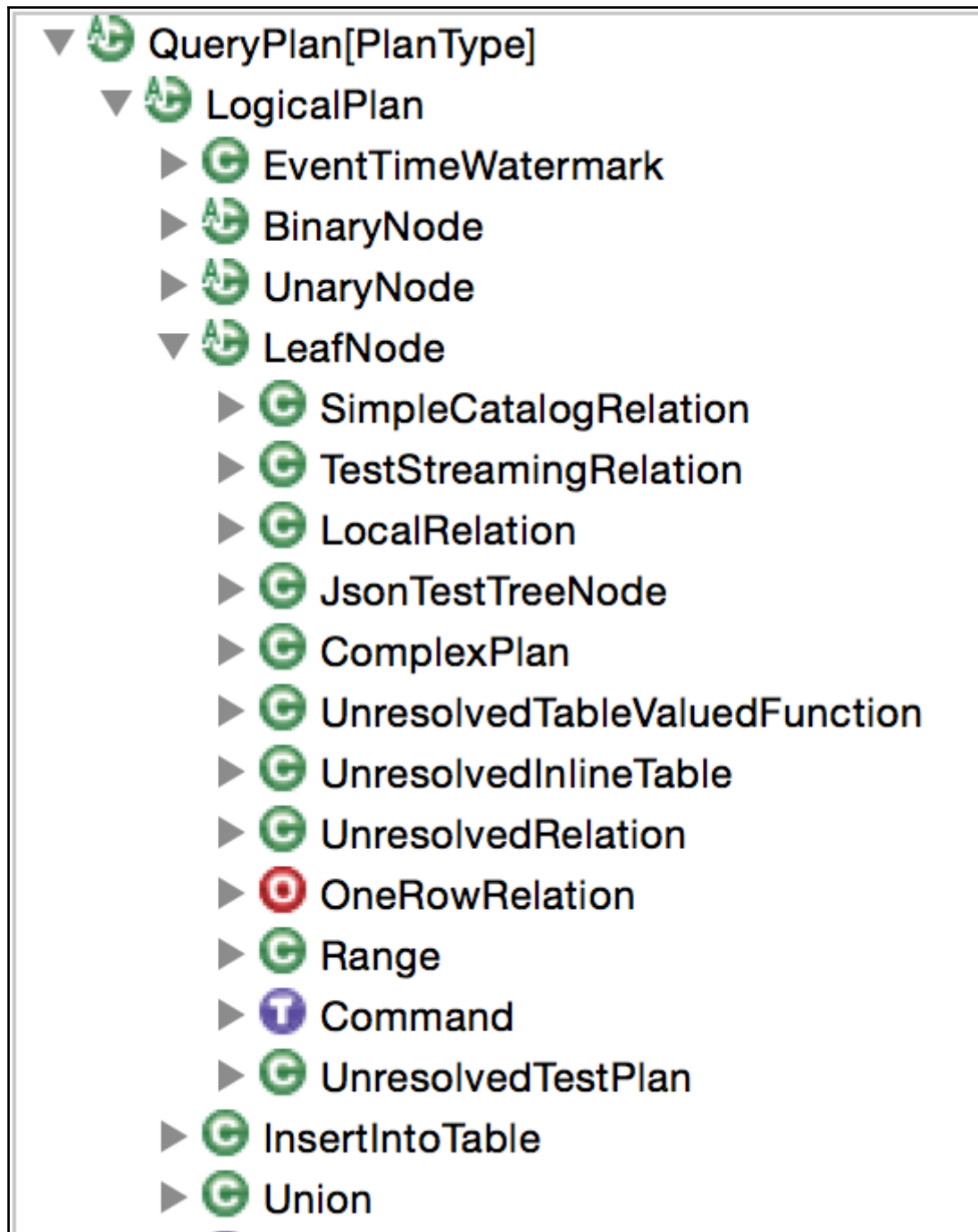








- ▶  TernaryExpression
- ▶  BinaryExpression
- ▶  UnaryExpression
- ▼  LeafExpression
  - ▶  BoundReference
  - ▶  NonFoldableLiteral
  - ▶  LeafMathExpression
  - ▶  NamePlaceholder
  - ▶  Literal
  - ▶  MutableExpression
  - ▶  NondeterministicExpression
  - ▶  OuterReference
  - ▶  Attribute
  - ▶  LambdaVariable
  - ▶  InputFileBlockLength
  - ▶  InputFileBlockStart
  - ▶  InputFileName
  - ▶  UnresolvedOrdinal
  - ▶  GetColumnByOrdinal
- ▼  Star
  - ▶  ResolvedStar
  - ▶  UnresolvedStar



# Project Tungsten

UnsafeRow
null bit set
values (fixed length)
values (variable length)

localhost:4040/storage/

spark 2.0.0 Jobs Stages Storage Environment Executors SQL

### Storage

RDDs

RDD Name	Storage Level	Cached Partitions	Fraction Cached	Size in Memory
LocalTableScan [value#1]	Memory Deserialized 1x Replicated	8	100%	998.4 KB
ParallelCollectionRDD	Memory Deserialized 1x Replicated	8	100%	3.8 MB

localhost:4040/jobs/

spark 2.0.0 Jobs Stages Storage Environment Executors SQL

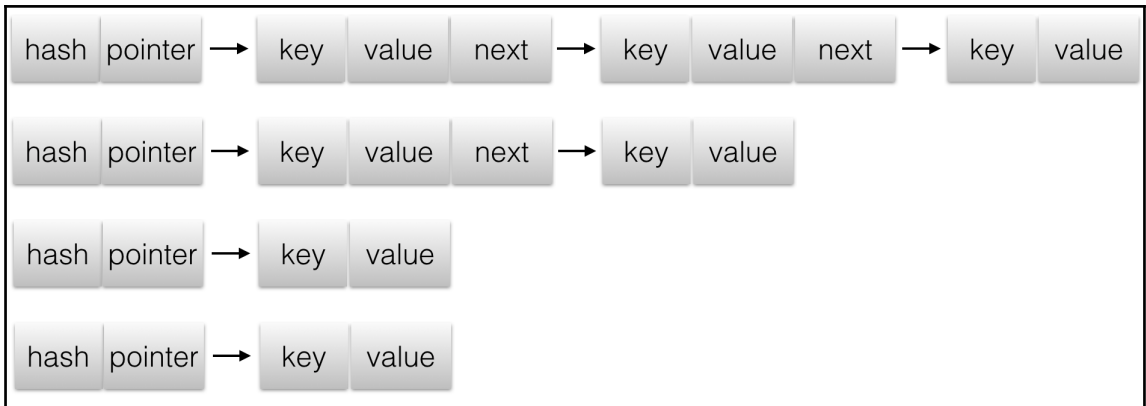
### Spark Jobs (?)

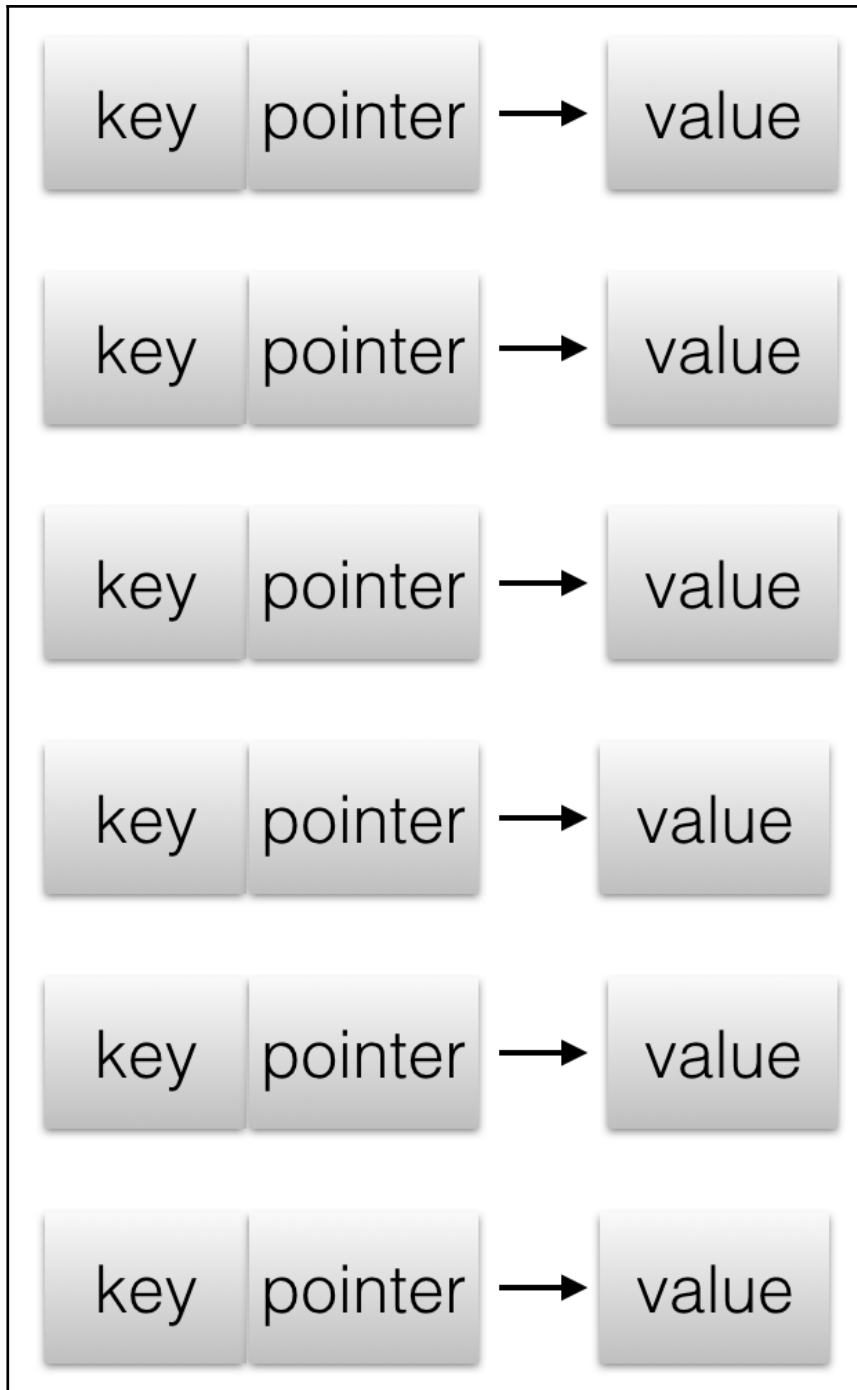
User: romeokienzler  
 Total Uptime: 1.4 min  
 Scheduling Mode: FIFO  
 Completed Jobs: 3

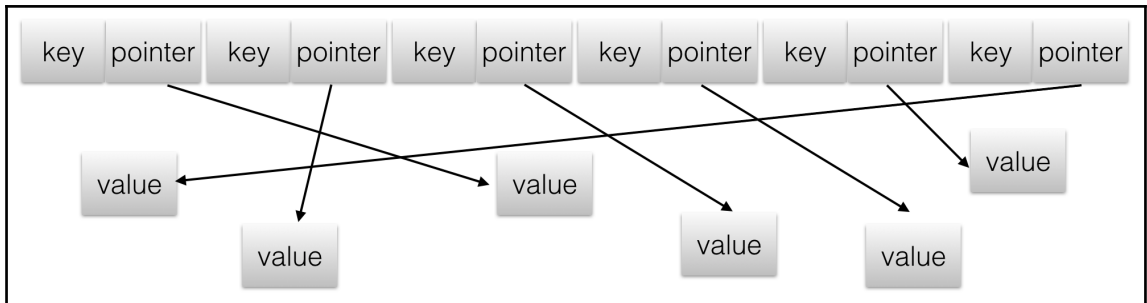
▶ Event Timeline

#### Completed Jobs (3)

Job Id	Description	Submitted	Duration
2	count at <console>:32	2017/01/04 07:04:43	0.7 s
1	count at <console>:28	2017/01/04 07:04:37	2 s
0	count at <console>:29	2017/01/04 07:04:16	0.5 s







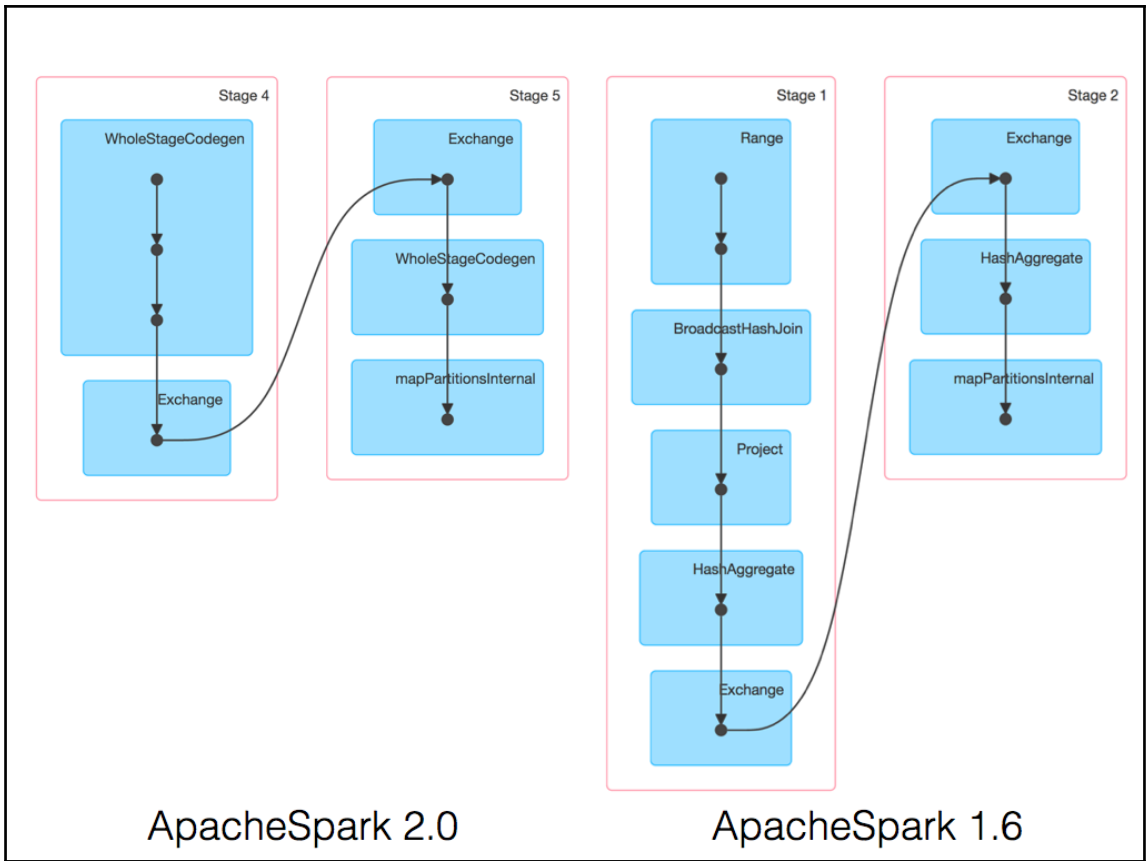
Row 1	Field 1	Field 2	Field 3
Row 2	Field 1	Field 2	Field 3
Row 3	Field 1	Field 2	Field 3

	Row 1	Row 2	Row 3
Column 1	Field 1	Field 1	Field 1
Column 2	Field 2	Field 2	Field 2
Column 3	Field 3	Field 3	Field 3

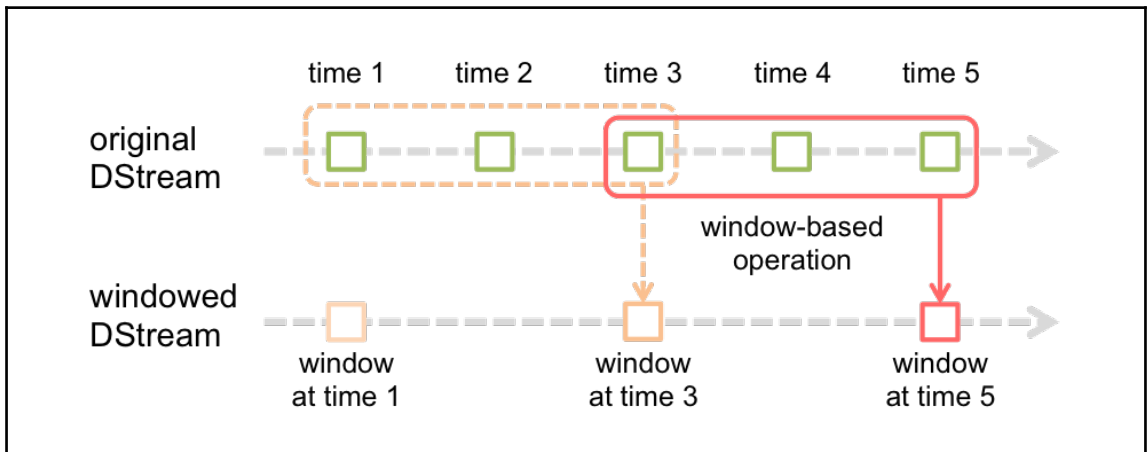
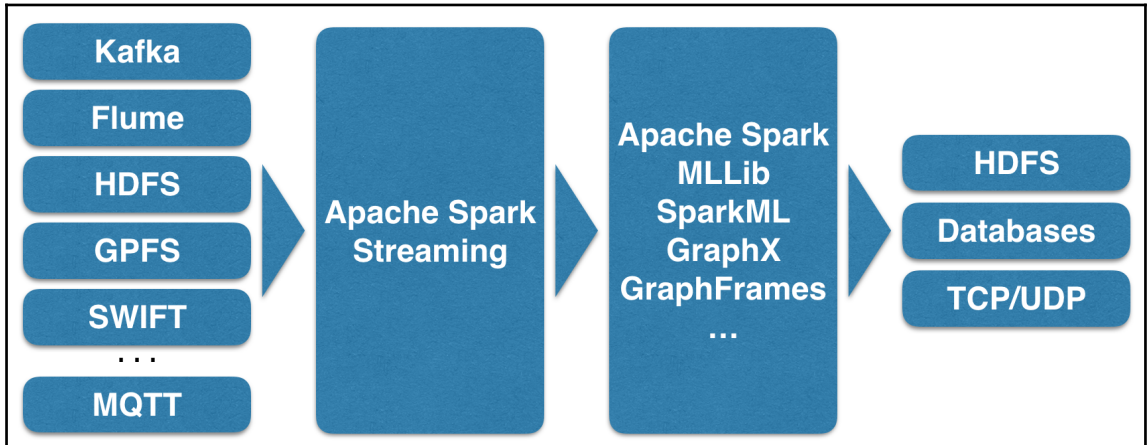
The screenshot shows a web browser window with the URL `localhost:4040/jobs/`. The browser's address bar shows the page title "Are @ - strings al..." and the path `http://192.168.1.1/...`. The page header includes the Apache Spark 2.0.0 logo and a navigation menu with the following items: Jobs, Stages, Storage, Environment, Executors, and SQL. The main content area is titled "Spark Jobs (?)". Below the title, the following information is displayed: User: romeokienzler, Total Uptime: 2.6 min, Scheduling Mode: FIFO, and Completed Jobs: 4. A link for "Event Timeline" is also present. The section "Completed Jobs (4)" contains a table with the following data:

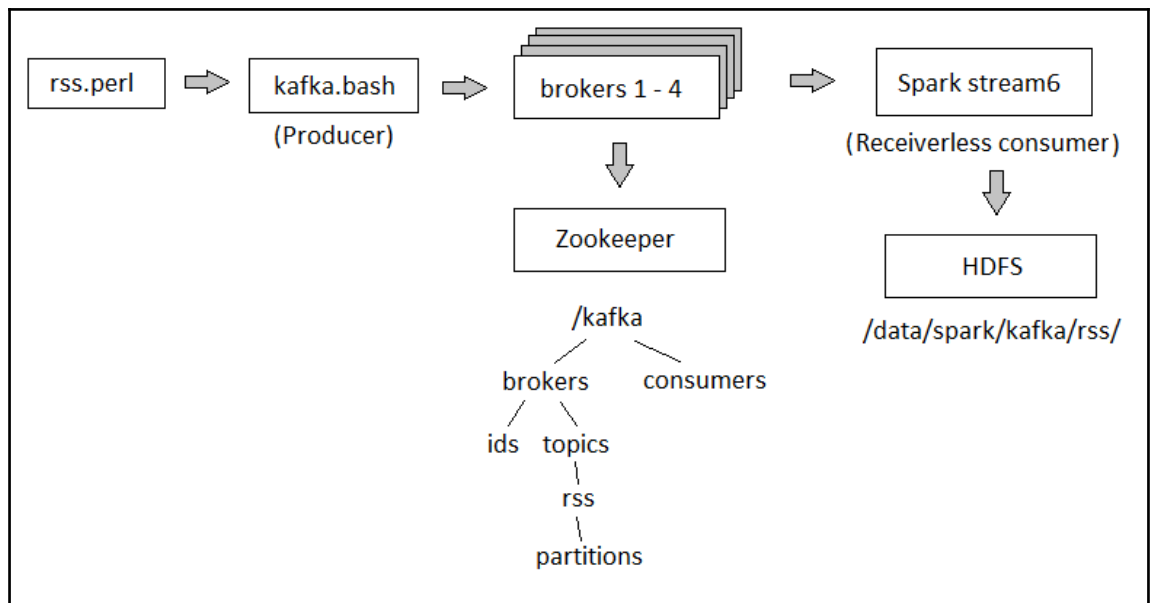
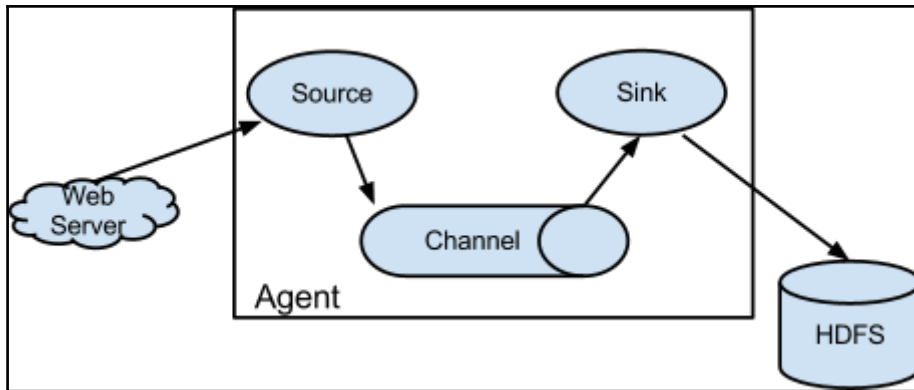
Job Id	Description	Submitted	Duration
3	<a href="#">count at &lt;console&gt;:24</a>	2017/01/04 13:11:14	0.5 s
2	<a href="#">run at ThreadPoolExecutor.java:1142</a>	2017/01/04 13:11:14	17 ms
1	<a href="#">count at &lt;console&gt;:24</a>	2017/01/04 13:09:25	31 s
0	<a href="#">run at ThreadPoolExecutor.java:1142</a>	2017/01/04 13:09:24	0.4 s





# Apache Spark Streaming





# Structured Streaming

```
Romeos-MacBook-Pro:chapter6 romeokienzler$ spark-shell --packages org.apache.bahir:spark-sql-streaming-mqtt_2.11:2.1.0,org.eclipse.paho:org.eclipse.paho.client.mqttv3:1.1.0
Ivy Default Cache set to: /Users/romeokienzler/.ivy2/cache
The jars for the packages stored in: /Users/romeokienzler/.ivy2/jars
:: loading settings :: url = jar:file:/Users/romeokienzler/Documents/runtimes/spark-2.1.0-bin-hadoop2.7/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
org.apache.bahir#spark-sql-streaming-mqtt_2.11 added as a dependency
org.eclipse.paho#org.eclipse.paho.client.mqttv3 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent:1.0
  confs: [default]
  found org.apache.bahir#spark-sql-streaming-mqtt_2.11:2.1.0 in local-m2-cache
  found org.apache.spark#spark-tags_2.11:2.1.0 in local-m2-cache
  found org.scalatest#scalatest_2.11:2.2.6 in local-m2-cache
  found org.scala-lang#scala-reflect;2.11.8 in local-m2-cache
  found org.scala-lang.modules#scala-xml_2.11;1.0.2 in local-m2-cache
  found org.spark-project.spark#unused;1.0.0 in local-m2-cache
  found org.eclipse.paho#org.eclipse.paho.client.mqttv3;1.1.0 in central
:: resolution report :: resolve 7237ms :: artifacts dl 8ms
  :: modules in use:
  org.apache.bahir#spark-sql-streaming-mqtt_2.11:2.1.0 from local-m2-cache in [default]
  org.apache.spark#spark-tags_2.11:2.1.0 from local-m2-cache in [default]
  org.eclipse.paho#org.eclipse.paho.client.mqttv3;1.1.0 from central in [default]
  org.scala-lang#scala-reflect;2.11.8 from local-m2-cache in [default]
  org.scala-lang.modules#scala-xml_2.11;1.0.2 from local-m2-cache in [default]
  org.scalatest#scalatest_2.11:2.2.6 from local-m2-cache in [default]
  org.spark-project.spark#unused;1.0.0 from local-m2-cache in [default]
-----
|         |         | modules |         | artifacts |
|         |         | search | dl | evicted |         | dl |
|-----|-----|-----|---|-----|-----|---|
| default | 7 | 1 | 1 | 0 | 7 | 0 |
-----
:: retrieving :: org.apache.spark#spark-submit-parent
  confs: [default]
  0 artifacts copied, 7 already retrieved (0kB/9ms)
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). For SparkR, use setLogLevel(newLevel).
17/07/10 08:37:49 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/07/10 08:38:00 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Spark context Web UI available at http://192.168.0.100:4040
Spark context available as 'sc' (master = local[*], app id = local-1499668675024).
Spark session available as 'spark'.
Welcome to

  ____
 /  __ \
/   /  \
/_____/

 version 2.1.0

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_65)
Type in expressions to have them evaluated.
Type :help for more information.

scala> █
```

```
scala> val mqtt_host = "pcoyha.messaging.internetofthings.ibmcloud.com"
mqtt_host: String = pcoyha.messaging.internetofthings.ibmcloud.com

scala> val org = "pcoyha"
org: String = pcoyha

scala> val apiKey = "a-pcoyha-0aigc1k8ub"
apiKey: String = a-pcoyha-0aigc1k8ub

scala> val apiToken = "&wuypVX2yNgVLAclR8"
apiToken: String = &wuypVX2yNgVLAclR8

scala> var randomSessionId = scala.util.Random.nextInt(10000)
randomSessionId: Int = 8270

scala>

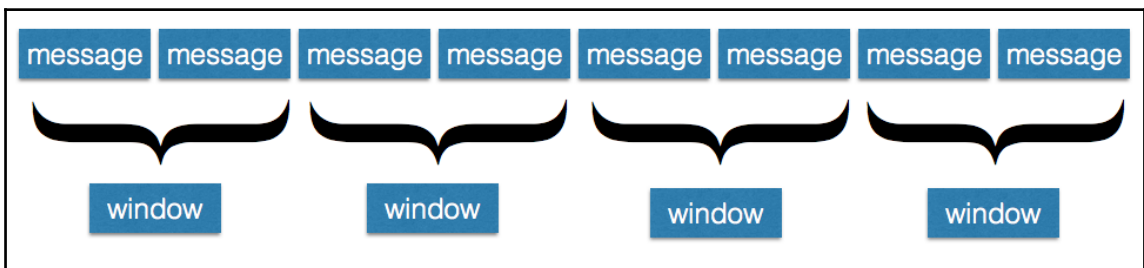
scala> :paste
// Entering paste mode (ctrl-D to finish)

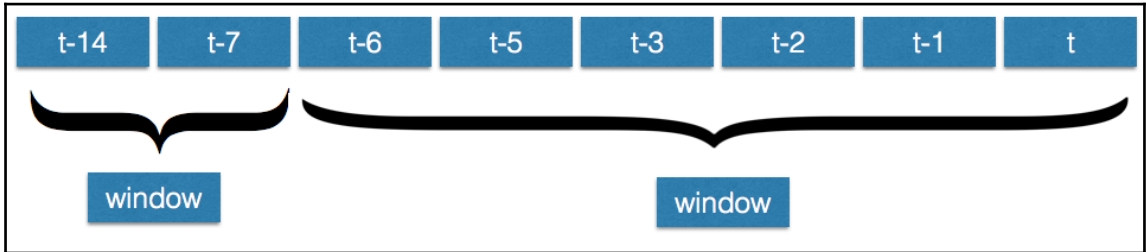
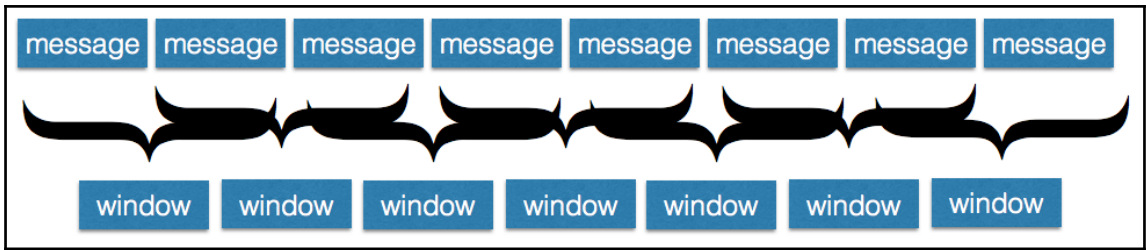
val df = spark.readStream
  .format("org.apache.bahir.sql.streaming.mqtt.MQTTStreamSourceProvider")
  .option("username", apiKey)
  .option("password", apiToken)
  .option("clientId", "a:"+org+": "+apiKey+randomSessionId)
  .option("topic", "iot-2/type/WashingMachine/id/Washer01/evt/voltage/fmt/json")
  .load("tcp://" + mqtt_host + ":1883")

// Exiting paste mode, now interpreting.

df: org.apache.spark.sql.DataFrame = [value: string, timestamp: timestamp]
```

```
scala> val query = df.writeStream
  |   .outputMode("append")
  |   .format("console")
  |   .start()
query: org.apache.spark.sql.streaming.StreamingQuery = Streaming Query [id = a2377c24-c274-476e-bc2b-07d57bab1877, runId = 387ca22f-138c-4456-9243-9218766a6f13] [state = ACTIVE]
```





Batch: 332	
value	timestamp
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...

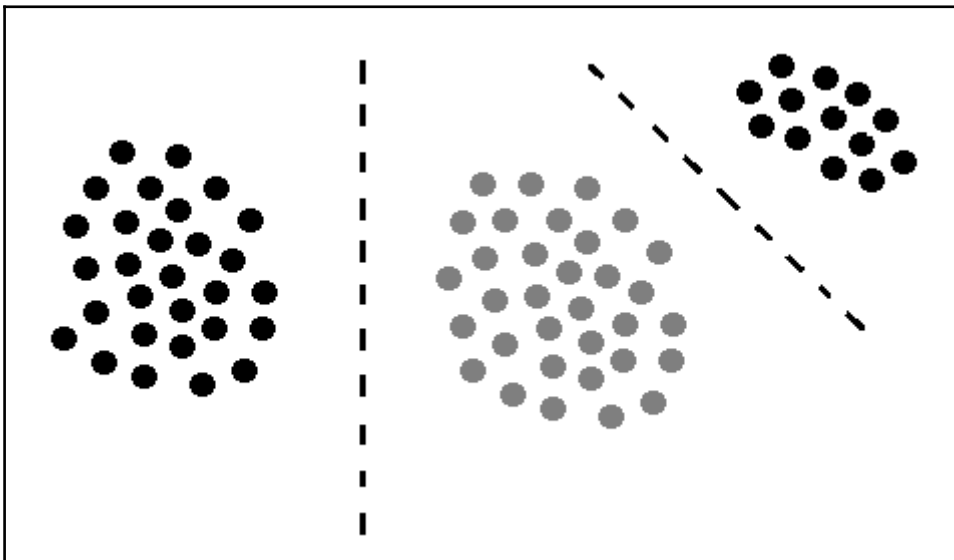
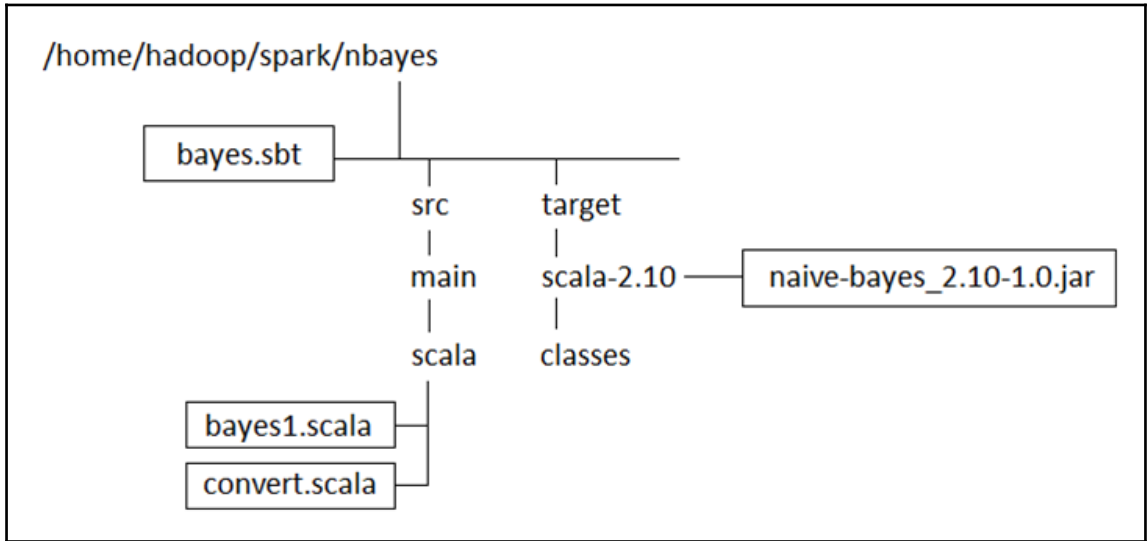
  

Batch: 333	
value	timestamp
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...

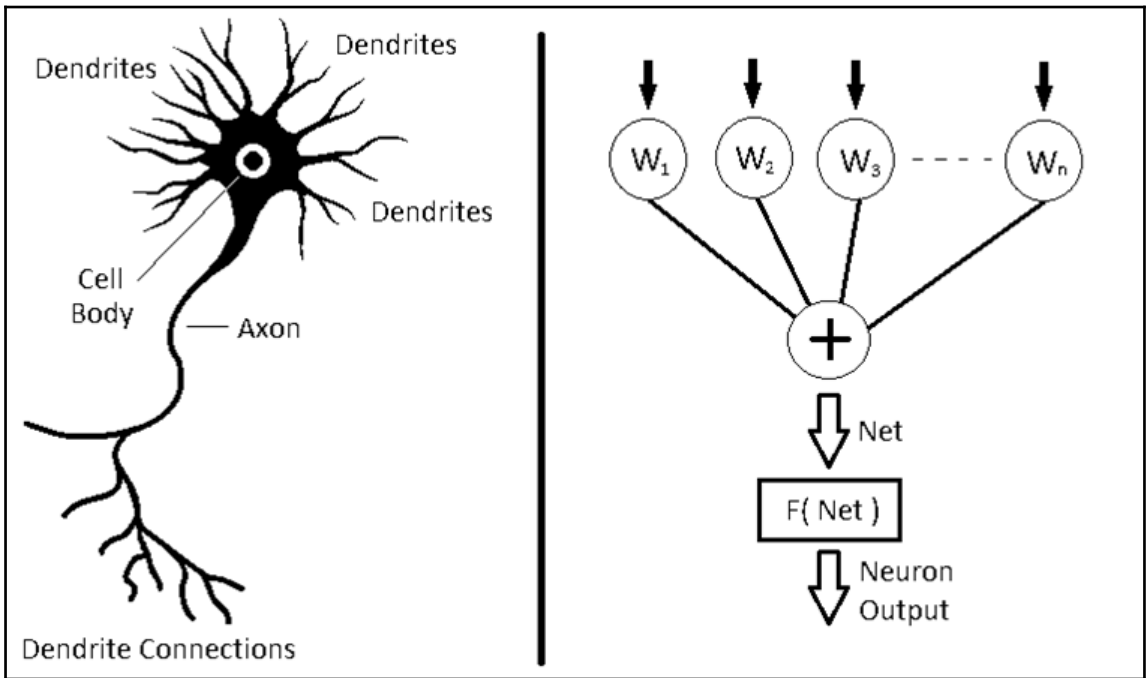
Batch: 334	
value	timestamp
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...
{\"d\":{\"voltage\":2...}	2017-04-26 05:31:...

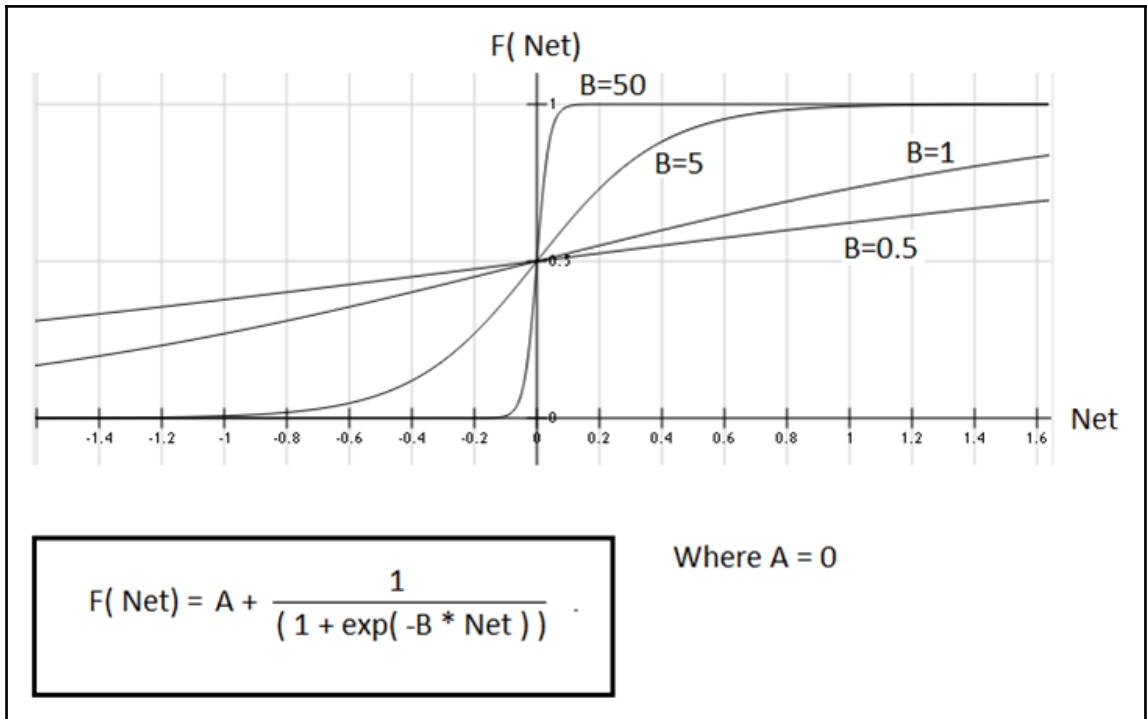
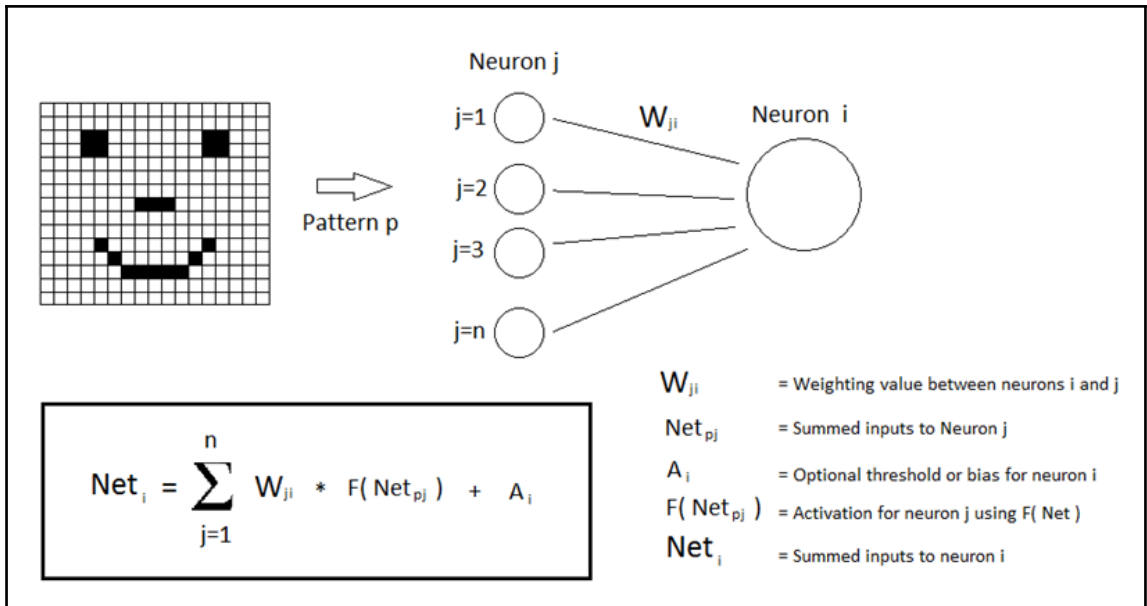
# Apache Spark MLlib

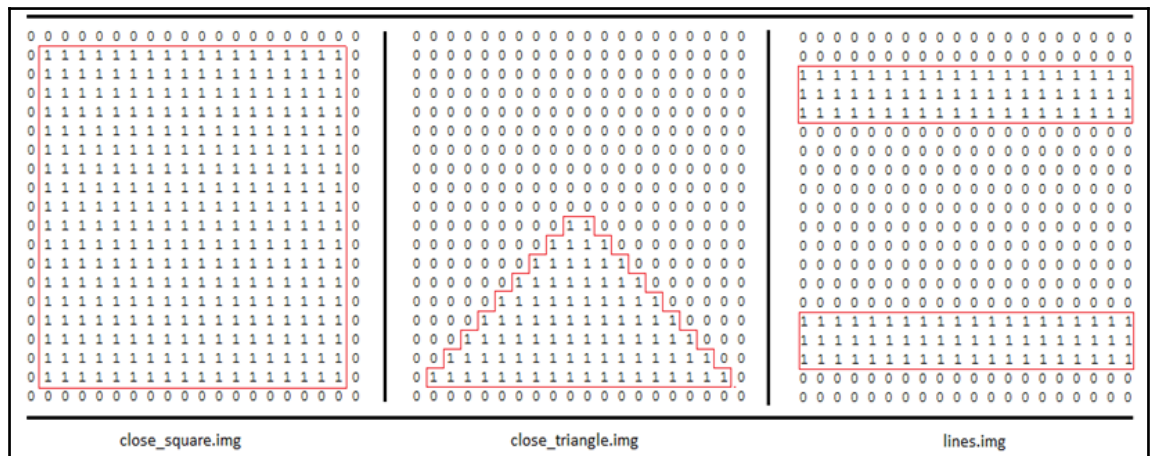
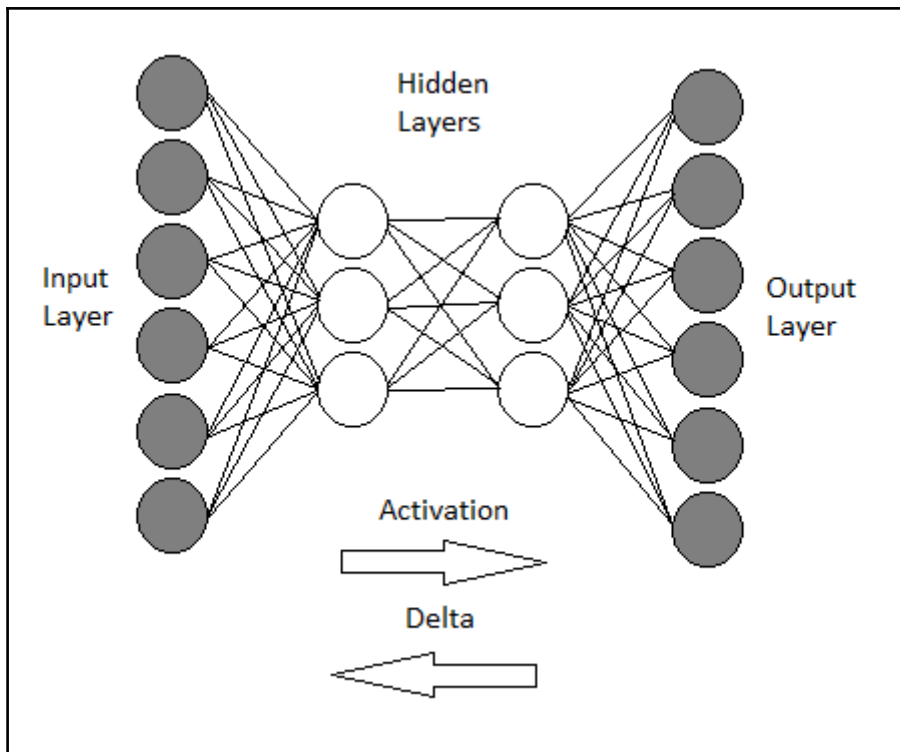


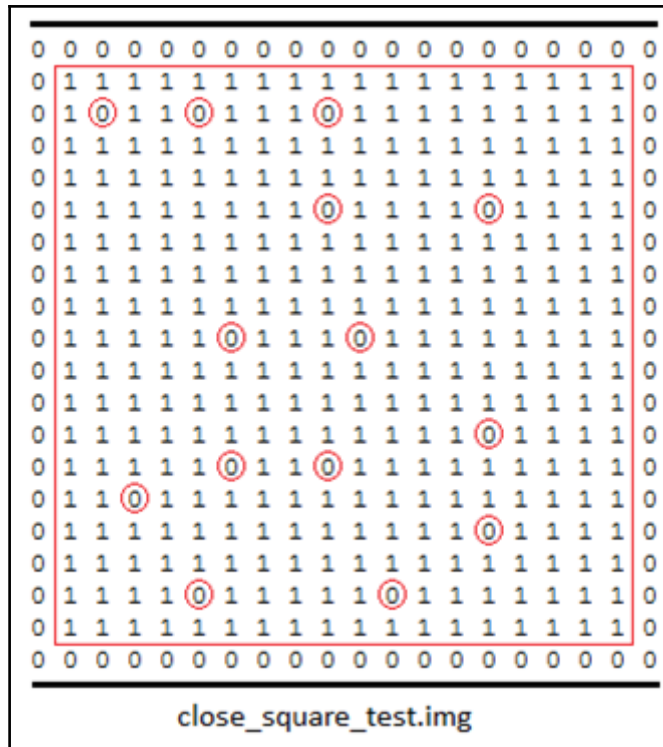


$$P(\text{Spam} | \text{Buy}) = \frac{P(\text{Buy} | \text{Spam}) * P(\text{Spam})}{P(\text{Buy} | \text{Spam}) * P(\text{Spam}) + P(\text{Buy} | \text{Not Spam}) * P(\text{Not Spam})}$$









hc2nn.semtech-solutions.co.nz:19080
Search

**Spark Master at spark://hc2nn.semtech-solutions.co.nz:8077**  
1.3.0-SNAPSHOT

URL: spark://hc2nn.semtech-solutions.co.nz:8077  
 REST URL: spark://hc2nn.semtech-solutions.co.nz:8066 (cluster mode)  
 Workers: 4  
 Cores: 8 Total, 8 Used  
 Memory: 3.1 GB Total, 2.7 GB Used  
 Applications: 1 Running, 2 Completed  
 Drivers: 0 Running, 0 Completed  
 Status: ALIVE

### Workers

Worker Id	Address	State	Cores	Memory
<a href="#">worker-20150422141206-hc2r1m2.semtech-solutions.co.nz-8078</a>	hc2r1m2.semtech-solutions.co.nz:8078	ALIVE	2 (2 Used)	783.0 MB (700.0 MB Used)
<a href="#">worker-20150422141207-hc2r1m4.semtech-solutions.co.nz-8078</a>	hc2r1m4.semtech-solutions.co.nz:8078	ALIVE	2 (2 Used)	783.0 MB (700.0 MB Used)
<a href="#">worker-20150422141208-hc2r1m1.semtech-solutions.co.nz-8078</a>	hc2r1m1.semtech-solutions.co.nz:8078	ALIVE	2 (2 Used)	783.0 MB (700.0 MB Used)
<a href="#">worker-20150422141208-hc2r1m3.semtech-solutions.co.nz-8078</a>	hc2r1m3.semtech-solutions.co.nz:8078	ALIVE	2 (2 Used)	783.0 MB (700.0 MB Used)

### Running Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
<a href="#">app-20150422143345-0002</a>	ANN 1	8	700.0 MB	2015/04/22 14:33:45	hadoop	RUNNING	3 s

### Completed Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
<a href="#">app-20150422142620-0001</a>	ANN 1	8	700.0 MB	2015/04/22 14:26:20	hadoop	FINISHED	1.5 min
<a href="#">app-20150422142411-0000</a>	ANN 1	8	700.0 MB	2015/04/22 14:24:11	hadoop	FINISHED	1.8 min

hc2nn.semtech-solutions.co.nz:19080/app/?appId=app-20150422143345-0002


**Spark** 1.3.0-SNAPSHOT **Application: ANN 1**

ID: app-20150422143345-0002  
Name: ANN 1  
User: hadoop  
Cores: 100 (8 granted, 92 left)  
Executor Memory: 700.0 MB  
Submit Date: Wed Apr 22 14:33:45 NZST 2015  
State: RUNNING  
[Application Detail UI](#)

**Executor Summary**

ExecutorID	Worker	Cores	Memory	State	Logs
2	<a href="#">worker-20150422141208-hc2r1m1.semtech-solutions.co.nz-8078</a>	2	700	RUNNING	<a href="#">stdout</a> <a href="#">stderr</a>
1	<a href="#">worker-20150422141208-hc2r1m3.semtech-solutions.co.nz-8078</a>	2	700	RUNNING	<a href="#">stdout</a> <a href="#">stderr</a>
3	<a href="#">worker-20150422141206-hc2r1m2.semtech-solutions.co.nz-8078</a>	2	700	RUNNING	<a href="#">stdout</a> <a href="#">stderr</a>
0	<a href="#">worker-20150422141207-hc2r1m4.semtech-solutions.co.nz-8078</a>	2	700	RUNNING	<a href="#">stdout</a> <a href="#">stderr</a>

← hc2r1m2.semtech-solutions.co.nz:19081
🔍 Search


Spark Worker at hc2r1m2.semtech-solutions.co.nz:8078

**ID:** worker-20150222113812-hc2r1m2.semtech-solutions.co.nz-8078  
**Master URL:** spark://hc2nn.semtech-solutions.co.nz:8077  
**Cores:** 2 (2 Used)  
**Memory:** 783.0 MB (700.0 MB Used)

[Back to Master](#)

### Running Executors (1)

ExecutorID	Cores	State	Memory	Job Details	Logs
2	2	LOADING	700.0 MB	<b>ID:</b> app-20150222134613-0002 <b>Name:</b> ANN 1 <b>User:</b> root	<a href="#">stdout stderr</a>

### Finished Executors (2)

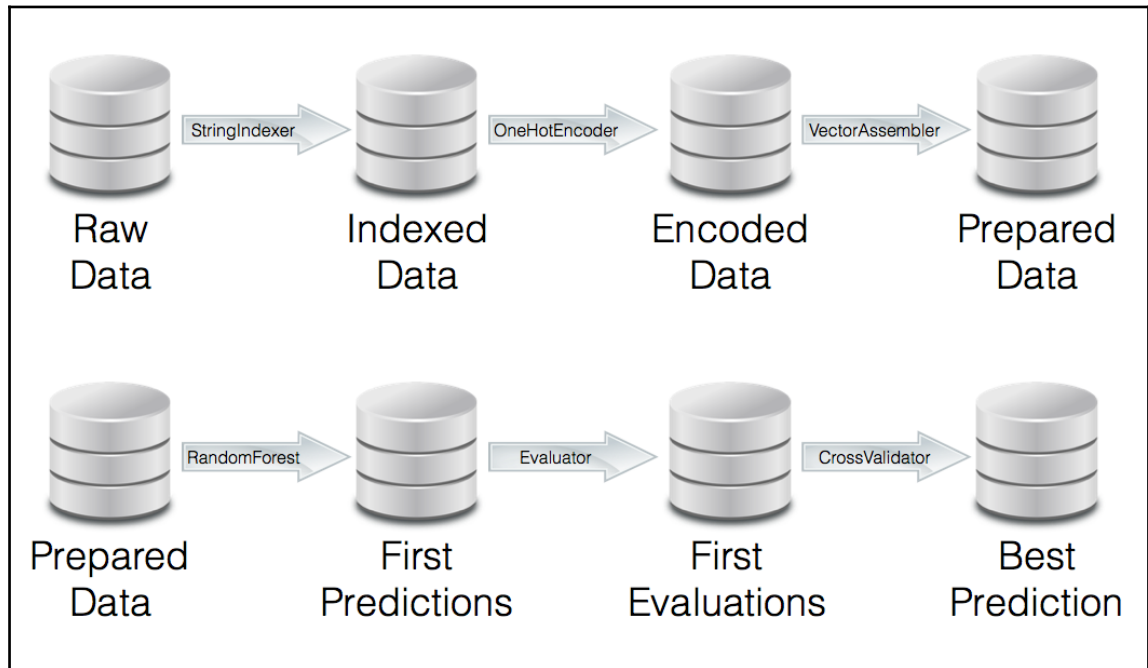
ExecutorID	Cores	State	Memory	Job Details	Logs
2	2	KILLED	700.0 MB	<b>ID:</b> app-20150222125002-0000 <b>Name:</b> ANN 1 <b>User:</b> root	<a href="#">stdout stderr</a>
2	2	KILLED	700.0 MB	<b>ID:</b> app-20150222125935-0001 <b>Name:</b> ANN 1 <b>User:</b> root	<a href="#">stdout stderr</a>

$$\arg \min_{\mathbf{s}} \sum_{i=1}^K \sum_{\mathbf{x} \in \mathbf{s}_i} \|\mathbf{x} - \mathbf{B}_i\|^2$$

where  $\mathbf{B}_i$  is the mean of members of  $\mathbf{s}_i$



# Apache SparkML



train_date.csv	11 Aug 2016, 18:12	2.89 GB	comma-separated values
train_categorical.csv	11 Aug 2016, 17:50	2.68 GB	comma-separated values
train_numeric.csv	11 Aug 2016, 17:41	2.14 GB	comma-separated values
train_date.parquet	26 Apr 2017, 16:52	890.5 MB	Document
train_numeric.parquet	26 Apr 2017, 16:27	257.4 MB	Document
train_categorical.parquet	26 Apr 2017, 17:06	40.6 MB	Document

```

scala> var df_numeric = spark.read.parquet(basePath+"train_numeric.parquet")
df_numeric: org.apache.spark.sql.DataFrame = [Id: int, L0_S0_F0: double ... 968 more fields]

scala>

scala> var df_date = spark.read.parquet(basePath+"train_date.parquet")
df_date: org.apache.spark.sql.DataFrame = [Id: int, L0_S0_D1: double ... 1155 more fields]

scala>

scala> var df_categorical = spark.read.parquet(basePath+"train_categorical.parquet")
df_categorical: org.apache.spark.sql.DataFrame = [Id: int, L0_S1_F25: string ... 2139 more fields]
  
```

```
[scala> dfcat.select("L0_S22_F545").distinct.show
+-----+
|L0_S22_F545|
+-----+
|          T16|
|  T12582912|
|          null|
|    T48576|
|  T16777232|
|          T512|
|    T589824|
|    T1372|
|          T8|
|  T16777557|
|          T32|
|    T6553|
|  T-18748192|
|          T96|
+-----+
```

```
scala> dfnum.show
```

```
+-----+-----+-----+-----+
| Id|L0_S0_F0|L0_S0_F2|L0_S0_F4|Response|
+-----+-----+-----+-----+
|  4|    0.03|-0.034|-0.197|      0|
|  6|   null|   null|   null|      0|
|  7|   0.088|  0.086|  0.003|      0|
|  9|  -0.036|-0.064|  0.294|      0|
| 11|  -0.055|-0.086|  0.294|      0|
| 13|   0.003|  0.019|  0.294|      0|
| 14|   null|   null|   null|      0|
| 16|   null|   null|   null|      0|
| 18|  -0.016|-0.041|-0.179|      0|
| 23|   null|   null|   null|      0|
| 26|   0.016|  0.093|-0.015|      0|
| 27|  -0.062|-0.153|-0.197|      0|
| 28|  -0.075|-0.093|  0.367|      0|
| 31|  -0.003|-0.093|-0.161|      0|
| 34|  -0.016|-0.138|-0.197|      0|
| 38|   0.252|  0.25|  0.003|      0|
| 41|   null|   null|   null|      0|
| 44|  -0.016|-0.041|  0.003|      0|
| 47|   null|   null|   null|      0|
| 49|   0.088|  0.033|  0.33|      0|
+-----+-----+-----+-----+
```

```
only showing top 20 rows
```

```
root
```

```
|-- label: integer (nullable = true)
|-- L0_S22_F545: string (nullable = true)
|-- L0_S0_F0: double (nullable = true)
|-- L0_S0_F2: double (nullable = true)
|-- L0_S0_F4: double (nullable = true)
|-- L0_S22_F545Index: double (nullable = true)
```

```
scala> indexed.select("L0_S22_F545","L0_S22_F545Index").distinct.show
```

```
+-----+-----+
|L0_S22_F545|L0_S22_F545Index|
+-----+-----+
|  T12582912|          6.0|
|      T1372|         10.0|
|       T16 |          7.0|
|       T32 |          9.0|
|   T48576 |          2.0|
|T-18748192|          8.0|
|      NA  |          0.0|
|T16777557 |          1.0|
|T16777232 |         11.0|
|       T8 |          4.0|
|      T512|          3.0|
|   T589824|         13.0|
|      T6553|         12.0|
|       T96 |          5.0|
+-----+-----+
```

```
scala> encoded.select("L0_S22_F545Index","L0_S22_F545Vec").distinct.show
```

L0_S22_F545Index	L0_S22_F545Vec
11.0	(13, [11], [1.0])
2.0	(13, [2], [1.0])
8.0	(13, [8], [1.0])
3.0	(13, [3], [1.0])
10.0	(13, [10], [1.0])
6.0	(13, [6], [1.0])
7.0	(13, [7], [1.0])
12.0	(13, [12], [1.0])
9.0	(13, [9], [1.0])
4.0	(13, [4], [1.0])
5.0	(13, [5], [1.0])
0.0	(13, [0], [1.0])
1.0	(13, [1], [1.0])
13.0	(13, [], [])

```
scala> assembled.show
```

label	L0_S22_F545	L0_S0_F0	L0_S0_F2	L0_S0_F4	L0_S22_F545Index	L0_S22_F545Vec	features
0	NA	0.03	-0.034	-0.197	0.0	(13, [0], [1.0])	(16, [0, 13, 14, 15], ...
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])
0	NA	0.088	0.086	0.003	0.0	(13, [0], [1.0])	(16, [0, 13, 14, 15], ...
0	NA	-0.036	-0.064	0.294	0.0	(13, [0], [1.0])	(16, [0, 13, 14, 15], ...
0	NA	-0.055	-0.086	0.294	0.0	(13, [0], [1.0])	(16, [0, 13, 14, 15], ...
0	NA	0.003	0.019	0.294	0.0	(13, [0], [1.0])	(16, [0, 13, 14, 15], ...
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])
0	NA	-0.016	-0.041	-0.179	0.0	(13, [0], [1.0])	(16, [0, 13, 14, 15], ...
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])
0	NA	0.016	0.093	-0.015	0.0	(13, [0], [1.0])	(16, [0, 13, 14, 15], ...
0	NA	-0.062	-0.153	-0.197	0.0	(13, [0], [1.0])	(16, [0, 13, 14, 15], ...
0	NA	-0.075	-0.093	0.367	0.0	(13, [0], [1.0])	(16, [0, 13, 14, 15], ...
0	NA	-0.003	-0.093	-0.161	0.0	(13, [0], [1.0])	(16, [0, 13, 14, 15], ...
0	NA	-0.016	-0.138	-0.197	0.0	(13, [0], [1.0])	(16, [0, 13, 14, 15], ...
0	NA	0.252	0.25	0.003	0.0	(13, [0], [1.0])	(16, [0, 13, 14, 15], ...
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])
0	NA	-0.016	-0.041	0.003	0.0	(13, [0], [1.0])	(16, [0, 13, 14, 15], ...
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])
0	NA	0.088	0.033	0.33	0.0	(13, [0], [1.0])	(16, [0, 13, 14, 15], ...

only showing top 20 rows

```
scala> assembled.select("features").first.get(0)
res27: Any = (16, [0,13,14,15], [1.0,0.03,-0.034,-0.197])
```

```
scala> transformed.show
```

label	L0_S22_F545	L0_S0_F0	L0_S0_F2	L0_S0_F4	L0_S22_F545Index	L0_S22_F545Vec	features
0	NA	0.03	-0.034	-0.197	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])
0	NA	0.088	0.086	0.003	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...
0	NA	-0.036	-0.064	0.294	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...
0	NA	-0.055	-0.086	0.294	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...
0	NA	0.003	0.019	0.294	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])
0	NA	-0.016	-0.041	-0.179	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])
0	NA	0.016	0.093	-0.015	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...
0	NA	-0.062	-0.153	-0.197	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...
0	NA	-0.075	-0.093	0.367	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...
0	NA	-0.003	-0.093	-0.161	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...
0	NA	-0.016	-0.138	-0.197	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...
0	NA	0.252	0.25	0.003	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])
0	NA	-0.016	-0.041	0.003	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])
0	NA	0.088	0.033	0.33	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...

only showing top 20 rows

```
scala> result.show
```

label	L0_S22_F545	L0_S0_F0	L0_S0_F2	L0_S0_F4	L0_S22_F545Index	L0_S22_F545Vec	features	rawPrediction	probability	prediction
0	NA	0.03	-0.034	-0.197	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...	[19.8764711847913...	[0.99382355923956...	0.0
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])	[19.8734515671497...	[0.99367257835748...	0.0
0	NA	0.088	0.086	0.003	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...	[19.8936982048582...	[0.99468491024291...	0.0
0	NA	-0.036	-0.064	0.294	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...	[19.9103667119433...	[0.99551833559716...	0.0
0	NA	-0.055	-0.086	0.294	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...	[19.9128187397603...	[0.99564093698801...	0.0
0	NA	0.003	0.019	0.294	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...	[19.9021809064659...	[0.99510904532329...	0.0
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])	[19.8734515671497...	[0.99367257835748...	0.0
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])	[19.8734515671497...	[0.99367257835748...	0.0
0	NA	-0.016	-0.041	-0.179	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...	[19.8762685936784...	[0.99381342968392...	0.0
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])	[19.8734515671497...	[0.99367257835748...	0.0
0	NA	0.016	0.093	-0.015	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...	[19.8893959144095...	[0.99419679572047...	0.0
0	NA	-0.062	-0.153	-0.197	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...	[19.8900653890112...	[0.99450326945056...	0.0
0	NA	-0.075	-0.093	0.367	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...	[19.9155130528803...	[0.99577565264401...	0.0
0	NA	-0.003	-0.093	-0.161	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...	[19.8830065488786...	[0.99415032744393...	0.0
0	NA	-0.016	-0.138	-0.197	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...	[19.8787059668007...	[0.99393529834003...	0.0
0	NA	0.252	0.25	0.003	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...	[19.8899480526405...	[0.99449740263202...	0.0
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])	[19.8734515671497...	[0.99367257835748...	0.0
0	NA	-0.016	-0.041	0.003	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...	[19.8929530805276...	[0.99464769432638...	0.0
0	NA	0.0	0.0	0.0	0.0	(13, [0], [1.0])	(16, [0], [1.0])	[19.8734515671497...	[0.99367257835748...	0.0
0	NA	0.088	0.033	0.33	0.0	(13, [0], [1.0])	(16, [0,13,14,15], ...	[19.9092368840784...	[0.99546184420392...	0.0

only showing top 20 rows

```
scala> var aucTraining = evaluator.evaluate(result, evaluatorParamMap)
aucTraining: Double = 0.5424418446501833
```

```
[scala> evaluator.evaluate(newPredictions, evaluatorParamMap)
res6: Double = 0.5362224872557545
```

```
scala> rfStage.getNumTrees
res1: Int = 5
```

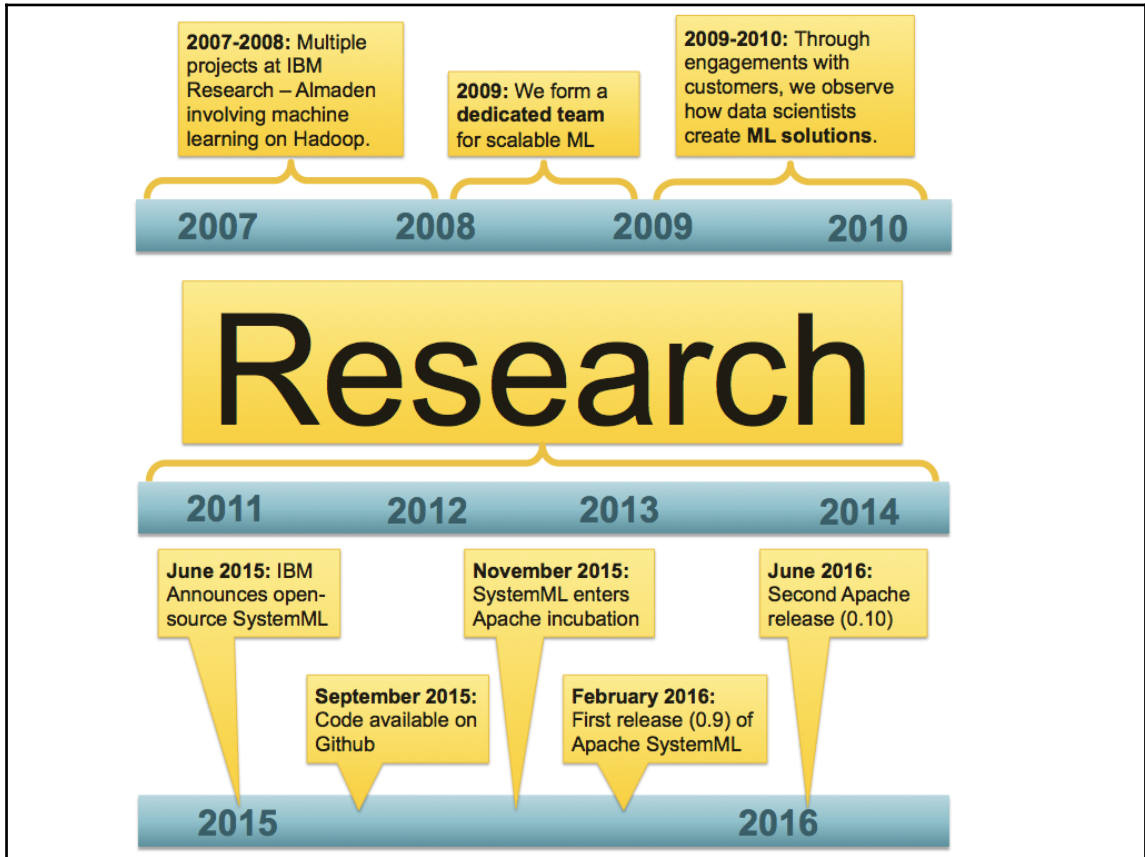
```
scala> rfStage.getFeatureSubsetStrategy
res2: String = auto
```

```
scala> rfStage.getImpurity
res3: String = entropy
```

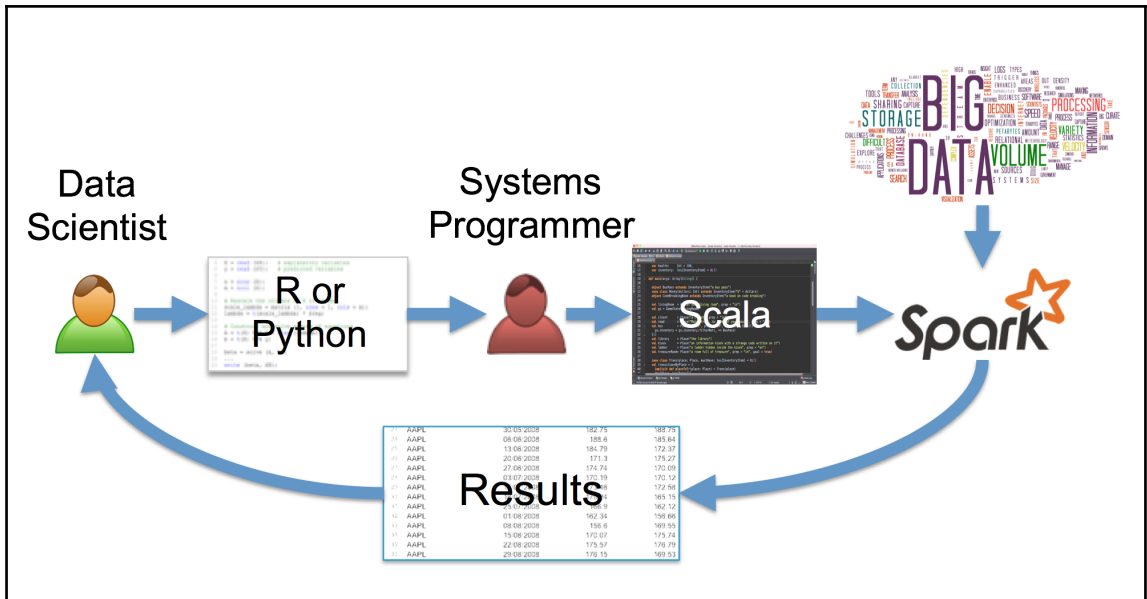
```
scala> rfStage.getMaxBins
res4: Int = 5
```

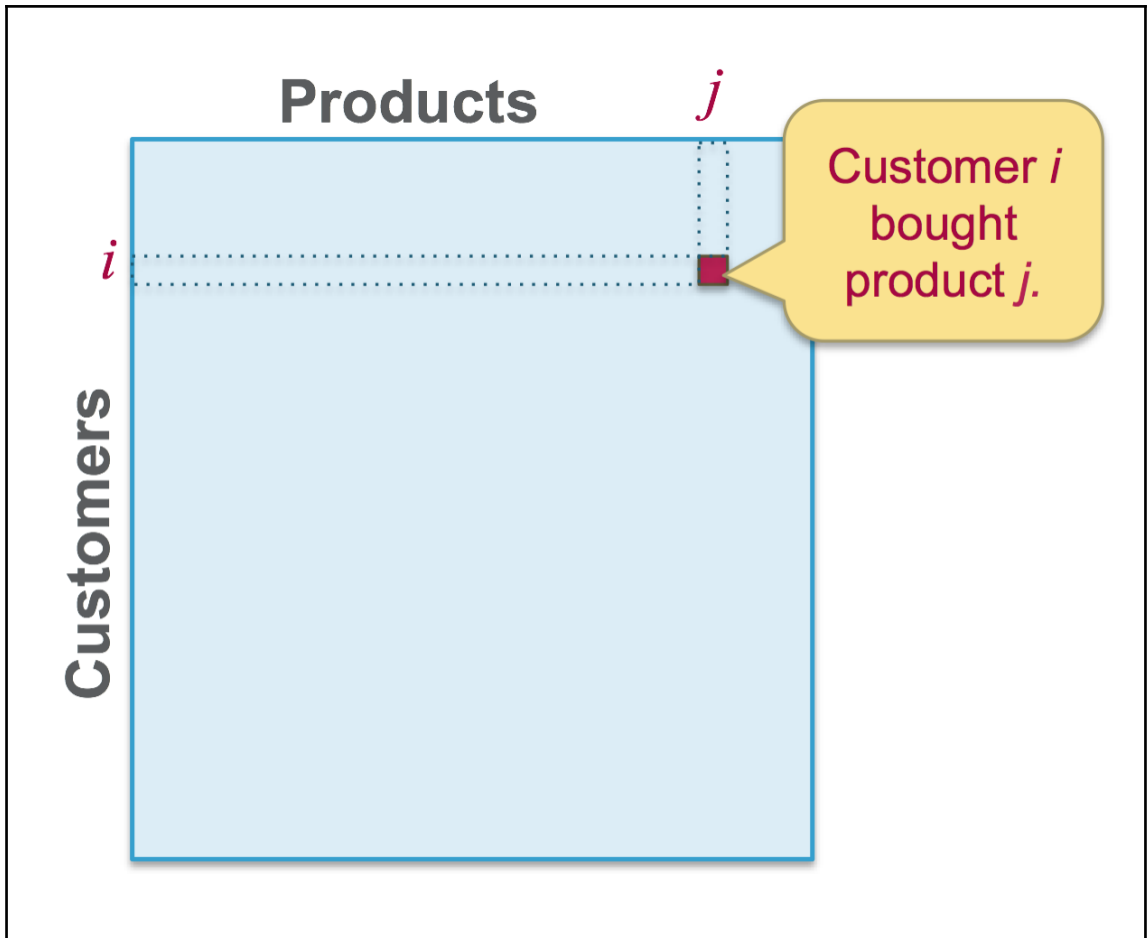
```
[scala> rfStage.getMaxDepth
res5: Int = 5
```

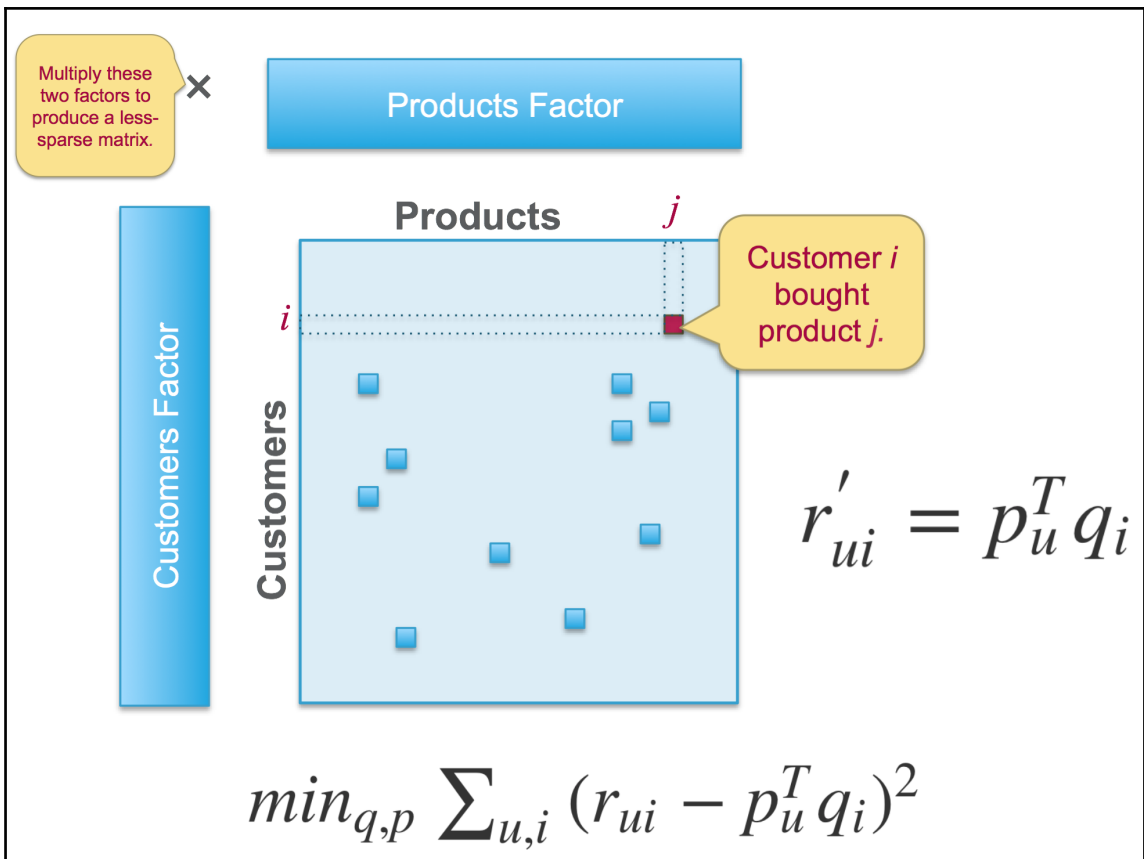
# Apache SystemML

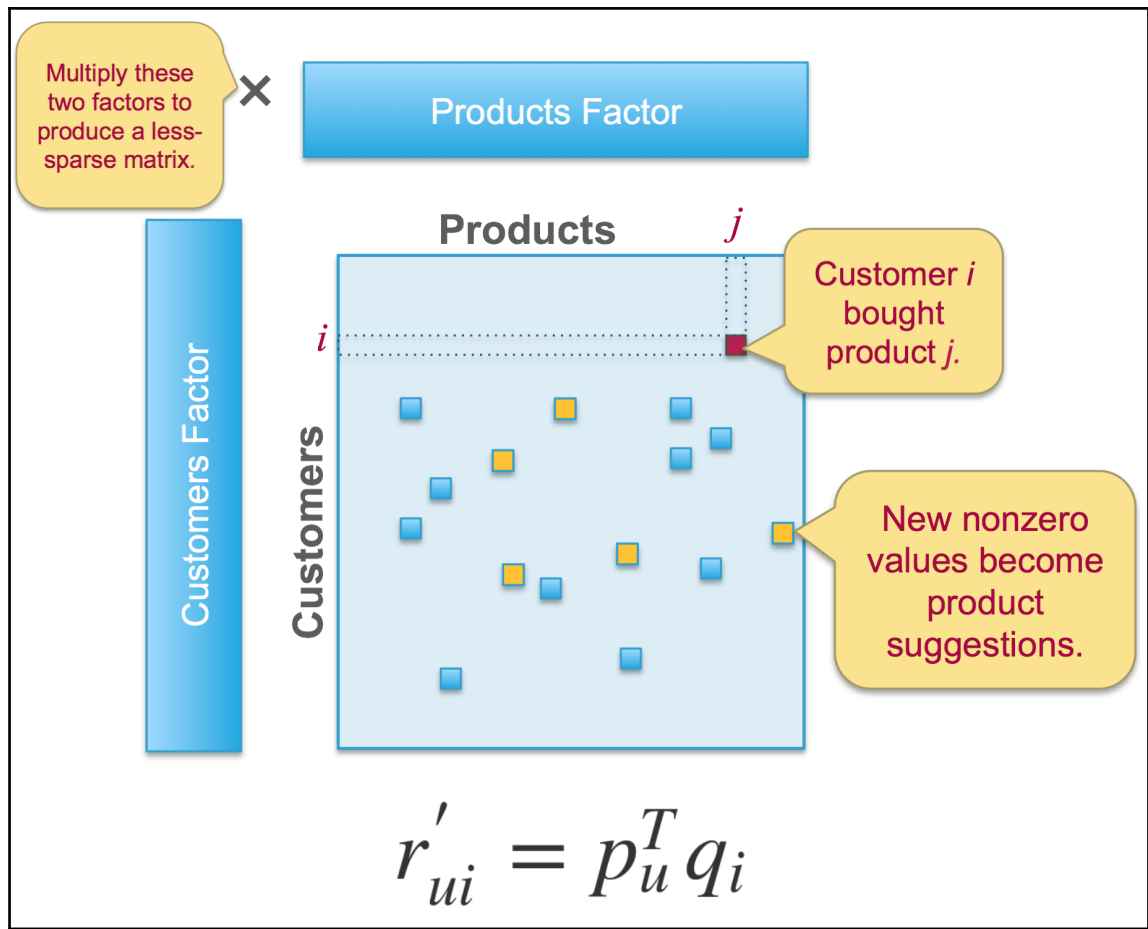






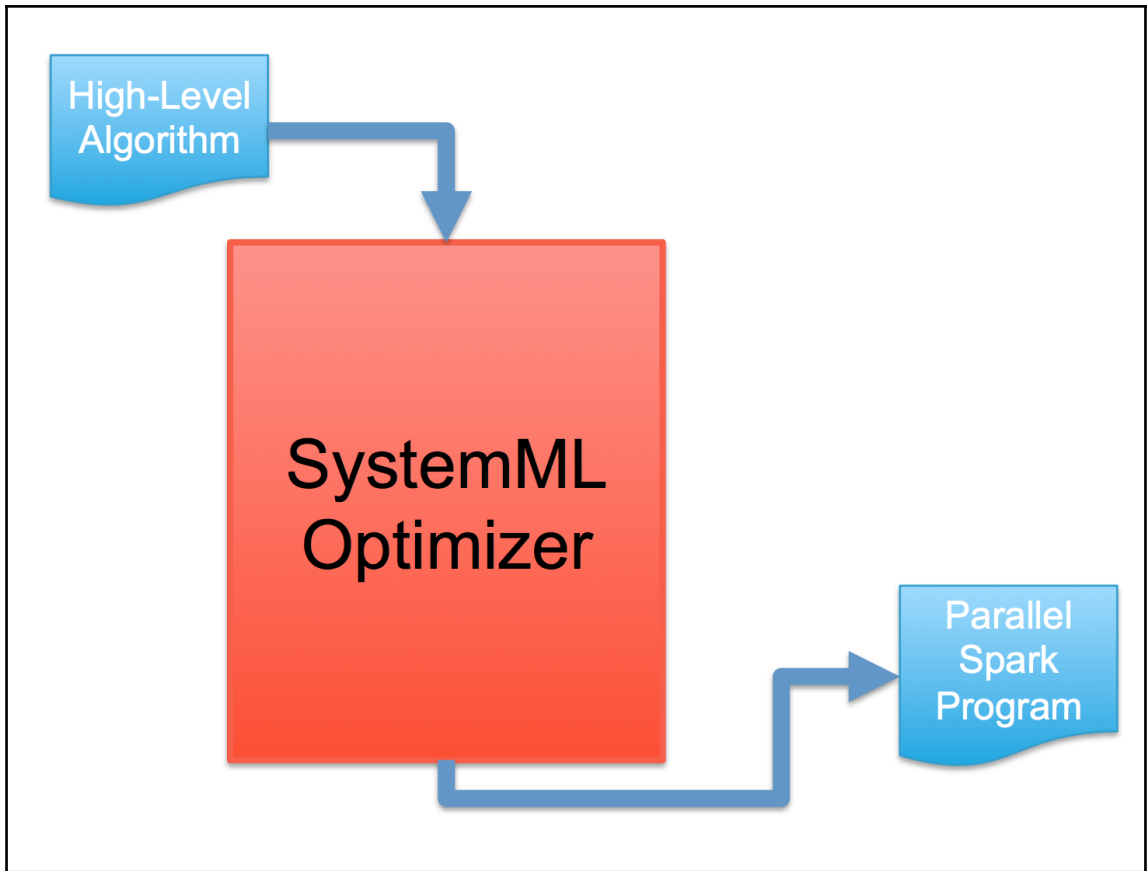


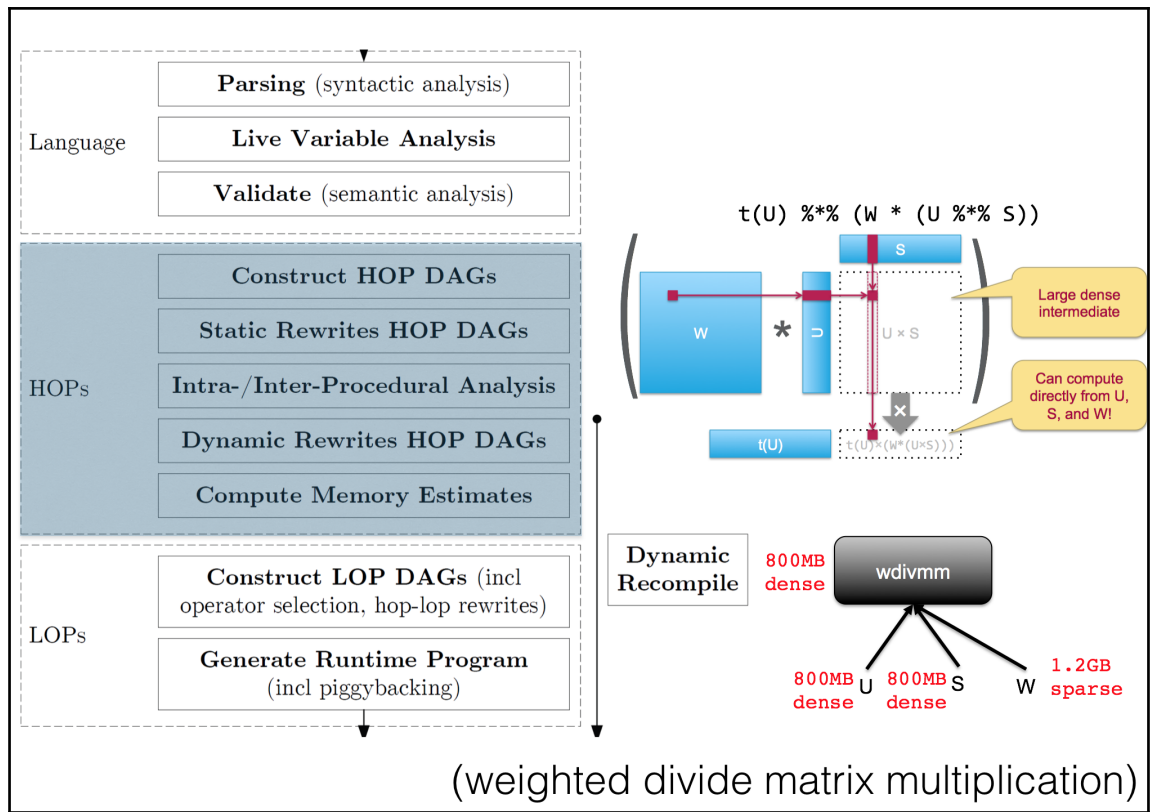


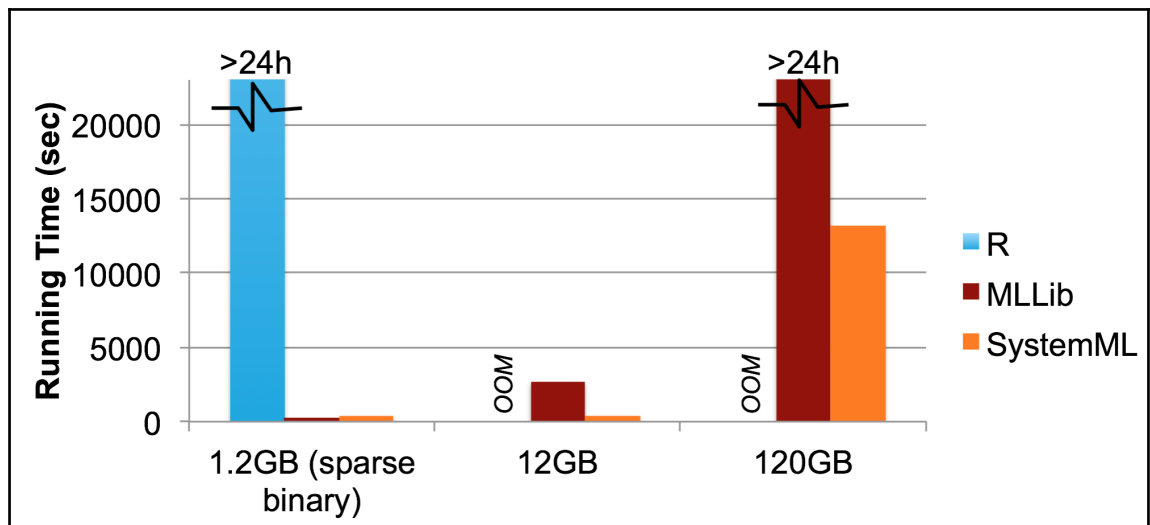
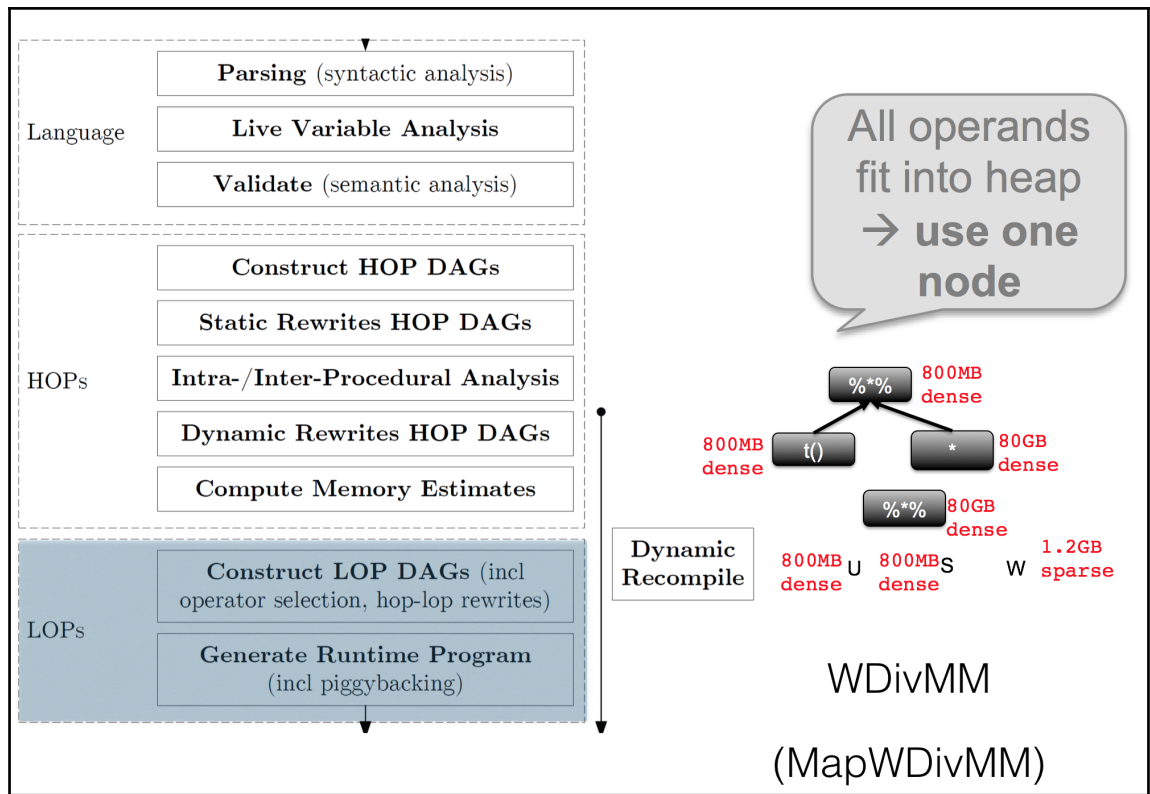


```
val model = ALS.train(ratings, rank, numIterations, 0.01)
```

```
U = rand(nrow(X), r, min = -1.0, max = 1.0);
V = rand(r, ncol(X), min = -1.0, max = 1.0);
while(i < mi) {
  i = i + 1; ii = 1;
  if (is_U)
    G = (W * (U %*% V - X)) %*% t(V) + lambda * U;
  else
    G = t(U) %*% (W * (U %*% V - X)) + lambda * V;
  norm_G2 = sum(G ^ 2); norm_R2 = norm_G2;
  R = -G; S = R;
  while(norm_R2 > 10E-9 * norm_G2 & ii <= mii) {
    if (is_U) {
      HS = (W * (S %*% V)) %*% t(V) + lambda * S;
      alpha = norm_R2 / sum (S * HS);
      U = U + alpha * S;
    } else {
      HS = t(U) %*% (W * (U %*% S)) + lambda * S;
      alpha = norm_R2 / sum (S * HS);
      V = V + alpha * S;
    }
    R = R - alpha * HS;
    old_norm_R2 = norm_R2; norm_R2 = sum(R ^ 2);
    S = R + (norm_R2 / old_norm_R2) * S;
    ii = ii + 1;
  }
  is_U = ! is_U;
}
```

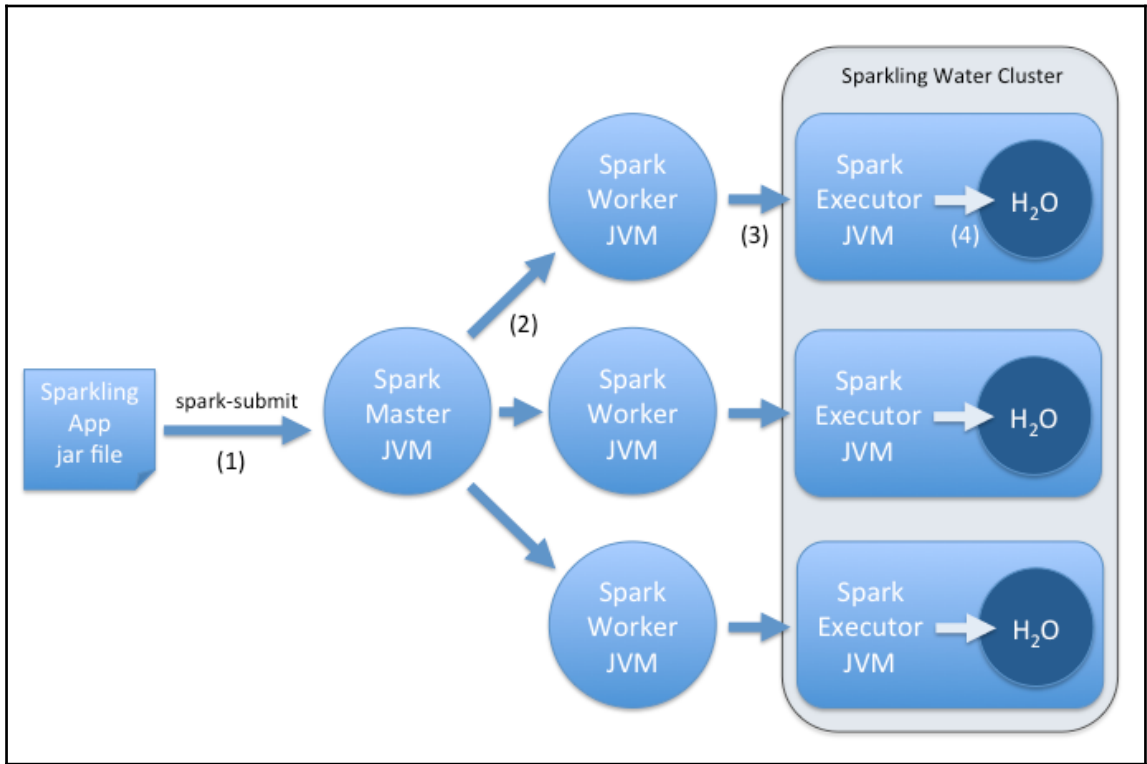


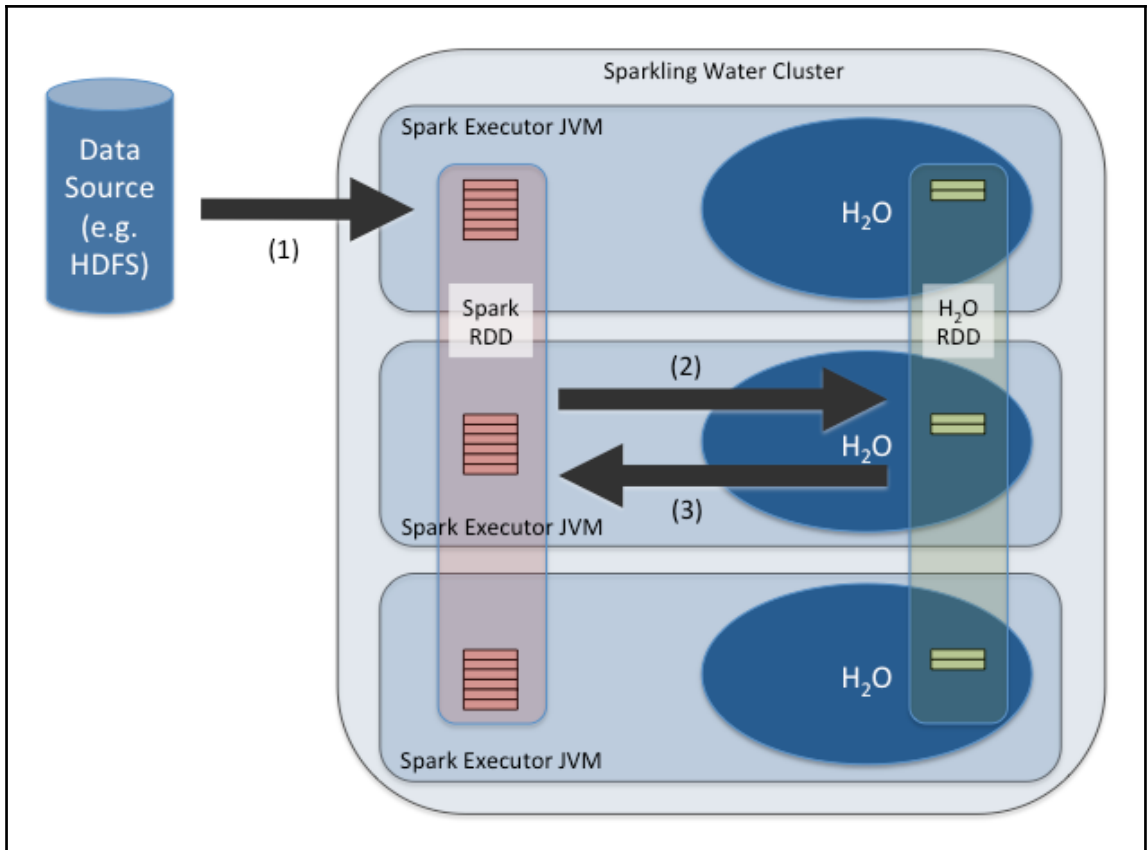


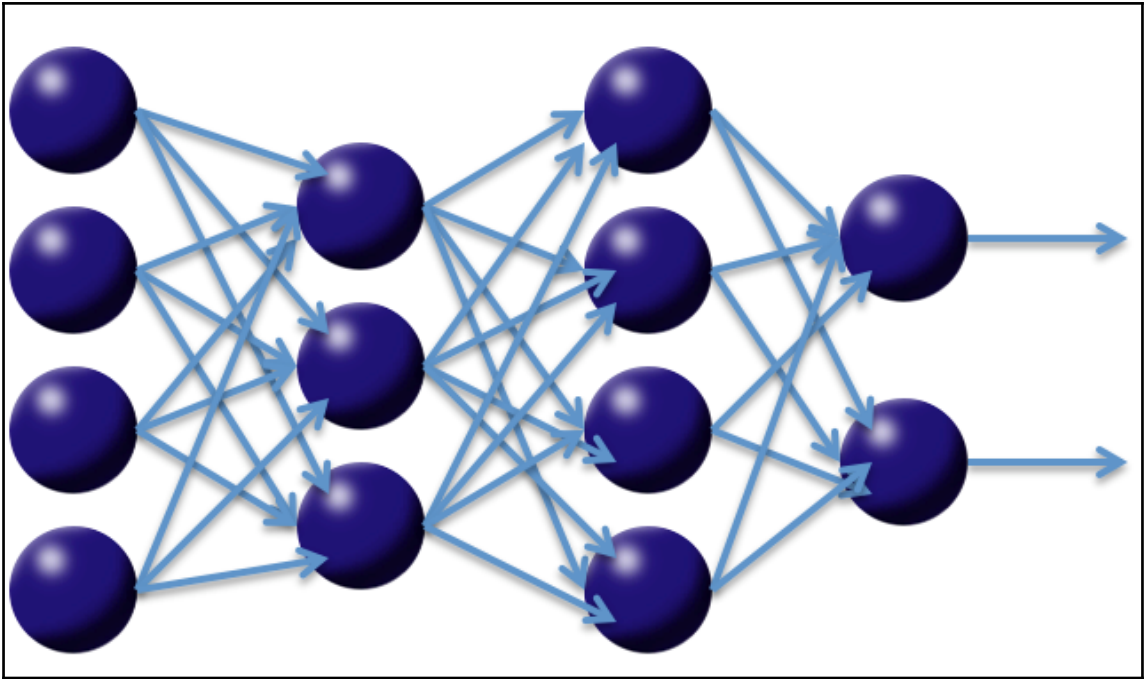


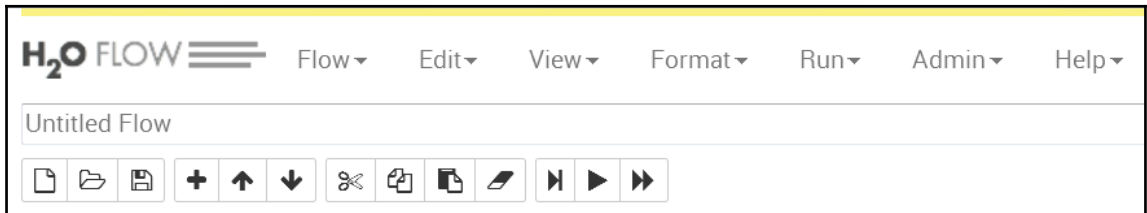
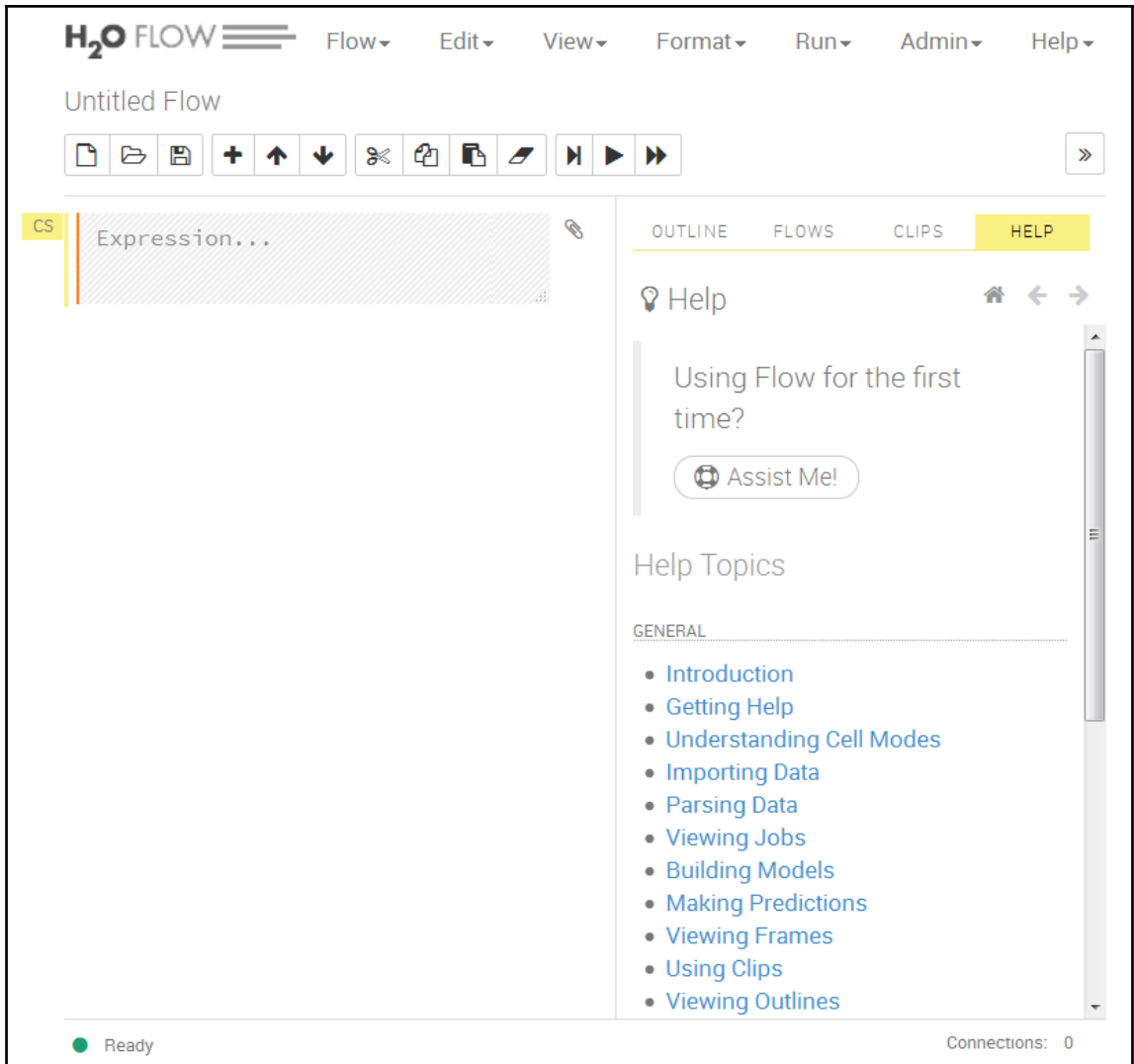


# Deep Learning on Apache Spark with DeepLearning4j and H2O









OUTLINE FLOWS CLIPS **HELP**

Help 🏠 ⬅️ ➡️

Using Flow for the first time?

Assist Me!

Help Topics

GENERAL

- [Introduction](#)
- [Getting Help](#)

GENERAL

- [Understanding Cell Modes](#)
- [Importing Data](#)
- [Parsing Data](#)
- [Viewing Jobs](#)
- [Building Models](#)
- [Making Predictions](#)
- [Viewing Frames](#)
- [Using Clips](#)
- [Viewing Outlines](#)
- [Saving Flows](#)
- [Troubleshooting](#)

PACKS

Flow packs are a great way to explore and learn H<sub>2</sub>O. Try out these Flows and run them in your browser.

[Browse installed packs...](#)

H<sub>2</sub>O REST API

- [Routes](#)
- [Schemas](#)

## sparkling-water-hadoop

CLOUD STATUS

HEALTHY
 CONSENSUS
 LOCKED

VERSION 0.3.0.1109    STARTED 3 minutes ago    NODES (USED / ALL) 4 / 4

⌵ Show advanced

NAME	PING	CORES	LOAD	DATA	DATA (%)	GC (FREE / TOTAL / MAX)	DISK (FREE / MAX)	DISK
192.168.1.105:54321	few seconds	2	0.120	- / -	NaN%	25.49 MB / 83.00 MB / 83.00 MB	39.02 GB / 49.21 GB	79%
192.168.1.108:54321	few seconds	2	0.200	- / -	NaN%	31.76 MB / 79.00 MB / 79.00 MB	39.08 GB / 49.21 GB	79%
192.168.1.109:54321	few seconds	2	0.070	- / -	NaN%	28.03 MB / 79.00 MB / 79.00 MB	39.41 GB / 49.21 GB	80%
192.168.1.110:54321	few seconds	2	0.080	- / -	NaN%	29.57 MB / 79.50 MB / 79.50 MB	39.30 GB / 49.21 GB	79%
TOTAL	-	8	0.470	- / -	NaN%	114.84 MB / 320.50 MB / 320.50 MB	156.81 GB / 196.86 GB	79%

## Job

TYPE **Model**

KEY **Q DeepLearningModel\_\_80e9dc14ccaa364696abec6f18194c5a**

DESCRIPTION **DeepLearning**

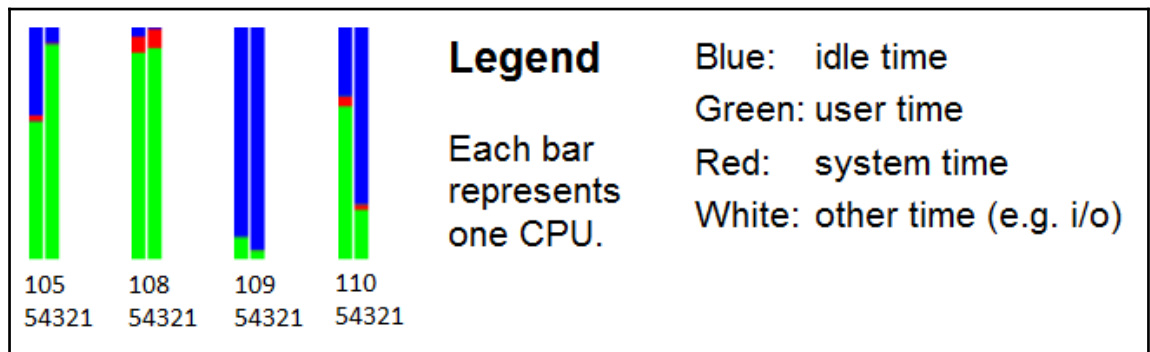
STATUS **RUNNING**

RUN TIME **00:00:32.949**


PROGRESS **11%**

Training at 8588 samples/s... Estimated time left: 9 min 8.882 sec

ACTIONS View Cancel Job



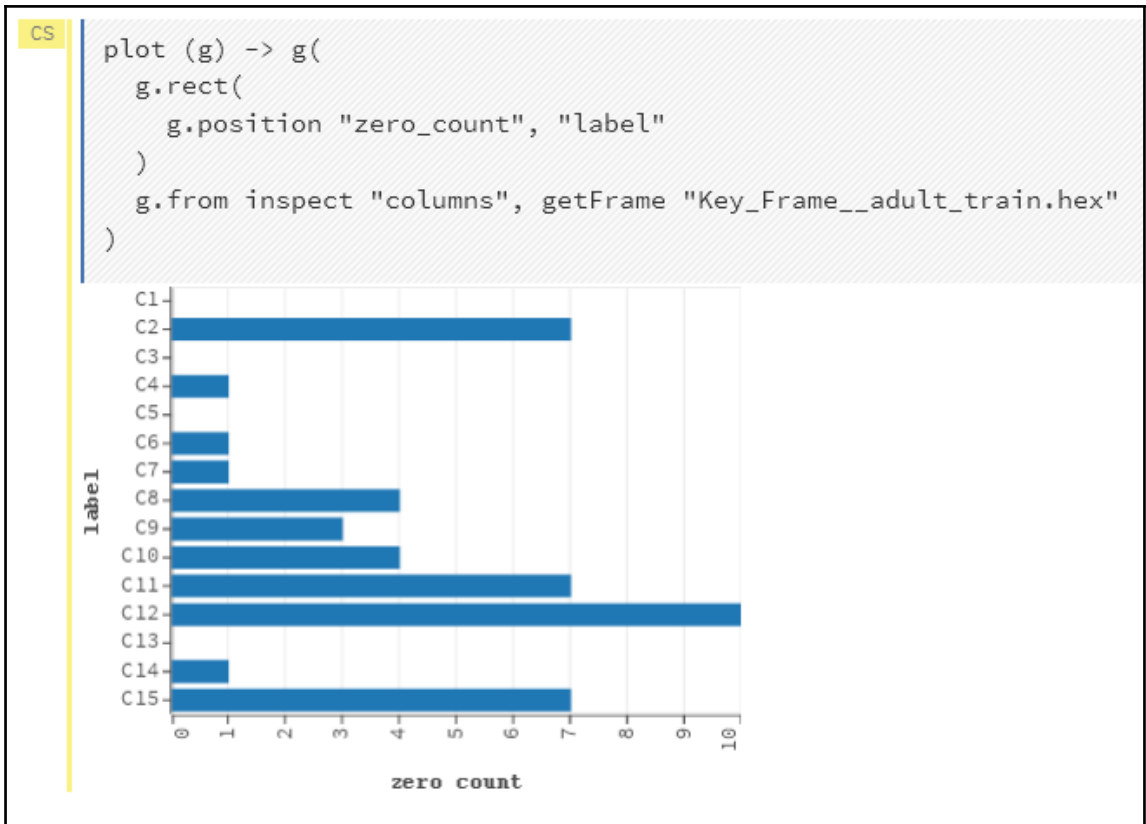
DATA PREVIEW						
Numeric	Enum	Numeric	Enum	Numeric	Enum	Enum
39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners

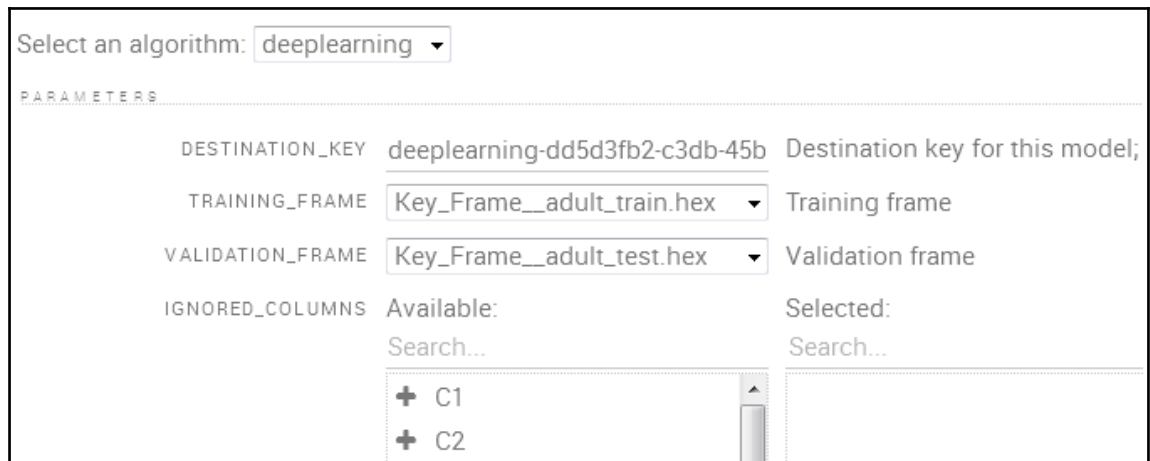
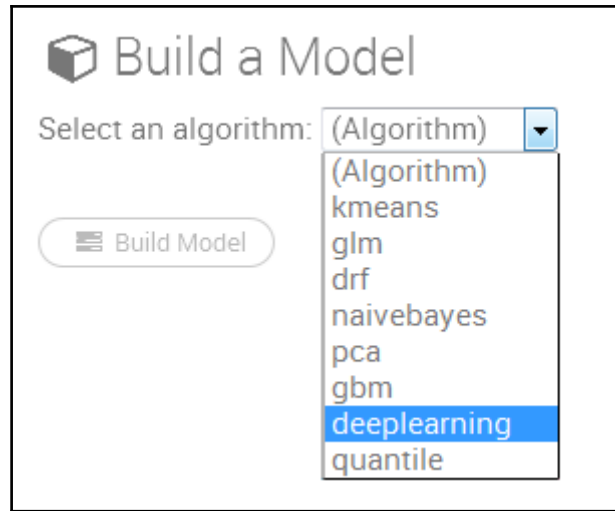
 Key\_Frame\_\_adult\_train.hex

ACTIONS: [View Data](#) [Inspect](#) [Build Model...](#) [Predict](#) [Download](#)

	ROWS	COLUMNS	COMPRESSED SIZE
	10	15	1KB

LABEL	MISSING_COUNT	ZERO_COUNT	POSITIVE_INFINITY_COUNT	NEGATIVE_INFINITY_COUNT	MIN	MAX	MEAN
C1	0	0	0	0	28	53	41.9
C2	0	7	0	0	0	2	0.4
C3	0	0	0	0	45781	338409	180924.4
C4	0	1	0	0	0	4	2.3
C5	0	0	0	0	5	14	11
C6	0	1	0	0	0	3	1.4







DROPNA20COLS	<input type="checkbox"/>	CHECKPOINT	
RESPONSE_COLUMN	C15	USE_ALL_FACTOR_LEVELS	<input checked="" type="checkbox"/>
N_FOLDS	0	TRAIN_SAMPLES_PER_ITERATION	-2
ACTIVATION	Rectifier	ADAPTIVE_RATE	<input checked="" type="checkbox"/>
HIDDEN	200, 200	RHO	0.99
EPOCHS	100	EPSILON	1e-8
VARIABLE_IMPORTANCES	<input checked="" type="checkbox"/>	INPUT_DROPOUT_RATIO	0
REPLICATE_TRAINING_DATA	<input checked="" type="checkbox"/>	L1	0
		L2	0

CS

```
buildModel 'deeplearning', {"destination_key":"deeplearning-ded16717-75cb-4014-a1fb-b99d257a3056","training_frame":"Key_Frame__adult_train2.hex","variable_importances":true,"activation":"Rectifier","hidden":[200,200],"epochs":"100","variable_importances":true,"replicate_training_data":true}
```

### Job

TYPE Model

KEY deeplearning-ded16717-75cb-4014-a1fb-b99d257a3056

DESCRIPTION DeepLearning

STATUS RUNNING

RUN TIME 00:00:57.911

PROGRESS 13%

Scoring on 10018 training samples, 16281 validation samples)

ACTIONS 🔍 View 🚫 Cancel Job

**Model** KEY: deeplearning-ded16717-75cb-4014-a1fb-b99d257a3056  
ALGORITHM: deeplearning

ACTIONS: [⚡ Predict...](#) [📄 Clone this model...](#) [☰ Inspect](#) [🗑 Delete](#)

▶ MODEL PARAMETERS

▼ TRAINING METRICS

MODEL_CATEGORY	AUC	GINI	MSE	DURATION_IN_MS	SCORING_TIME
Binomial	0.917392	0.834784	0.097503	0	0

▼ VALIDATION METRICS

MODEL_CATEGORY	AUC	GINI	MSE	DURATION_IN_MS	SCORING_TIME
Binomial	0.908921	0.817843	0.101004	0	0

▼ VARIABLE IMPORTANCES

Variable	Importance
C9.White	0.35
C14.United-States	0.32
C6.Married-civ-spouse	0.28
C11	0.25

**⚡ Predict**

KEY:

MODEL: deeplearning-ded16717-75cb-4014-a1fb-b99d257a3056

FRAME:  ▼

ACTIONS: [⚡ Predict](#)

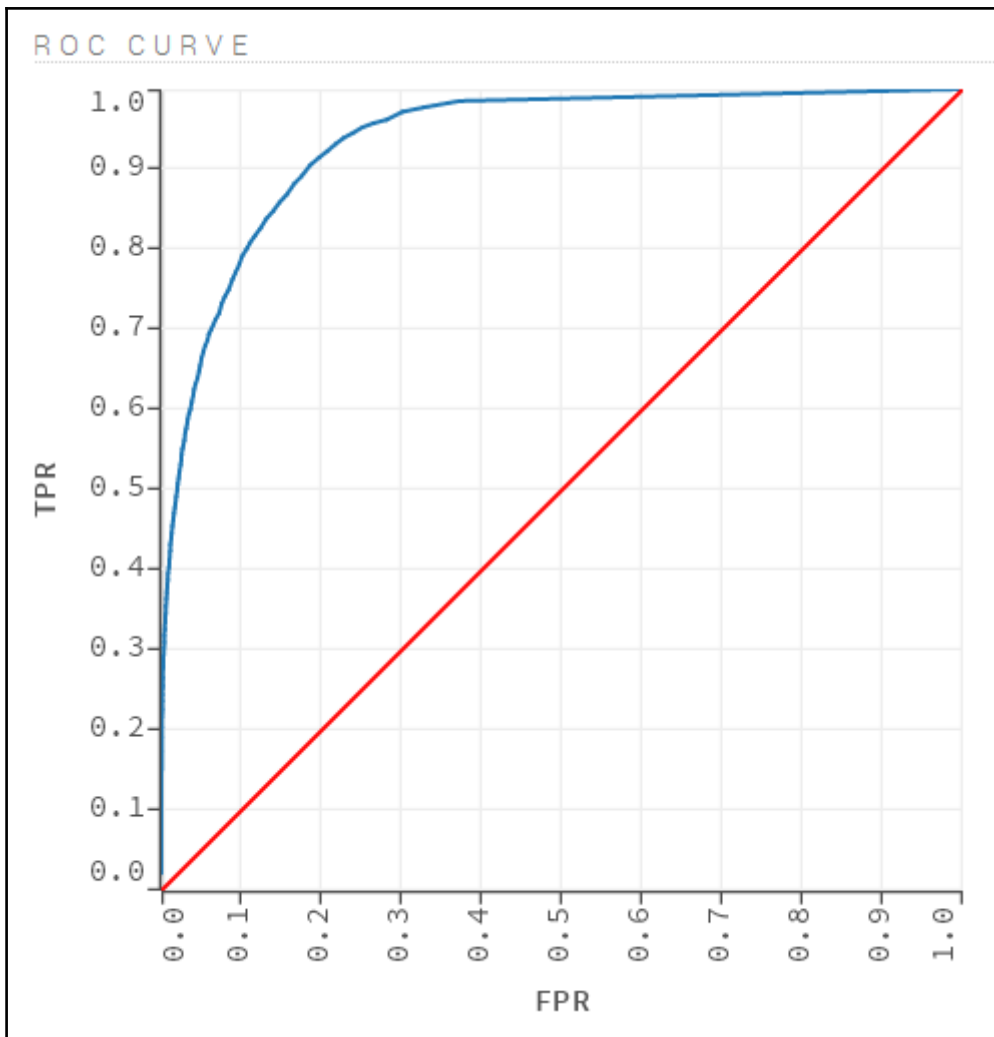
## Prediction

ACTIONS: [Inspect](#) [View Prediction Frame](#)

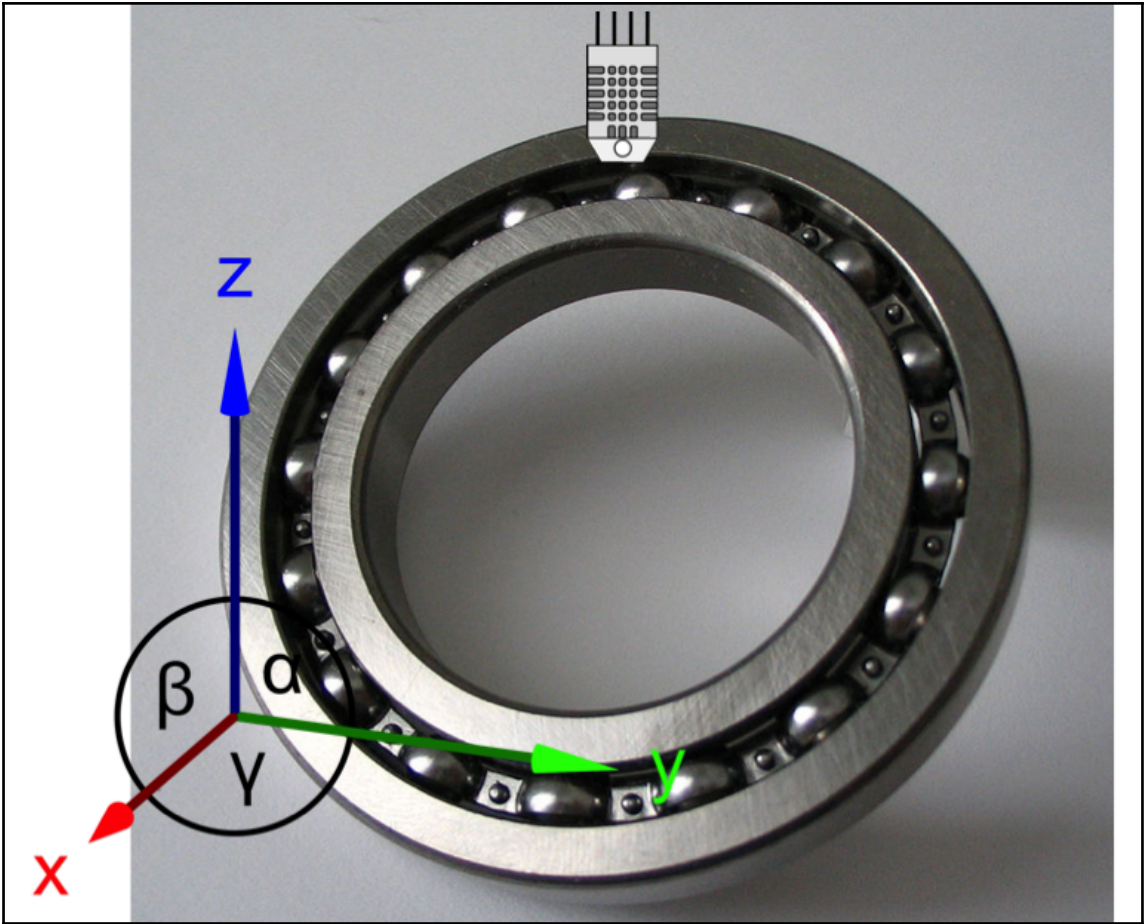
OUTPUT

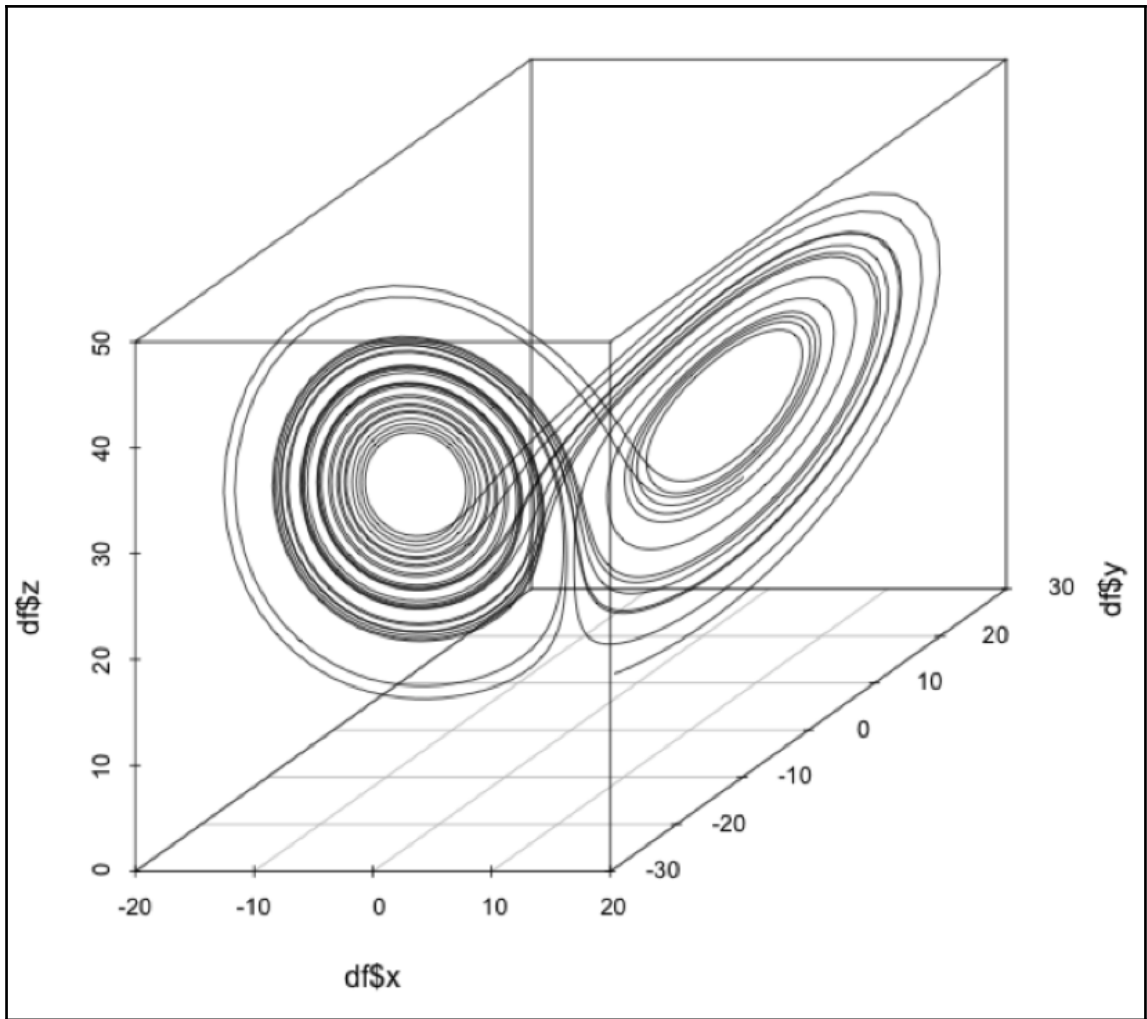
---

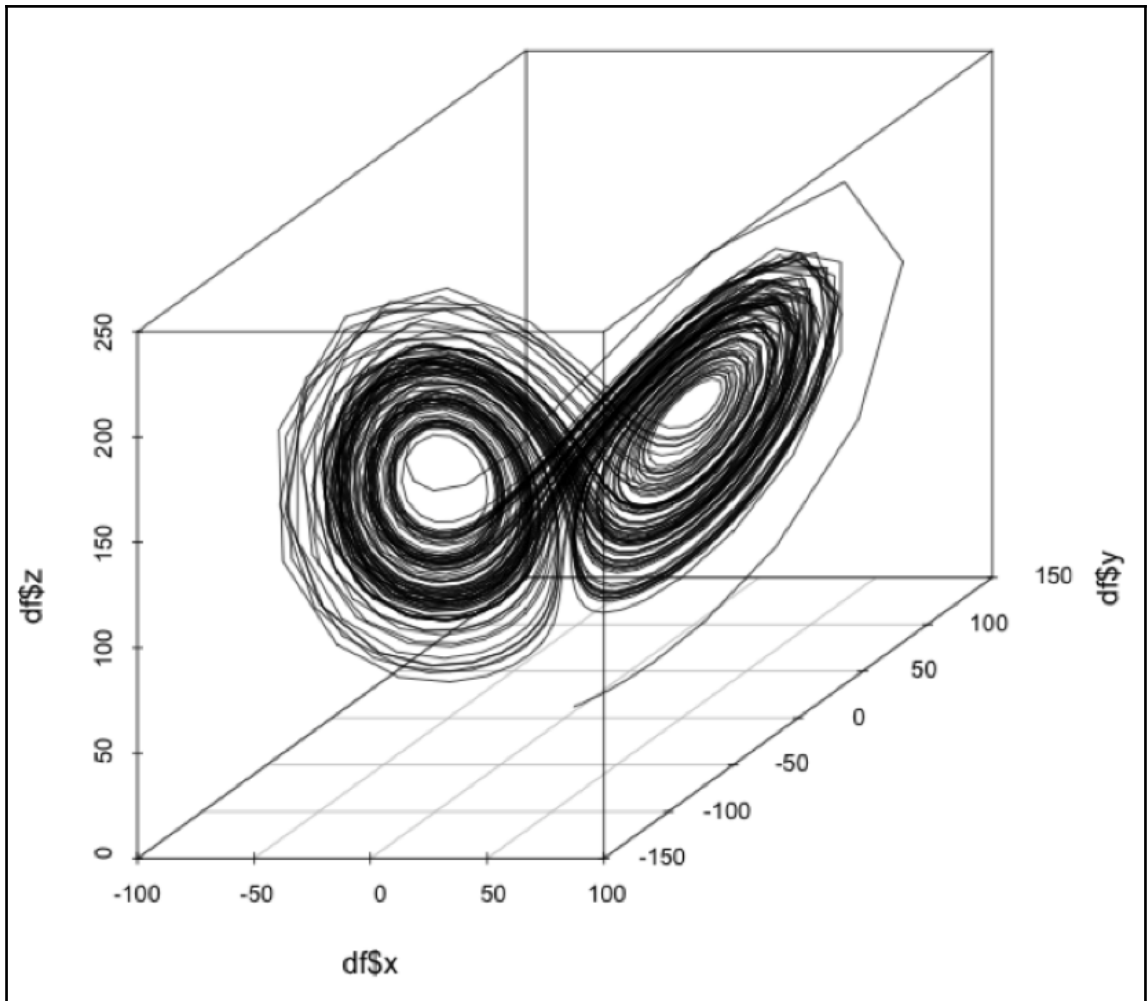
KEY	deeplearning-412a09ba-0c9a-4e80-8f70
FRAME	Key_Frame__adult_test.hex
MODEL_CATEGORY	Binomial
AUC	0.9366408391759876
GINI	0.8732816783519752
MSE	0.08692155822396484
DURATION_IN_MS	0
SCORING_TIME	0



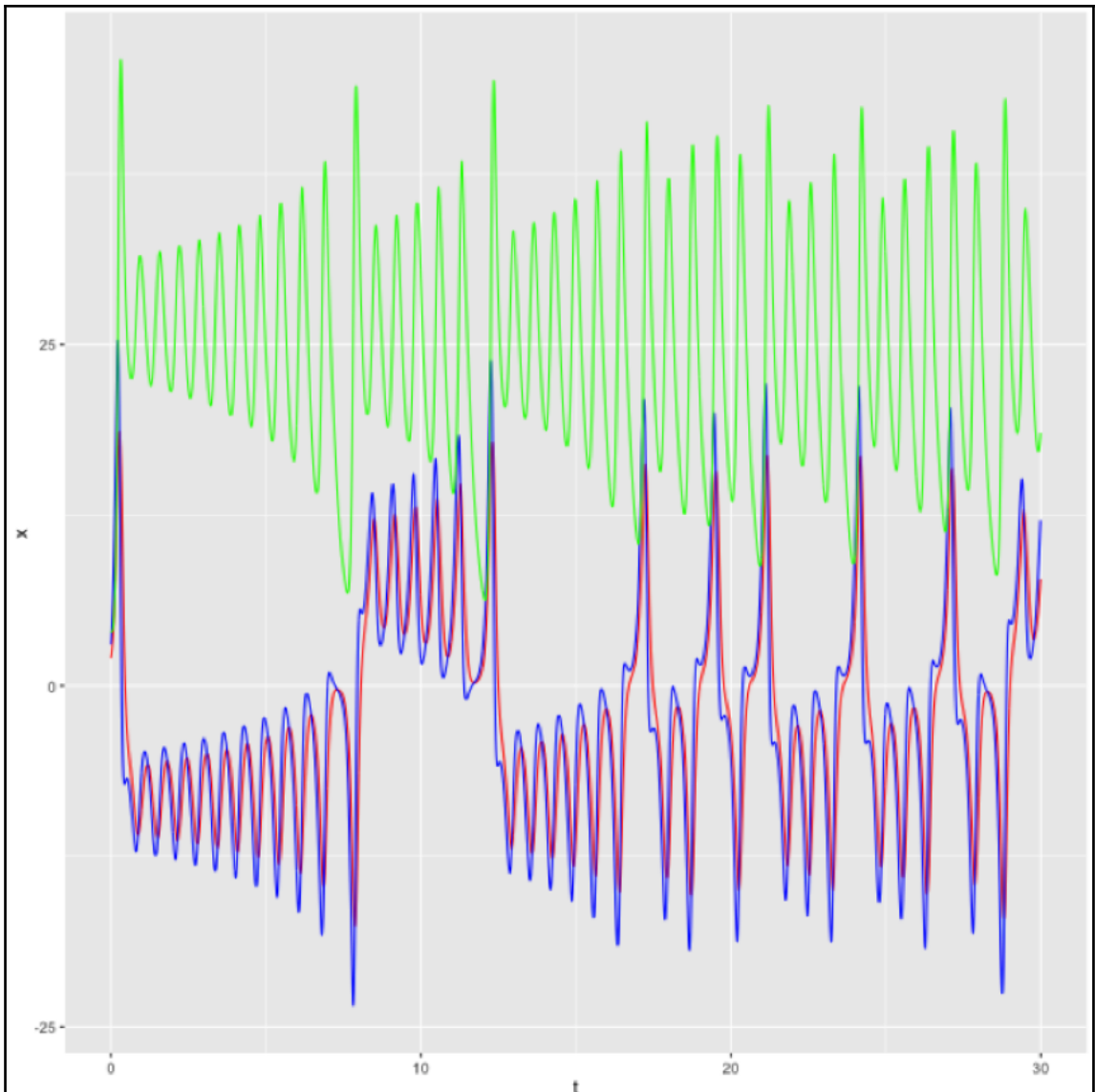
Input 1	Input 2	Output
0	0	0
0	1	1
1	0	1
1	1	0

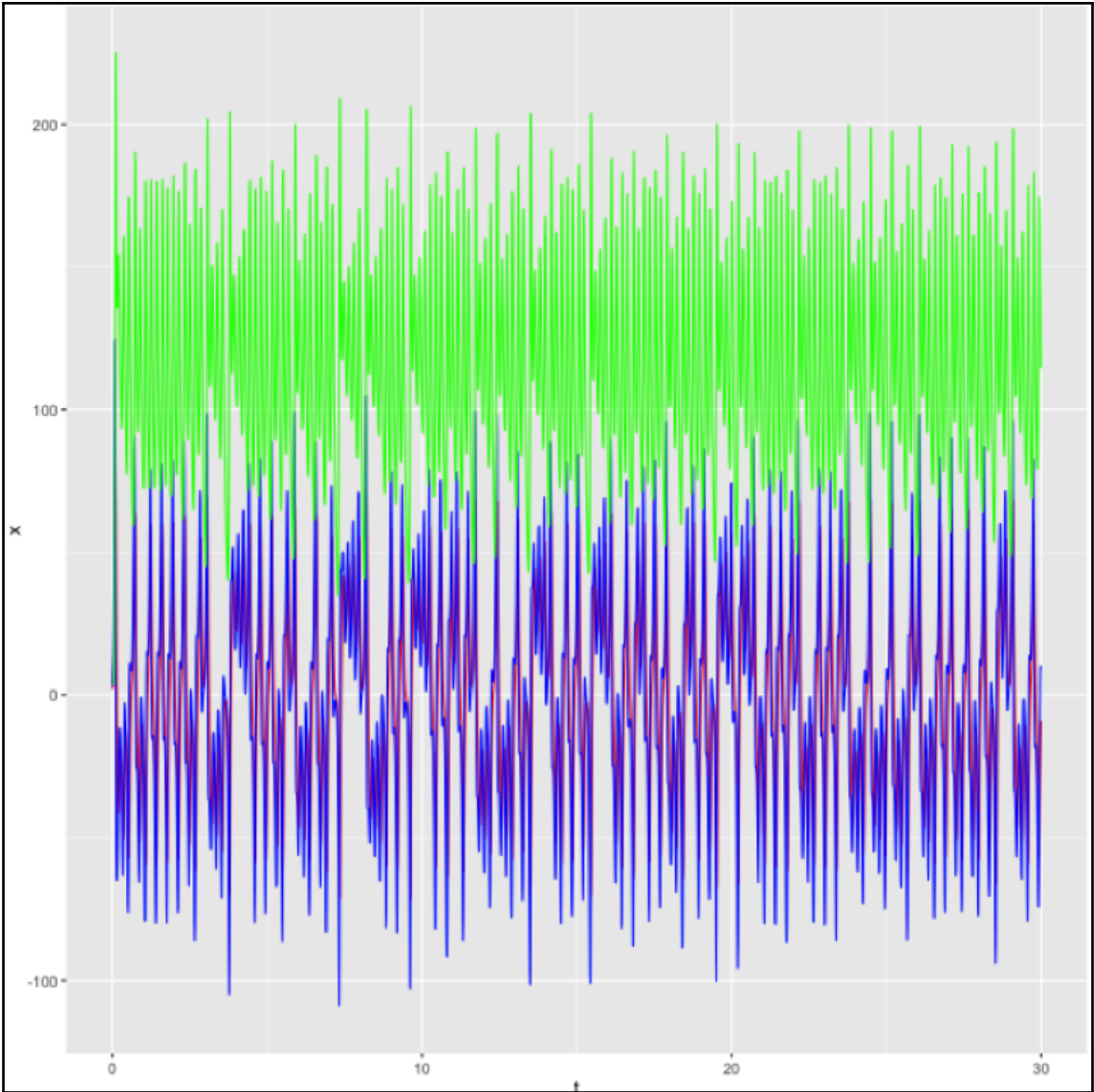


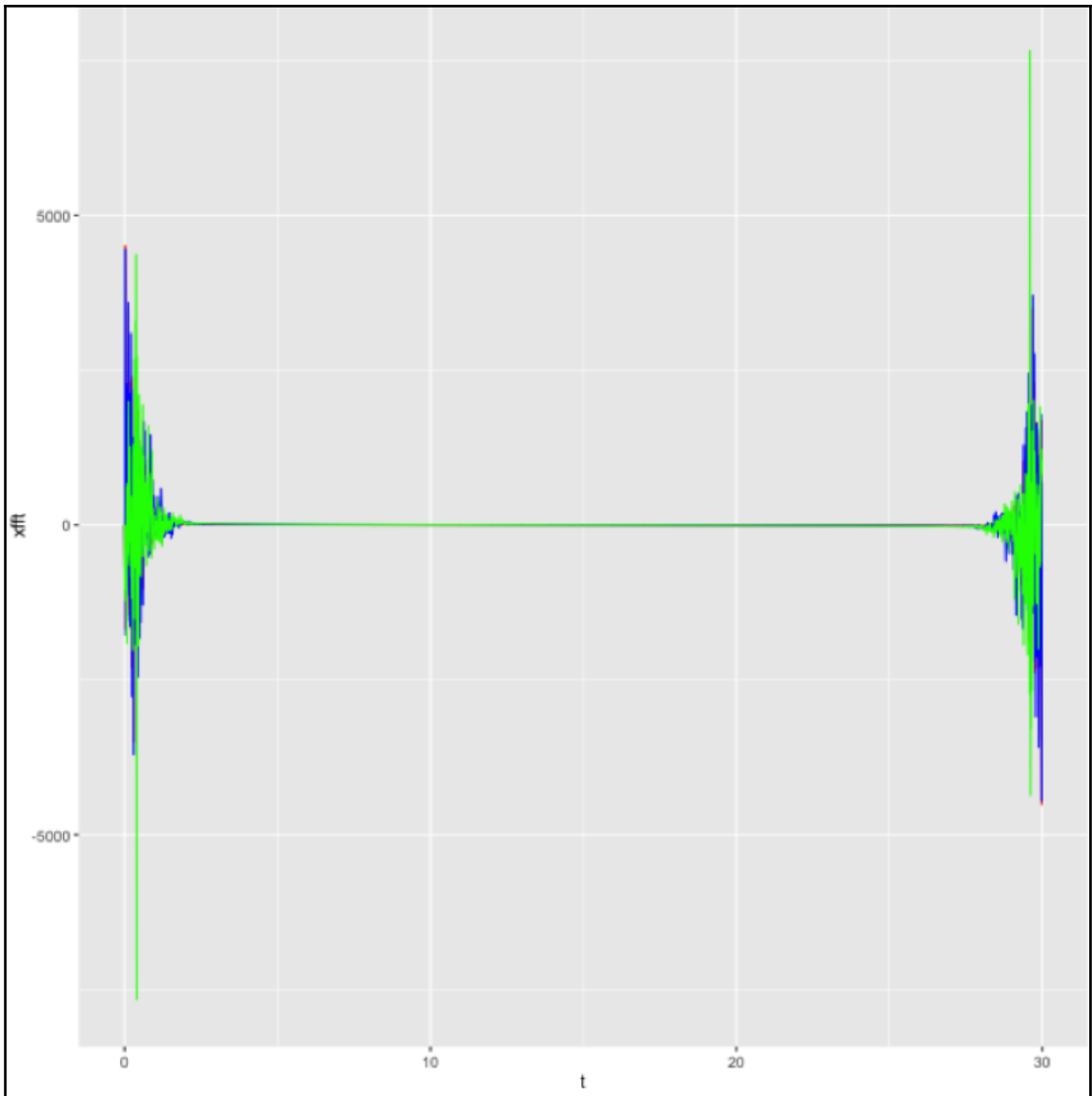


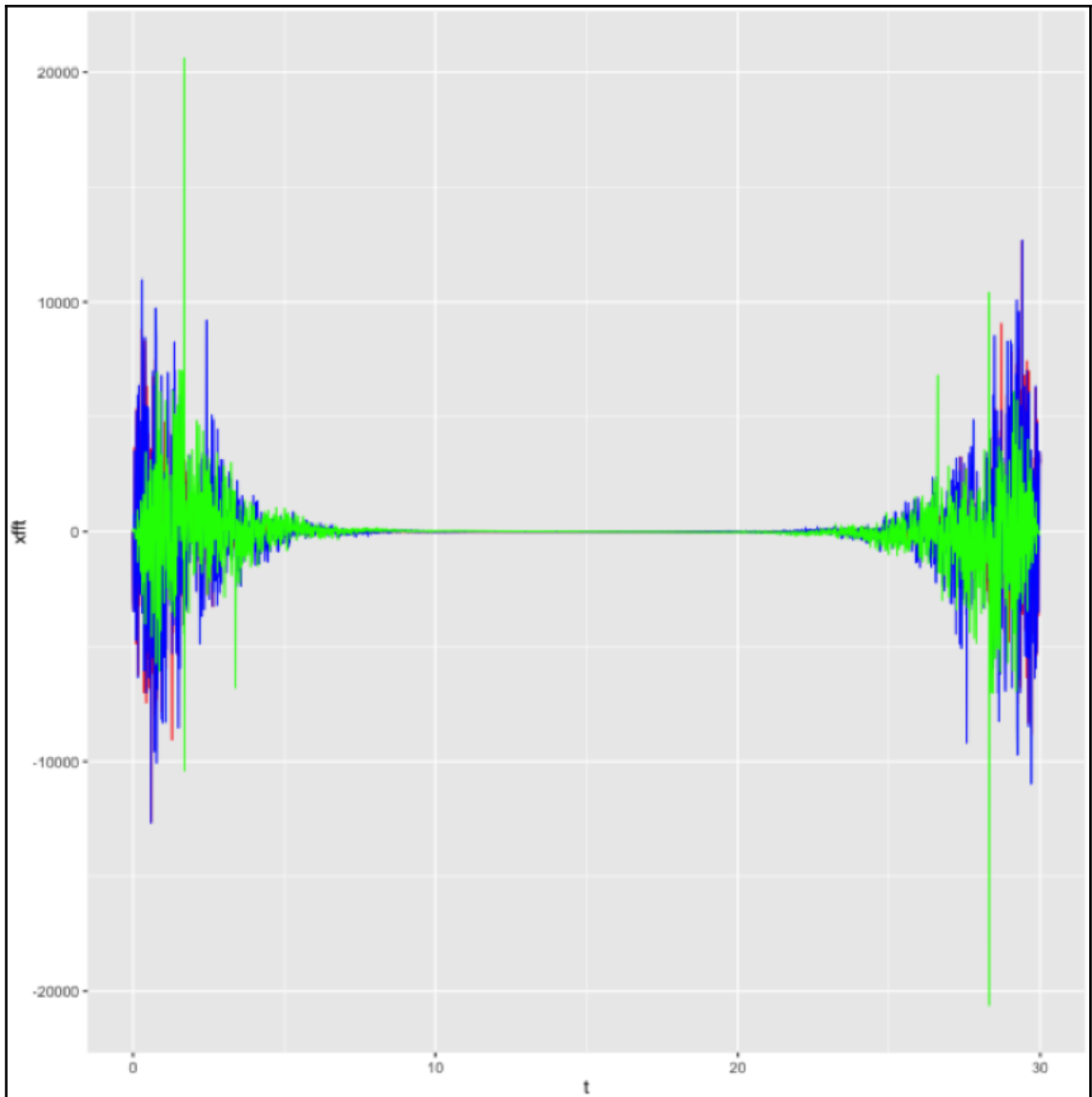


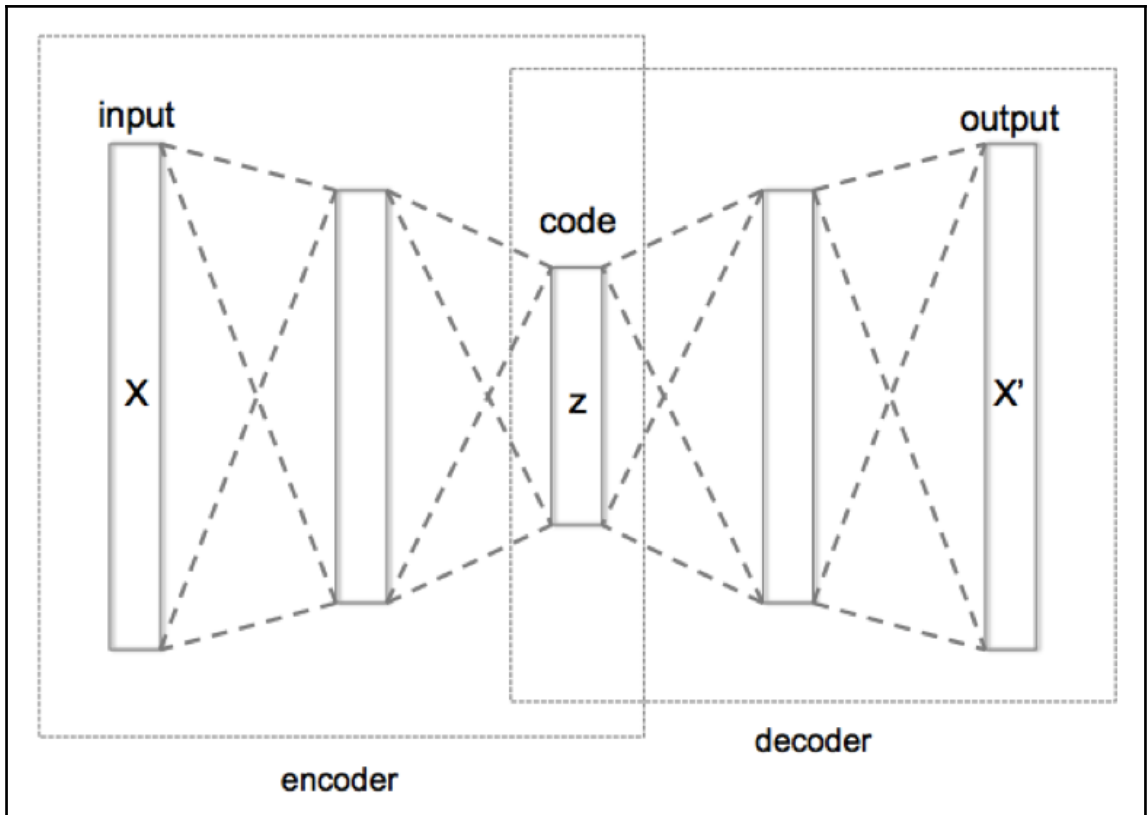


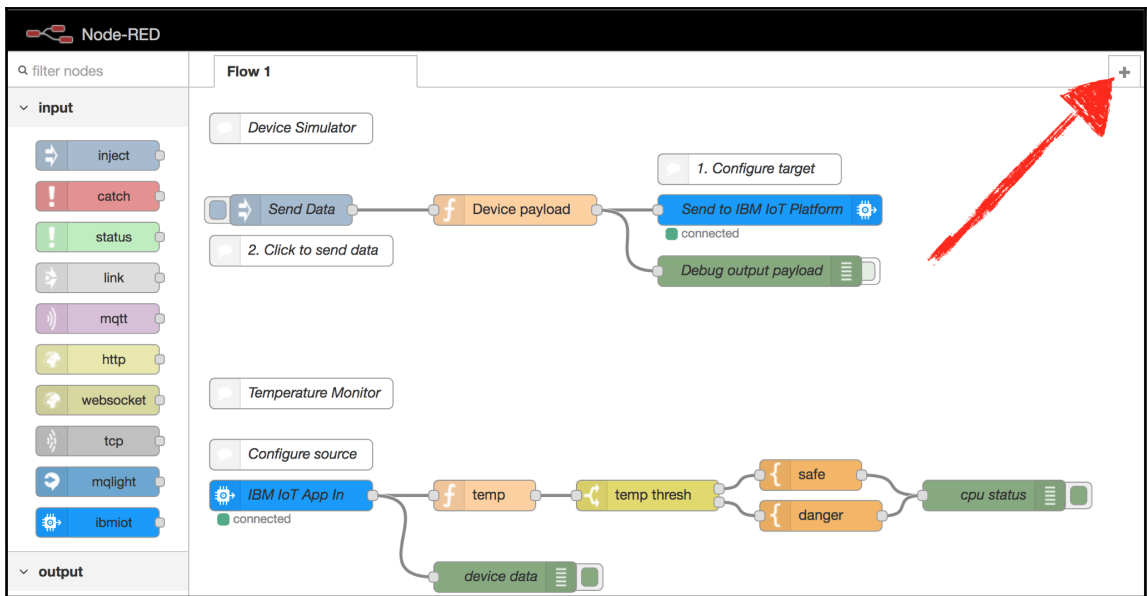
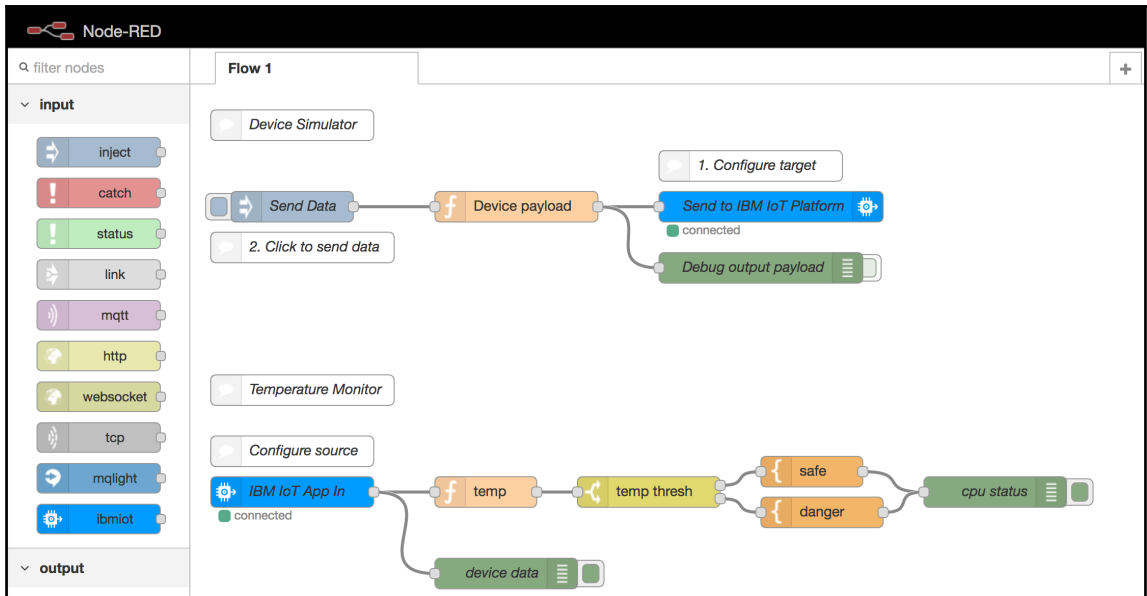


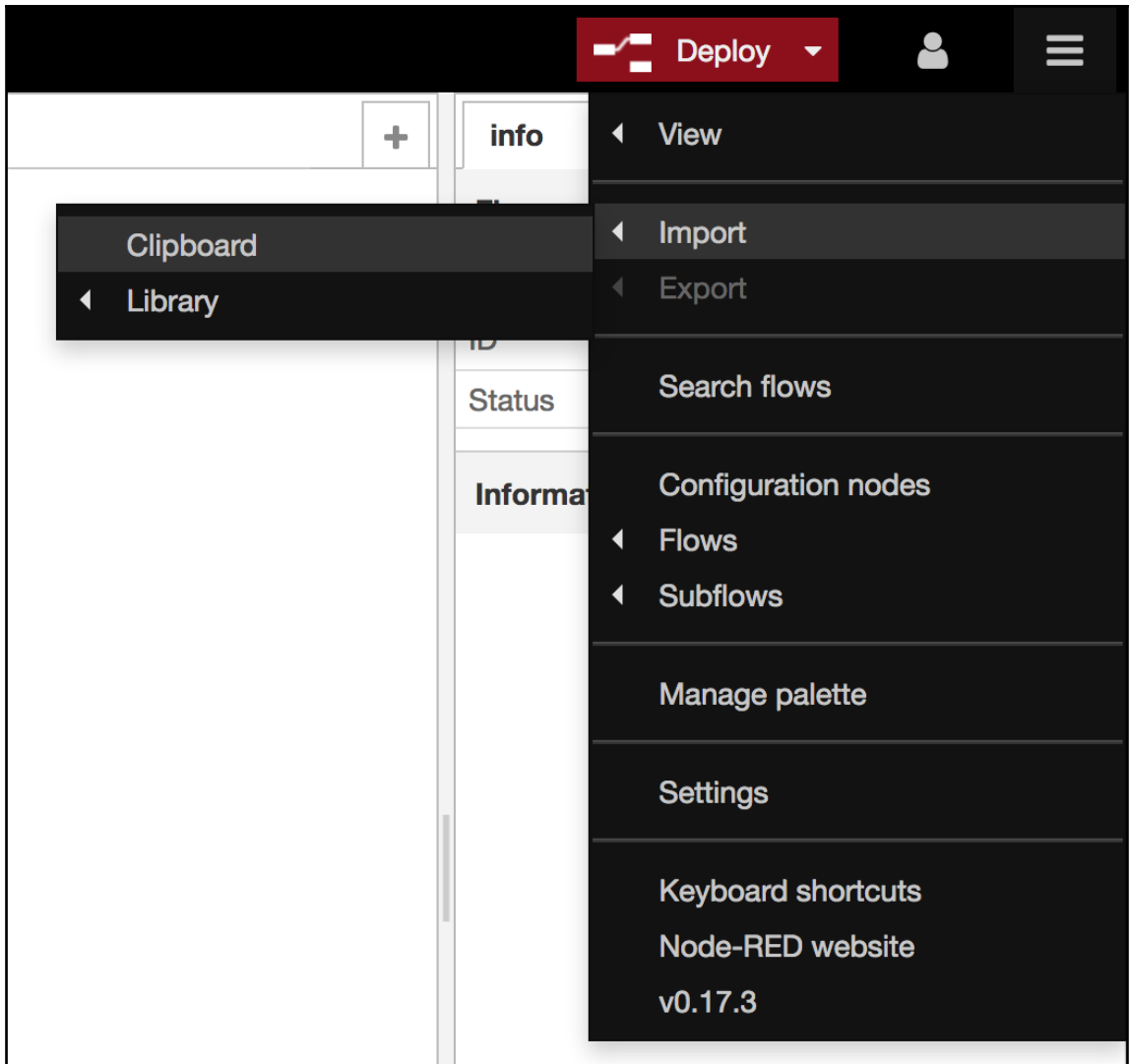












## Import nodes

```
routab":"","once":false,"x":128,"y":468.25,"wires":  
[[{"9f292ce3.075de"}]],  
{ "id":"d199e025.05a56", "type":"debug", "z":"92f63ac.5b29ac8", "name":  
"e":"","active":true, "console":"false", "complete":"false", "x":378.5, "y":  
574.25, "wires":[]}]
```

Import to

current flow

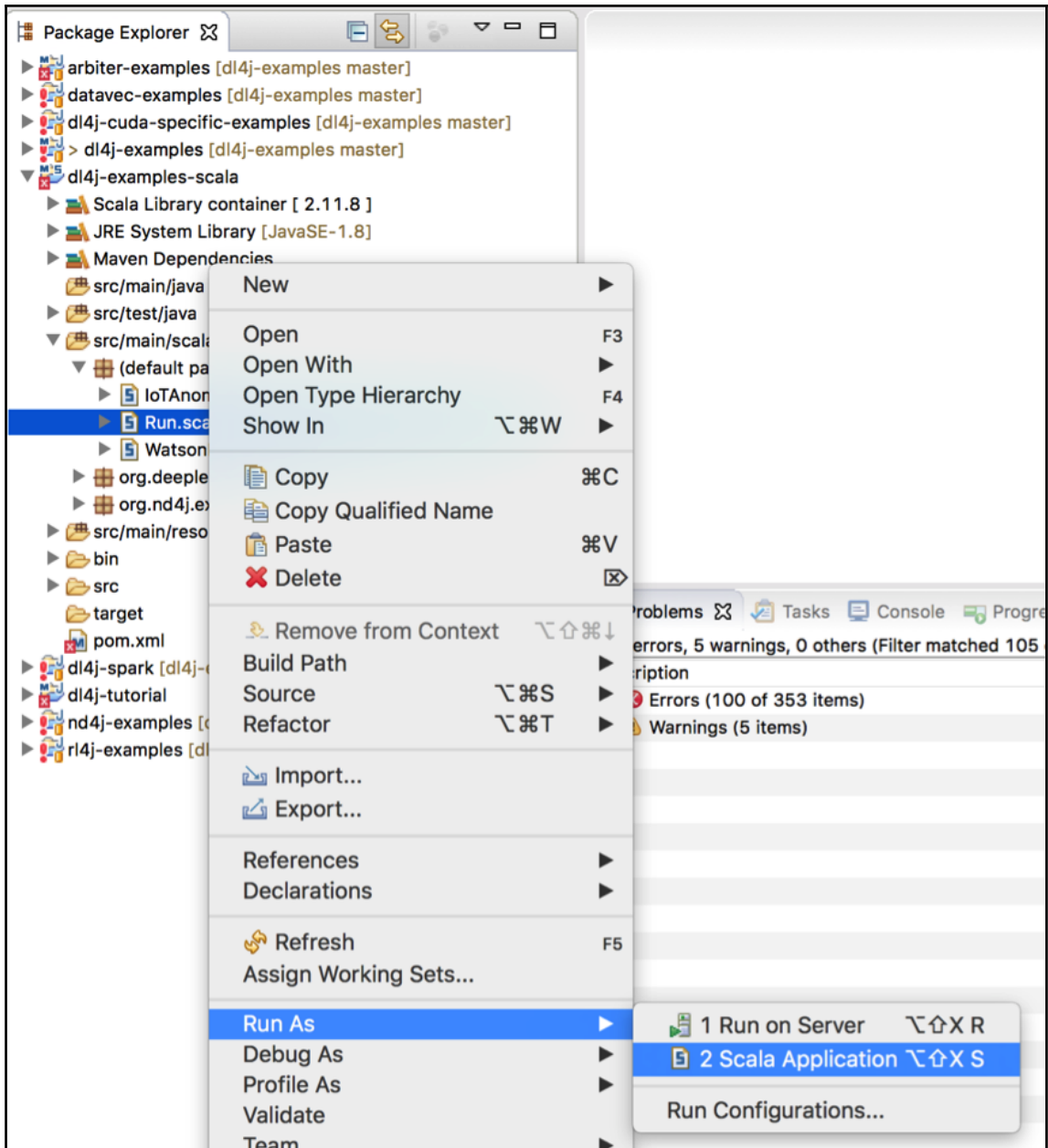
new flow

Cancel

Import

The screenshot shows the Node-RED web interface. On the left, the 'input' nodes list includes 'inject', 'catch', 'status', 'link', 'mqtt', 'http', 'websocket', 'tcp', 'mqlight', and 'ibmiot'. The 'output' nodes list includes 'debug', 'link', 'mqtt', and 'http response'. A red circle highlights the 'reset' node in the input list. The main workspace shows a flow with several nodes: 'timestamp v', a function node 'f', 'msg.payload', another function node 'f', 'limit to max 3000', 'IBM IoT' (connected), 'msg.payload', 'healthy', 'broken', another function node 'f', 'msg.payload', and another function node 'f'. The 'debug' node in the output list is also highlighted with a red circle. On the right, the 'debug' console shows several error messages: 'TypeError: Cannot read property 'temp' of undefined'. A red arrow points to the top of the console area.





```
Markers Properties Servers Data Source Explorer Snippets Console Progress
Run$ [Scala Application] /Library/Java/JavaVirtualMachines/jdk1.8.0_65.jdk/Contents/Home/bin/java (23 Jun 2017, 07:34:54)
jar:file:/Users/romeokienzler/.m2/repository/org/nd4j/nd4j-common/0.8.0/nd4j-common-0.8.0.jar!/
jar:file:/Users/romeokienzler/.m2/repository/org/nd4j/nd4j-native/0.8.0/nd4j-native-0.8.0-windows-x86_64.jar!/
07:34:59.589 [main] INFO org.reflections.Reflections - Reflections took 91 ms to scan 14 urls, producing 373 keys and 1449 values
Jun 23, 2017 7:34:59 AM com.ibm.iotf.client.AbstractClient createClient
INFO: main: Org ID = rwyrty
Client ID = a:rwyrty:a2g6k39sl6r5
Jun 23, 2017 7:34:59 AM com.ibm.iotf.client.AbstractClient connect
INFO: main: Connecting client a:rwyrty:a2g6k39sl6r5 to ssl://rwyrty.messaging.internetofthings.ibmcloud.com:8883 (attempt #1)...
Jun 23, 2017 7:35:09 AM com.ibm.iotf.client.AbstractClient connect
INFO: main: Successfully connected to the IBM Watson IoT Platform
Mainthread blocking...
```

```
Markers Properties Servers Data Source Explorer Snippets Console Progress
Run$ [Scala Application] /Library/Java/JavaVirtualMachines/jdk1.8.0_65.jdk/Contents/Home/bin/java (21 Jun 2017, 07:55:32)
Mainthread blocking...
Mainthread blocking...
Mainthread blocking...
Mainthread blocking...
Waiting for tumbling window to fill: 0
Waiting for tumbling window to fill: 100
Waiting for tumbling window to fill: 200
Waiting for tumbling window to fill: 300
Waiting for tumbling window to fill: 400
Waiting for tumbling window to fill: 500
Waiting for tumbling window to fill: 600
```

```
o.d.o.l.ScoreIterationListener - Score at iteration 0 is 392314.67211754626
o.d.o.l.ScoreIterationListener - Score at iteration 1 is 392312.42505880696
o.d.o.l.ScoreIterationListener - Score at iteration 2 is 392304.37555201456
o.d.o.l.ScoreIterationListener - Score at iteration 3 is 392282.8154061016
o.d.o.l.ScoreIterationListener - Score at iteration 4 is 392234.12018756795
o.d.o.l.ScoreIterationListener - Score at iteration 5 is 392139.0617437172
o.d.o.l.ScoreIterationListener - Score at iteration 6 is 391977.5299519522
o.d.o.l.ScoreIterationListener - Score at iteration 7 is 391732.0313246322
o.d.o.l.ScoreIterationListener - Score at iteration 8 is 391390.9885874034
o.d.o.l.ScoreIterationListener - Score at iteration 9 is 390949.9068585132
o.d.o.l.ScoreIterationListener - Score at iteration 10 is 390407.2208170733
o.d.o.l.ScoreIterationListener - Score at iteration 11 is 389766.7980037157
o.d.o.l.ScoreIterationListener - Score at iteration 12 is 389026.07803709887
o.d.o.l.ScoreIterationListener - Score at iteration 13 is 388180.6244370007
o.d.o.l.ScoreIterationListener - Score at iteration 14 is 387225.1191501263
o.d.o.l.ScoreIterationListener - Score at iteration 15 is 386161.5252982354
o.d.o.l.ScoreIterationListener - Score at iteration 16 is 384991.52117921936
o.d.o.l.ScoreIterationListener - Score at iteration 17 is 383719.54410875245
o.d.o.l.ScoreIterationListener - Score at iteration 18 is 382351.4820195209
o.d.o.l.ScoreIterationListener - Score at iteration 19 is 380891.5161926356
o.d.o.l.ScoreIterationListener - Score at iteration 20 is 379344.3968079217
o.d.o.l.ScoreIterationListener - Score at iteration 21 is 377715.15848935687
o.d.o.l.ScoreIterationListener - Score at iteration 22 is 376010.77007275063
o.d.o.l.ScoreIterationListener - Score at iteration 23 is 374238.82619509666
o.d.o.l.ScoreIterationListener - Score at iteration 24 is 372405.6130678293
o.d.o.l.ScoreIterationListener - Score at iteration 25 is 370520.31821126794
o.d.o.l.ScoreIterationListener - Score at iteration 26 is 368582.4865625592
o.d.o.l.ScoreIterationListener - Score at iteration 27 is 366603.3535501402
o.d.o.l.ScoreIterationListener - Score at iteration 28 is 364589.54680638906
o.d.o.l.ScoreIterationListener - Score at iteration 29 is 362548.2627484569
o.d.o.l.ScoreIterationListener - Score at iteration 30 is 360485.1333794203
```

```
o.d.o.l.ScoreIterationListener - Score at iteration 969 is 425.6250059539804
o.d.o.l.ScoreIterationListener - Score at iteration 970 is 423.3333805602828
o.d.o.l.ScoreIterationListener - Score at iteration 971 is 421.4624281317155
o.d.o.l.ScoreIterationListener - Score at iteration 972 is 419.2211628537399
o.d.o.l.ScoreIterationListener - Score at iteration 973 is 417.4023765884213
o.d.o.l.ScoreIterationListener - Score at iteration 974 is 415.2149420317019
o.d.o.l.ScoreIterationListener - Score at iteration 975 is 413.440681888014
o.d.o.l.ScoreIterationListener - Score at iteration 976 is 411.2982748331003
o.d.o.l.ScoreIterationListener - Score at iteration 977 is 409.57493015976763
o.d.o.l.ScoreIterationListener - Score at iteration 978 is 407.4829112362655
o.d.o.l.ScoreIterationListener - Score at iteration 979 is 405.80495495344996
o.d.o.l.ScoreIterationListener - Score at iteration 980 is 403.7537126010534
o.d.o.l.ScoreIterationListener - Score at iteration 981 is 402.1170321945724
o.d.o.l.ScoreIterationListener - Score at iteration 982 is 400.1184546085061
o.d.o.l.ScoreIterationListener - Score at iteration 983 is 398.51955195517445
o.d.o.l.ScoreIterationListener - Score at iteration 984 is 396.5621405109357
o.d.o.l.ScoreIterationListener - Score at iteration 985 is 395.0127922092593
o.d.o.l.ScoreIterationListener - Score at iteration 986 is 393.09832353721544
o.d.o.l.ScoreIterationListener - Score at iteration 987 is 391.5845006616595
o.d.o.l.ScoreIterationListener - Score at iteration 988 is 389.7125865777661
o.d.o.l.ScoreIterationListener - Score at iteration 989 is 388.2405412156419
o.d.o.l.ScoreIterationListener - Score at iteration 990 is 386.410098613004
o.d.o.l.ScoreIterationListener - Score at iteration 991 is 384.97496987874547
o.d.o.l.ScoreIterationListener - Score at iteration 992 is 383.1866651763012
o.d.o.l.ScoreIterationListener - Score at iteration 993 is 381.78825748540373
o.d.o.l.ScoreIterationListener - Score at iteration 994 is 380.0387015850865
o.d.o.l.ScoreIterationListener - Score at iteration 995 is 378.67470982690736
o.d.o.l.ScoreIterationListener - Score at iteration 996 is 376.96384785301404
o.d.o.l.ScoreIterationListener - Score at iteration 997 is 375.642910710587
o.d.o.l.ScoreIterationListener - Score at iteration 998 is 373.96674036087035
o.d.o.l.ScoreIterationListener - Score at iteration 999 is 372.6741075529085
```

```
o.d.o.l.ScoreIterationListener - Score at iteration 1969 is 44.70903950884105
o.d.o.l.ScoreIterationListener - Score at iteration 1970 is 45.2380225365112
o.d.o.l.ScoreIterationListener - Score at iteration 1971 is 45.8889230951459
o.d.o.l.ScoreIterationListener - Score at iteration 1972 is 46.46854959043279
o.d.o.l.ScoreIterationListener - Score at iteration 1973 is 47.17536573123725
o.d.o.l.ScoreIterationListener - Score at iteration 1974 is 47.81004132107224
o.d.o.l.ScoreIterationListener - Score at iteration 1975 is 48.57575742663159
o.d.o.l.ScoreIterationListener - Score at iteration 1976 is 49.27007691834137
o.d.o.l.ScoreIterationListener - Score at iteration 1977 is 50.101039358040985
o.d.o.l.ScoreIterationListener - Score at iteration 1978 is 50.85791943489326
o.d.o.l.ScoreIterationListener - Score at iteration 1979 is 51.757361700603106
o.d.o.l.ScoreIterationListener - Score at iteration 1980 is 52.58158597894784
o.d.o.l.ScoreIterationListener - Score at iteration 1981 is 53.556149354377496
o.d.o.l.ScoreIterationListener - Score at iteration 1982 is 54.453351759736925
o.d.o.l.ScoreIterationListener - Score at iteration 1983 is 55.506758667805
o.d.o.l.ScoreIterationListener - Score at iteration 1984 is 56.4806440514167
o.d.o.l.ScoreIterationListener - Score at iteration 1985 is 57.61902297956854
o.d.o.l.ScoreIterationListener - Score at iteration 1986 is 58.67646590078081
o.d.o.l.ScoreIterationListener - Score at iteration 1987 is 59.90408701956823
o.d.o.l.ScoreIterationListener - Score at iteration 1988 is 61.05119425005335
o.d.o.l.ScoreIterationListener - Score at iteration 1989 is 62.37603870187002
o.d.o.l.ScoreIterationListener - Score at iteration 1990 is 63.61836167255973
o.d.o.l.ScoreIterationListener - Score at iteration 1991 is 65.04312191353905
o.d.o.l.ScoreIterationListener - Score at iteration 1992 is 66.38540283674381
o.d.o.l.ScoreIterationListener - Score at iteration 1993 is 67.91764984573355
o.d.o.l.ScoreIterationListener - Score at iteration 1994 is 69.36393101040994
o.d.o.l.ScoreIterationListener - Score at iteration 1995 is 71.01005259072566
o.d.o.l.ScoreIterationListener - Score at iteration 1996 is 72.5647220289928
o.d.o.l.ScoreIterationListener - Score at iteration 1997 is 74.3255814684223
o.d.o.l.ScoreIterationListener - Score at iteration 1998 is 75.99247402491774
o.d.o.l.ScoreIterationListener - Score at iteration 1999 is 77.8737141122287
```

The screenshot displays the Node-RED web interface. On the left, the 'input' nodes list includes inject, catch, status, link, mqtt, http, websocket, tcp, mqlight, and ibmiot. The 'output' nodes list includes debug, link, mqtt, and http response. The main workspace shows a flow with the following nodes: 'timestamp', 'healthy', 'broken' (circled in red), 'reset', 'msg.payload', 'f' (function), 'limit to max 3000', and two 'IBM IoT' nodes. The 'broken' node is connected to a function node, which then connects to the 'limit to max 3000' node. The 'limit to max 3000' node connects to an 'IBM IoT' node. The 'timestamp' node connects to a function node, which connects to a 'msg.payload' node. The 'reset' node connects to a function node, which connects to a 'msg.payload' node. The right sidebar shows the 'debug' console with the following error messages:

```
function : (error)
"TypeError: Cannot read property 'temp' of undefined"
07/07/2017, 23:46:21 node: device data
iot-2/type/0.16.2/id/lorenz/ev/osc/rtm/json : msg : Object
+ { topic: "iot-2/type/0.16.2/id/lorenz/ev/", payload: object, deviceId: "lorenz", deviceType: "0.16.2", eventType: "osc" }
07/07/2017, 23:46:21 node: 64fa444.511944
iot-2/type/0.16.2/id/lorenz/ev/osc/rtm/json : msg.payload : Object
+ { x: 12.86756805837483, y: 17.767281997817488, z: 27.52203015992025 }
function : (error)
"TypeError: Cannot read property 'temp' of undefined"
07/07/2017, 23:46:21 node: device data
iot-2/type/0.16.2/id/lorenz/ev/osc/rtm/json : msg : Object
+ { topic: "iot-2/type/0.16.2/id/lorenz/ev/", payload: object, deviceId: "lorenz", deviceType: "0.16.2", eventType: "osc" }
07/07/2017, 23:46:21 node: 64fa444.511944
iot-2/type/0.16.2/id/lorenz/ev/osc/rtm/json : msg.payload : Object
+ { x: 13.259538549529597, y: 17.675845018101423, z: 28.80988190283415 }
```

```
o.d.o.l.ScoreIterationListener - Score at iteration 2969 is 10891.930373337565
o.d.o.l.ScoreIterationListener - Score at iteration 2970 is 10897.650306476724
o.d.o.l.ScoreIterationListener - Score at iteration 2971 is 10903.295572131516
o.d.o.l.ScoreIterationListener - Score at iteration 2972 is 10909.297284415812
o.d.o.l.ScoreIterationListener - Score at iteration 2973 is 10915.302549022963
o.d.o.l.ScoreIterationListener - Score at iteration 2974 is 10921.546926582616
o.d.o.l.ScoreIterationListener - Score at iteration 2975 is 10927.696193981876
o.d.o.l.ScoreIterationListener - Score at iteration 2976 is 10934.121015271206
o.d.o.l.ScoreIterationListener - Score at iteration 2977 is 10940.642727116407
o.d.o.l.ScoreIterationListener - Score at iteration 2978 is 10947.279991840644
o.d.o.l.ScoreIterationListener - Score at iteration 2979 is 10953.833701577507
o.d.o.l.ScoreIterationListener - Score at iteration 2980 is 10960.651412106614
o.d.o.l.ScoreIterationListener - Score at iteration 2981 is 10967.415789221683
o.d.o.l.ScoreIterationListener - Score at iteration 2982 is 10974.417499744583
o.d.o.l.ScoreIterationListener - Score at iteration 2983 is 10981.178323042242
o.d.o.l.ScoreIterationListener - Score at iteration 2984 is 10988.440476587411
o.d.o.l.ScoreIterationListener - Score at iteration 2985 is 10995.3844071995
o.d.o.l.ScoreIterationListener - Score at iteration 2986 is 11002.536337484873
o.d.o.l.ScoreIterationListener - Score at iteration 2987 is 11009.585159062908
o.d.o.l.ScoreIterationListener - Score at iteration 2988 is 11016.66864489814
o.d.o.l.ScoreIterationListener - Score at iteration 2989 is 11023.745910911426
o.d.o.l.ScoreIterationListener - Score at iteration 2990 is 11030.880953683412
o.d.o.l.ScoreIterationListener - Score at iteration 2991 is 11037.787551632187
o.d.o.l.ScoreIterationListener - Score at iteration 2992 is 11044.859483622726
o.d.o.l.ScoreIterationListener - Score at iteration 2993 is 11051.695858331497
o.d.o.l.ScoreIterationListener - Score at iteration 2994 is 11058.648678534704
o.d.o.l.ScoreIterationListener - Score at iteration 2995 is 11065.278833431727
o.d.o.l.ScoreIterationListener - Score at iteration 2996 is 11072.15876470129
o.d.o.l.ScoreIterationListener - Score at iteration 2997 is 11078.420031398582
o.d.o.l.ScoreIterationListener - Score at iteration 2998 is 11085.078628689993
o.d.o.l.ScoreIterationListener - Score at iteration 2999 is 11091.125671441947
```

```
Romeos-MacBook-Pro:dl4j-examples-scala romeokienzler$ mvn package
[INFO] Scanning for projects...
[INFO]
[INFO] -----
[INFO] Building DeepLearning4j Examples 0.8-SNAPSHOT
[INFO] -----
```

```
[INFO] Attaching shaded artifact.
[INFO]
[INFO] --- maven-assembly-plugin:2.4:single (make-jar-with-dependencies) @ dl4j-examples-scala ---
[INFO] Building jar: /Users/romeokienzler/Documents/tmp/deleteme6/deeplearning4j-examples-parent/dl4j-examples-scala/target/dl4j-examples-scala-0.8-SNAPSHOT-jar-with-dependencies.jar
[INFO] -----
[INFO] BUILD SUCCESS
[INFO] -----
[INFO] Total time: 02:25 min
[INFO] Finished at: 2017-07-07T16:25:13+02:00
[INFO] Final Memory: 85M/1571M
[INFO] -----
```

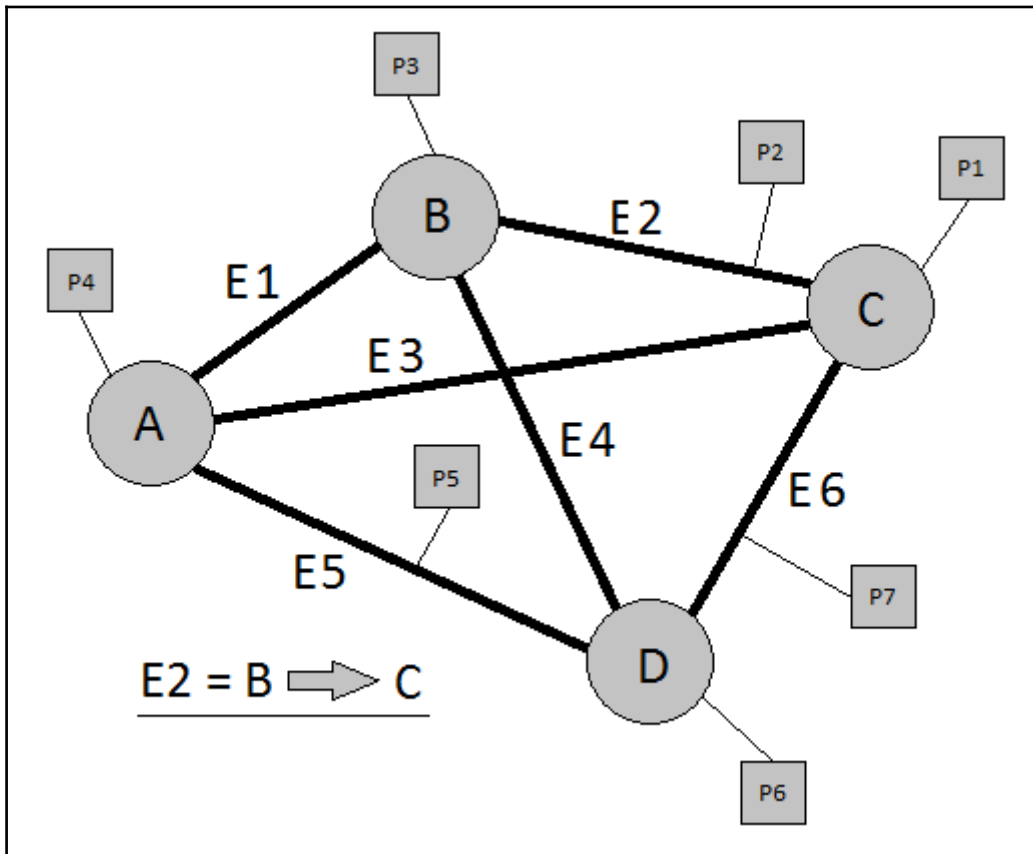
```
Romeos-MacBook-Pro:dl4j-examples-scala romeokienzler$ cd target/
Romeos-MacBook-Pro:target romeokienzler$ scp -P 2222 dl4j-examples-scala-0.8-SNAPSHOT-jar-with-dependencies.jar root@localhost:/root/
root@localhost's password:
dl4j-examples-scala-0.8-SNAPSHOT-jar-with-dependencies.jar 100% 396MB 93.5MB/s 00:04
```

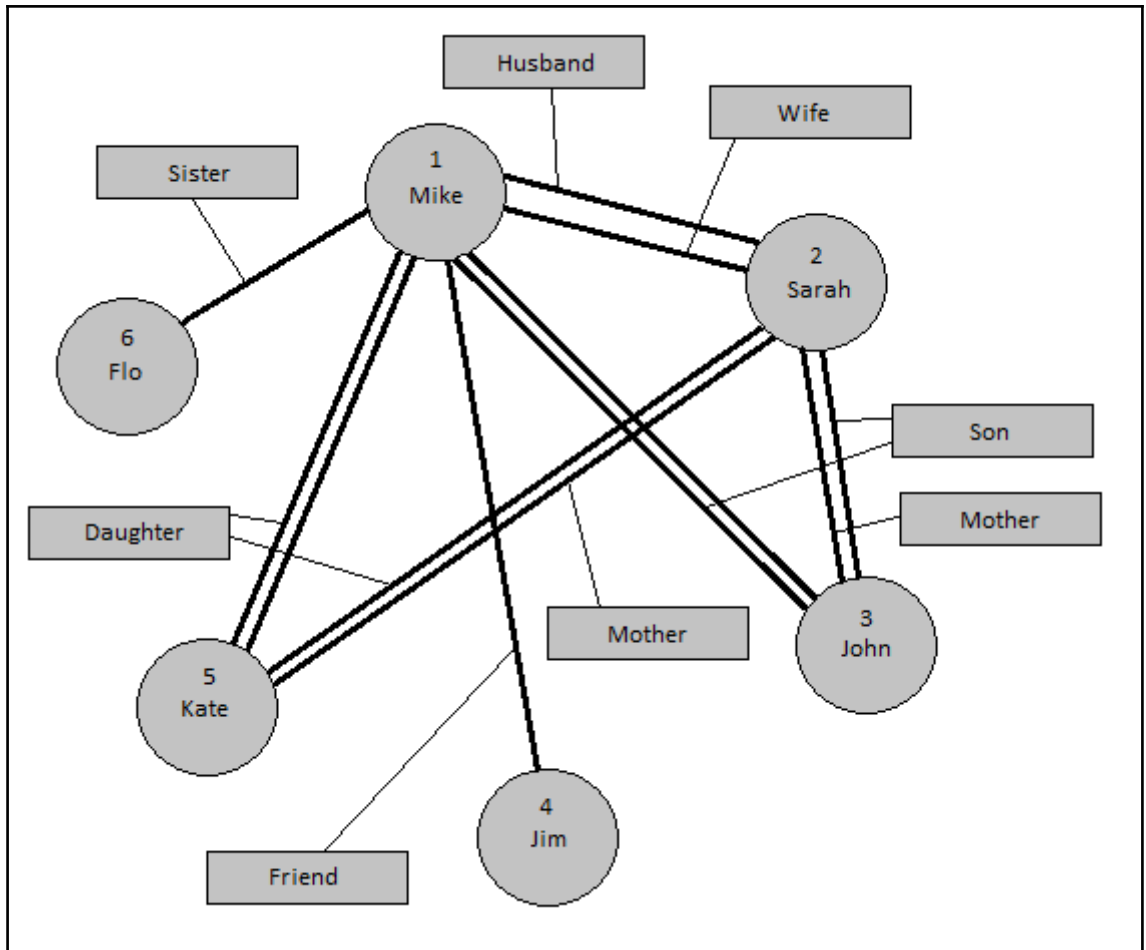
```
[root@sandbox ~]# export SPARK_MAJOR_VERSION=2
[root@sandbox ~]# spark-submit --master yarn --num-executors 1 --driver-memory 1000m --executor-memory 1250m --executor-cores 1 --class Run dl4j-examples-scala-0.8-SNAPSHOT-jar-with-dependencies.jar true
SPARK_MAJOR_VERSION is set to 2, using Spark2
17/07/07 17:57:33 INFO HdfsBackend: Loaded [cpuBackend] backend
17/07/07 17:57:33 INFO NativeOpsHolder: Number of threads used for NativeOps: 4
17/07/07 17:57:34 INFO Reflections: Reflections took 349 ms to scan 1 urls, producing 29 keys and 189 values
17/07/07 17:57:34 INFO HdfsBackend: Number of threads used for HDFS: 4
17/07/07 17:57:34 INFO DefaultOpExecutor: Backend used: [CPU]; OS: [Linux]
17/07/07 17:57:34 INFO DefaultOpExecutor: Cores: [4]; Memory: [0.9GB];
17/07/07 17:57:34 INFO DefaultOpExecutor: Blas vendor: [OPENBLAS]
```

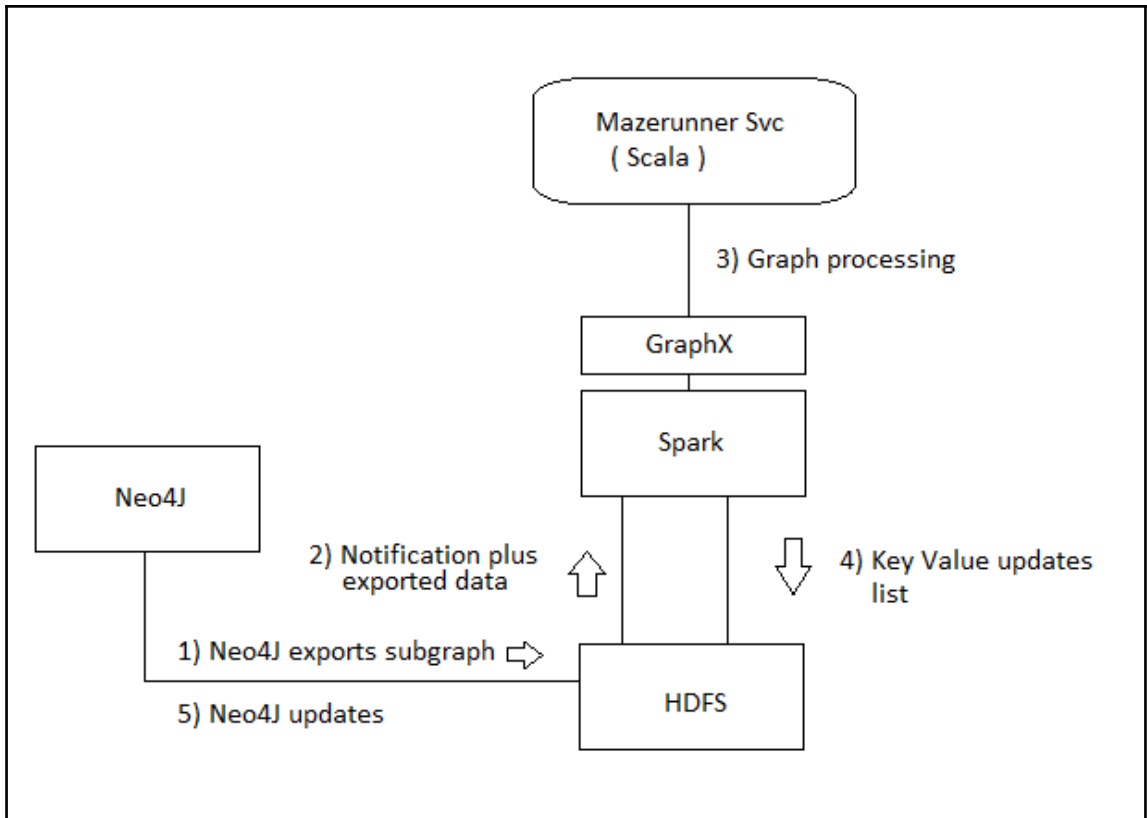
```
17/07/07 14:52:08 INFO LoggerUtility: main: Org ID = rwyrty
Client ID = a:rwyrty:a2g6k39sl6r5
17/07/07 14:52:09 INFO LoggerUtility: main: Connecting client a:rwyrty:a2g6k39sl6r5 to ssl://rwyrty.messaging.internetofthings.ibmcloud.com:8883 (attempt #1)...
17/07/07 14:52:10 INFO LoggerUtility: main: Successfully connected to the IBM Watson IoT Platform
Mainthread blocking...
Mainthread blocking...
Mainthread blocking...
Waiting for tumbling window to fill: 0
Waiting for tumbling window to fill: 100
Waiting for tumbling window to fill: 200
Waiting for tumbling window to fill: 300
Mainthread blocking...
Waiting for tumbling window to fill: 400
Waiting for tumbling window to fill: 500
```



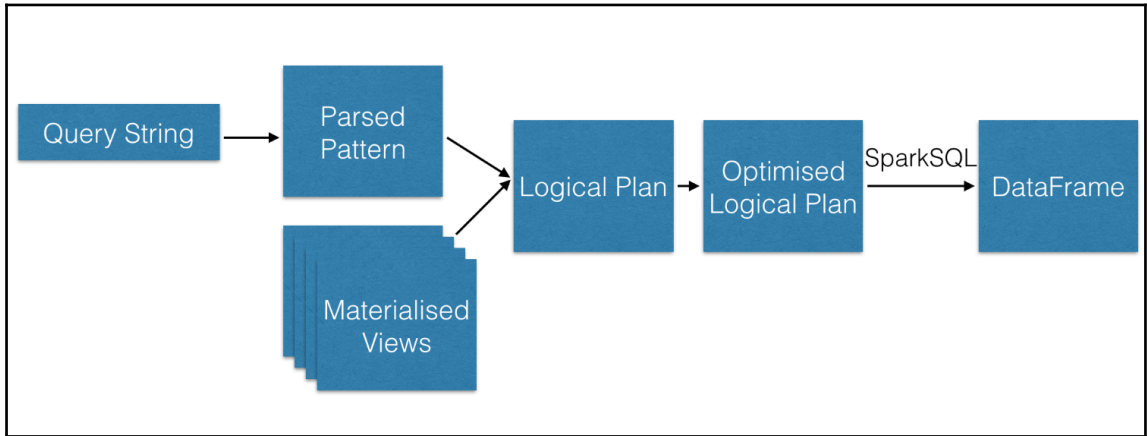
# Apache Spark GraphX

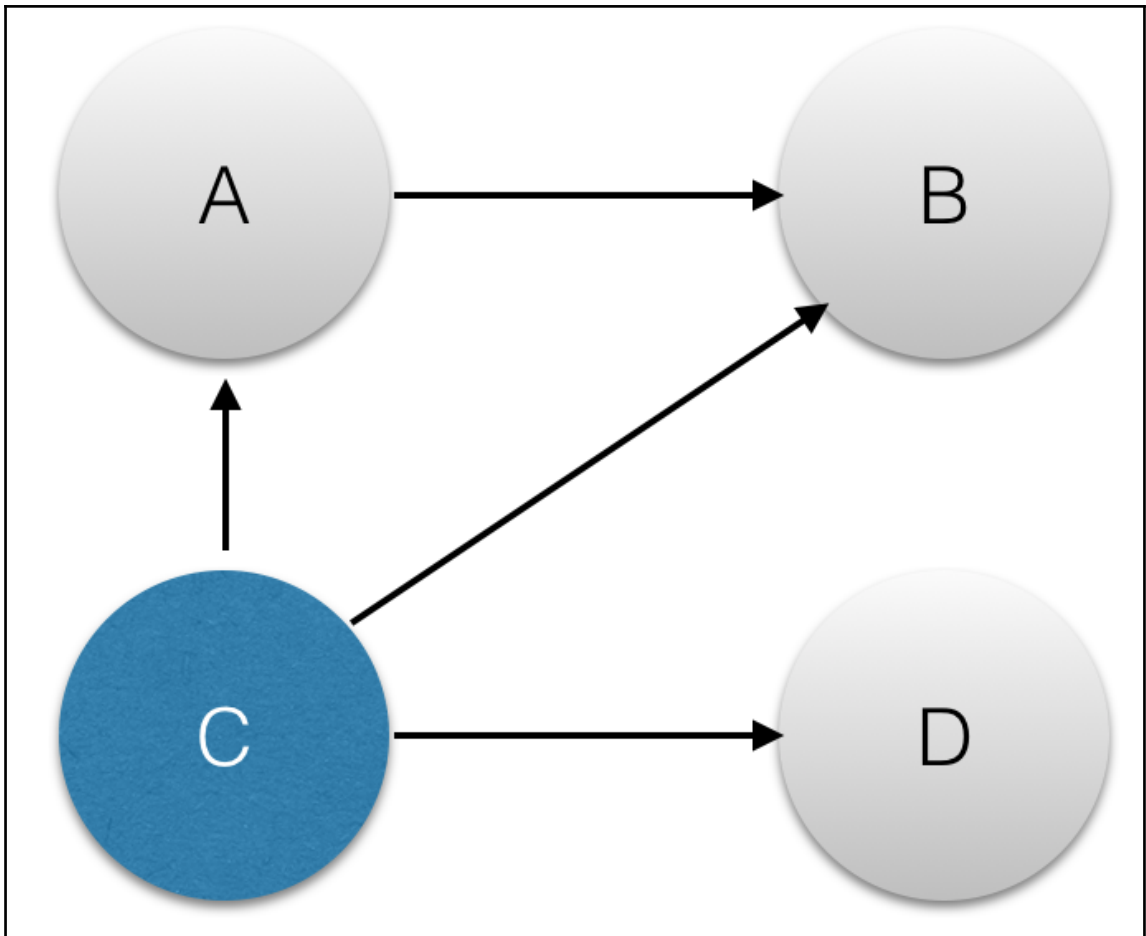


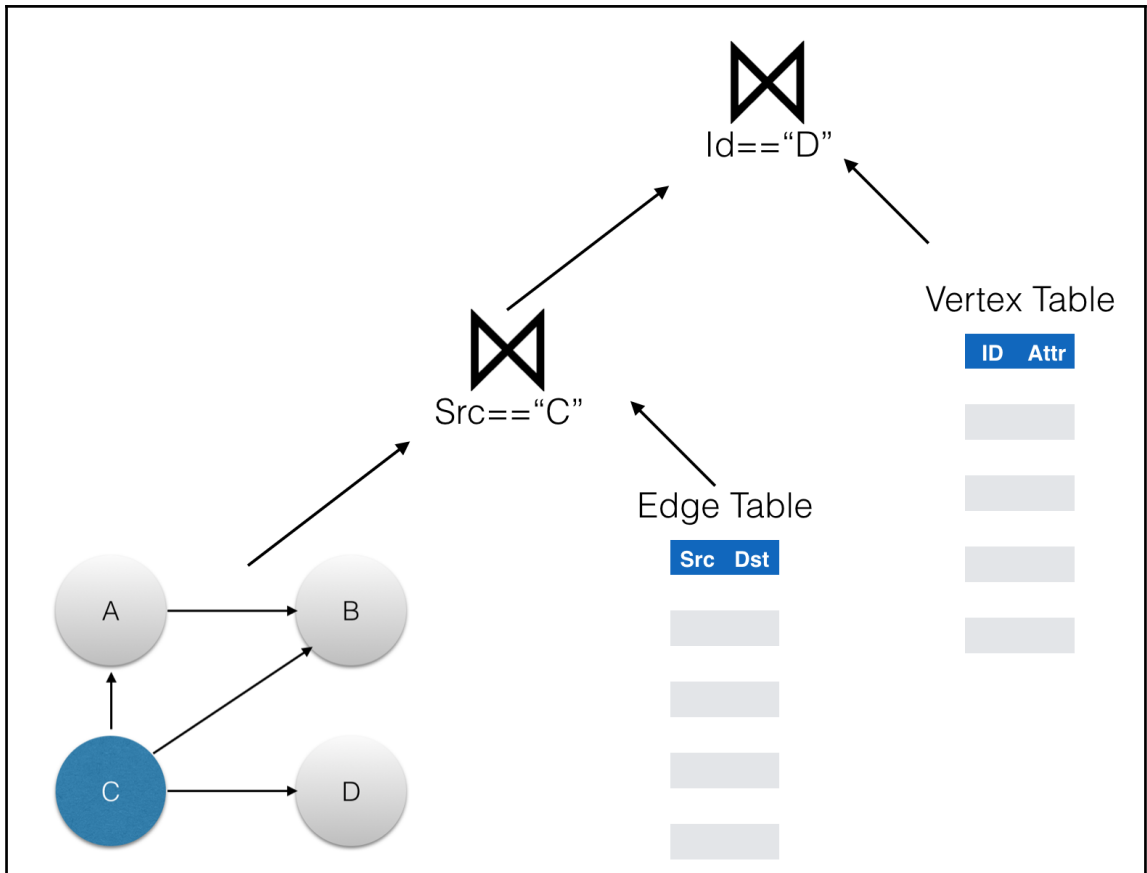


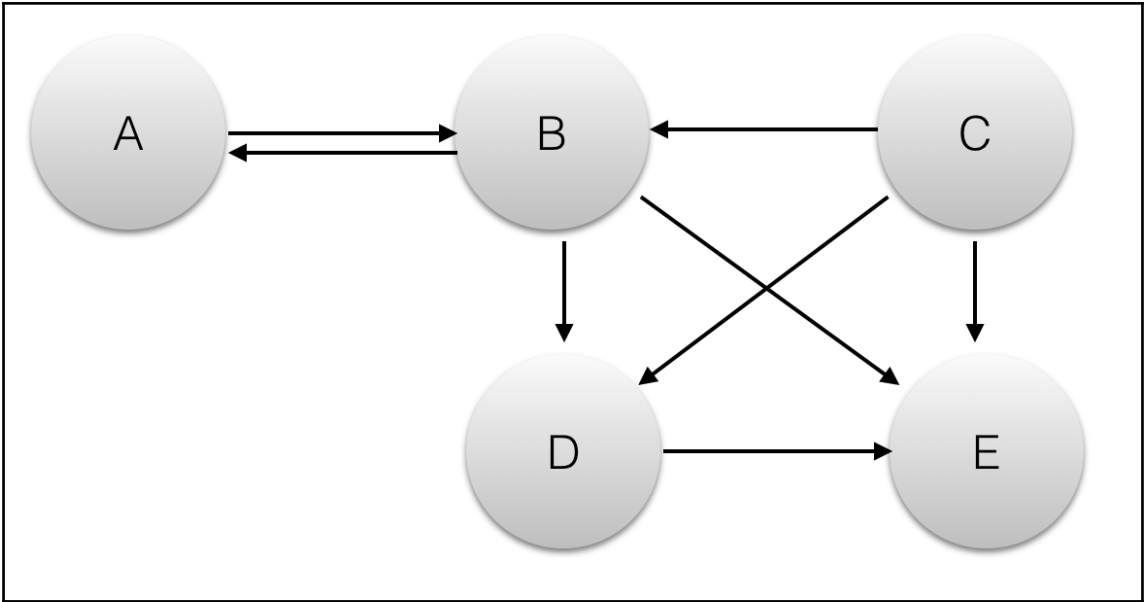


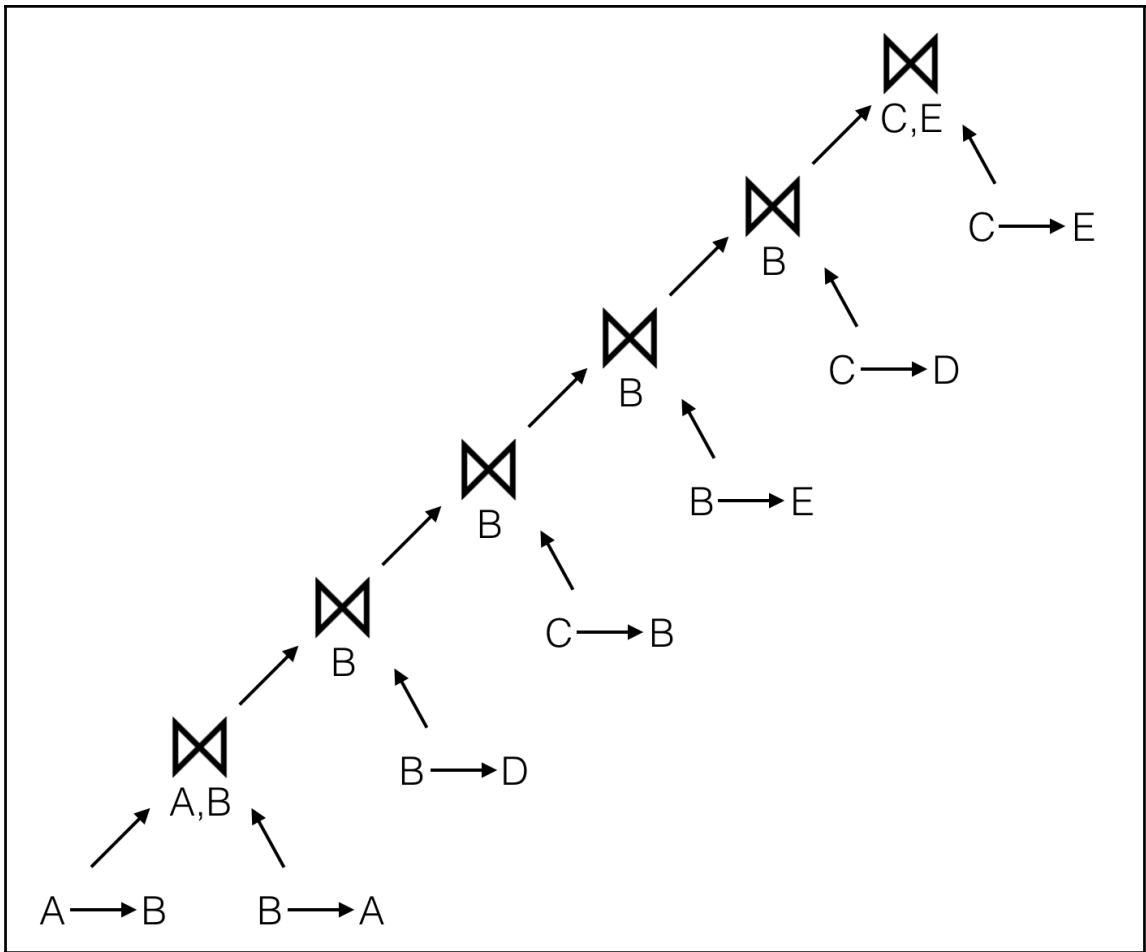
# Apache Spark GraphFrames



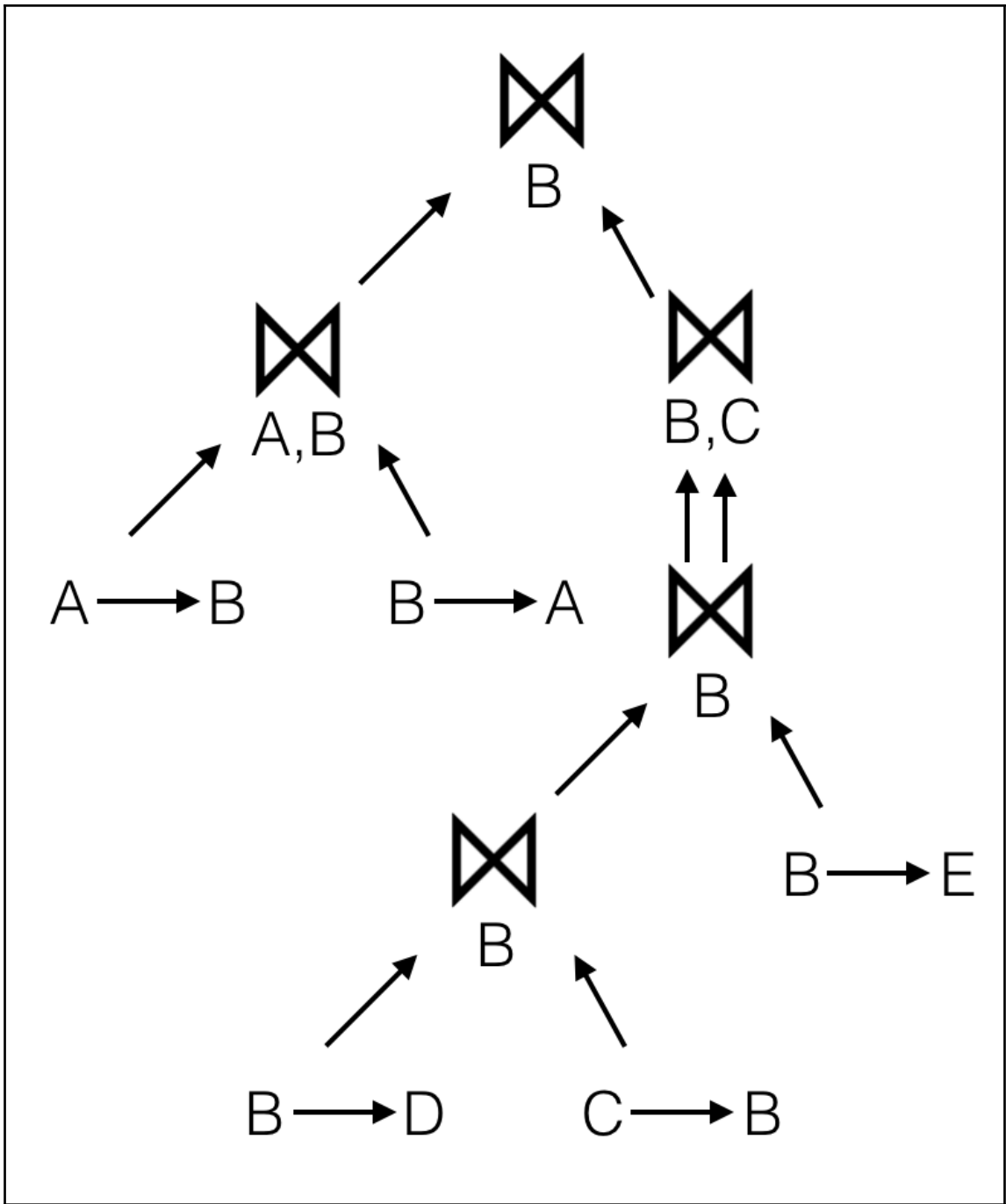










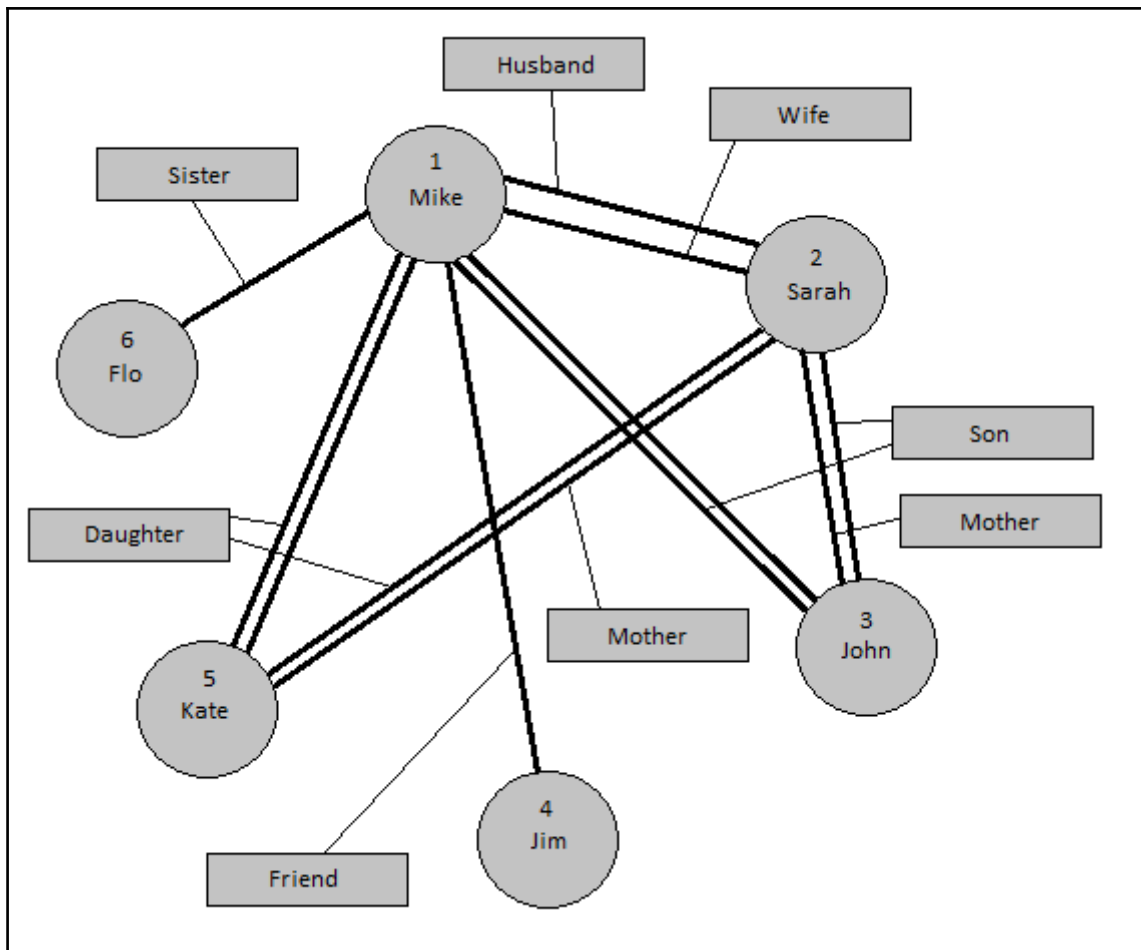


```
[scala> graph.vertices.filter("attr > 40").show
+---+-----+-----+
| id| name|attr|
+---+-----+-----+
|  1| Mike| 48|
|  2| Sarah| 45|
|  4|  Jim| 53|
|  6|  Flo| 52|
+---+-----+-----+
```

```
+---+-----+-----+-----+-----+
| id| name|attr| pagerank|
+---+-----+-----+-----+-----+
|  1| Mike| 48|1.7447770383026542|
|  2| Sarah| 45|1.5460757395935596|
|  5| Kate| 22|1.0800834145716334|
|  3| John| 25|1.0800834145716334|
|  4|  Jim| 53|          0.15|
|  6|  Flo| 52|          0.15|
+---+-----+-----+-----+-----+
```

```
scala> results.select("id", "count").show()
```

```
+---+-----+  
| id|count|  
+---+-----+  
|  3|    1|  
|  5|    1|  
|  6|    0|  
|  1|    2|  
|  4|    0|  
|  2|    2|  
+---+-----+
```



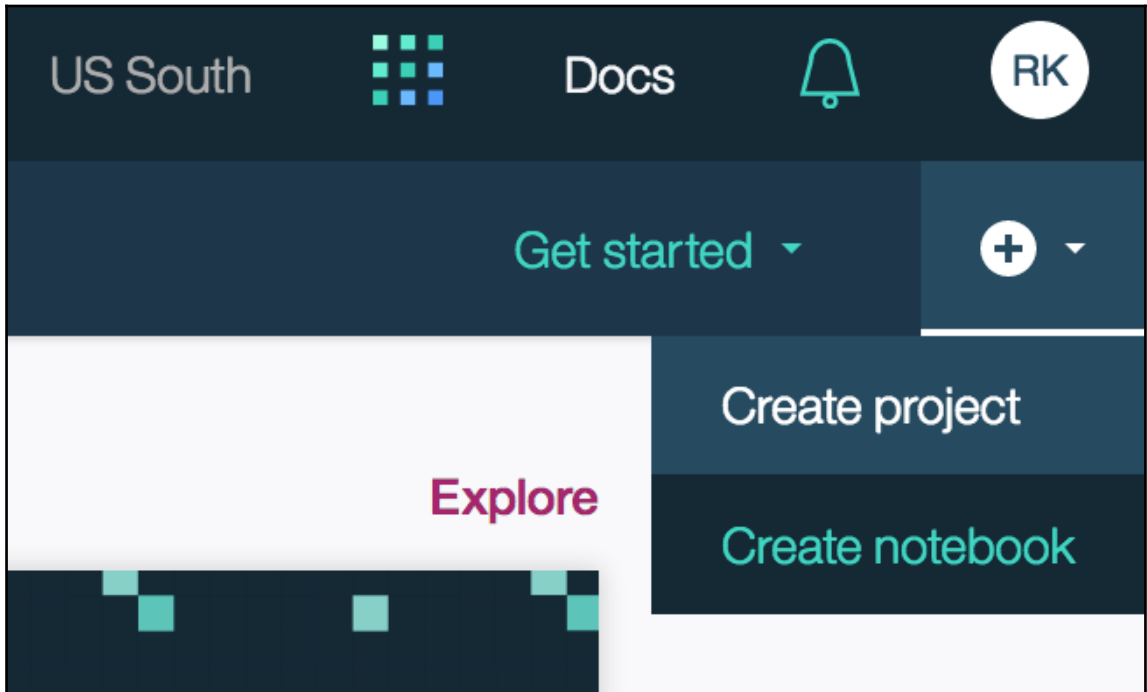
```
scala> result.select("id", "component").orderBy("component").show()
+---+-----+
| id| component|
+---+-----+
|  1|154618822656|
|  2|154618822656|
|  3|154618822656|
|  4|154618822656|
|  5|154618822656|
|  6|154618822656|
+---+-----+
```

```
scala> result.select("id", "component").orderBy("component").show()
17/06/18 07:21:05 WARN Executor: 1 block locks were not released by TID = 13624:
[rdd_1002_0]
+----+-----+
| id|   component|
+----+-----+
|  5| 154618822656|
|  2| 154618822656|
|  3| 154618822656|
|  1| 154618822656|
|  6| 644245094400|
|  4|1425929142272|
+----+-----+
```

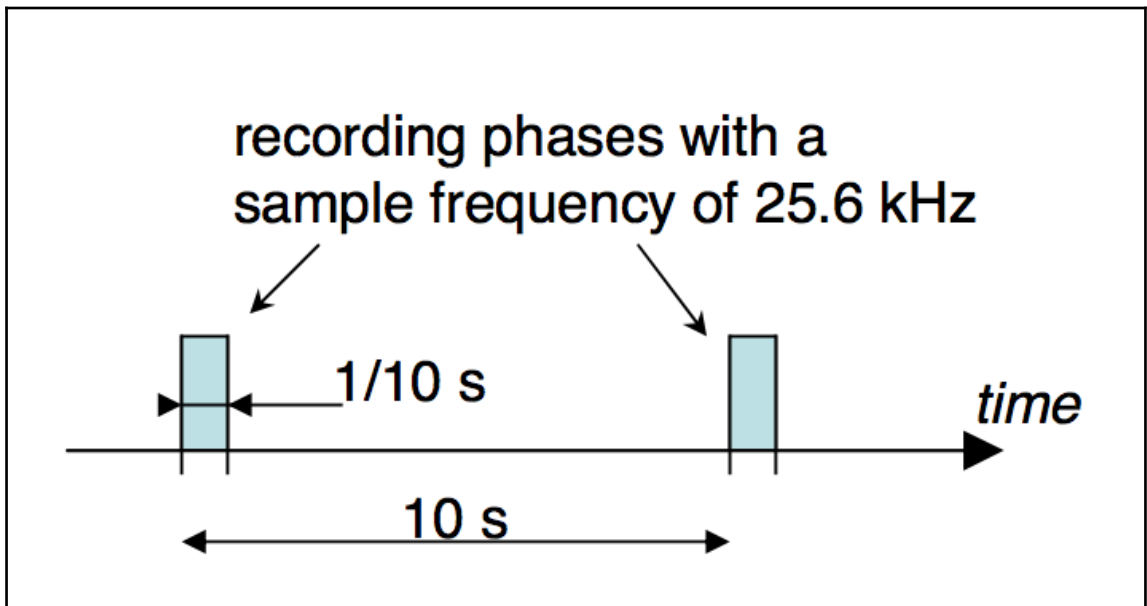
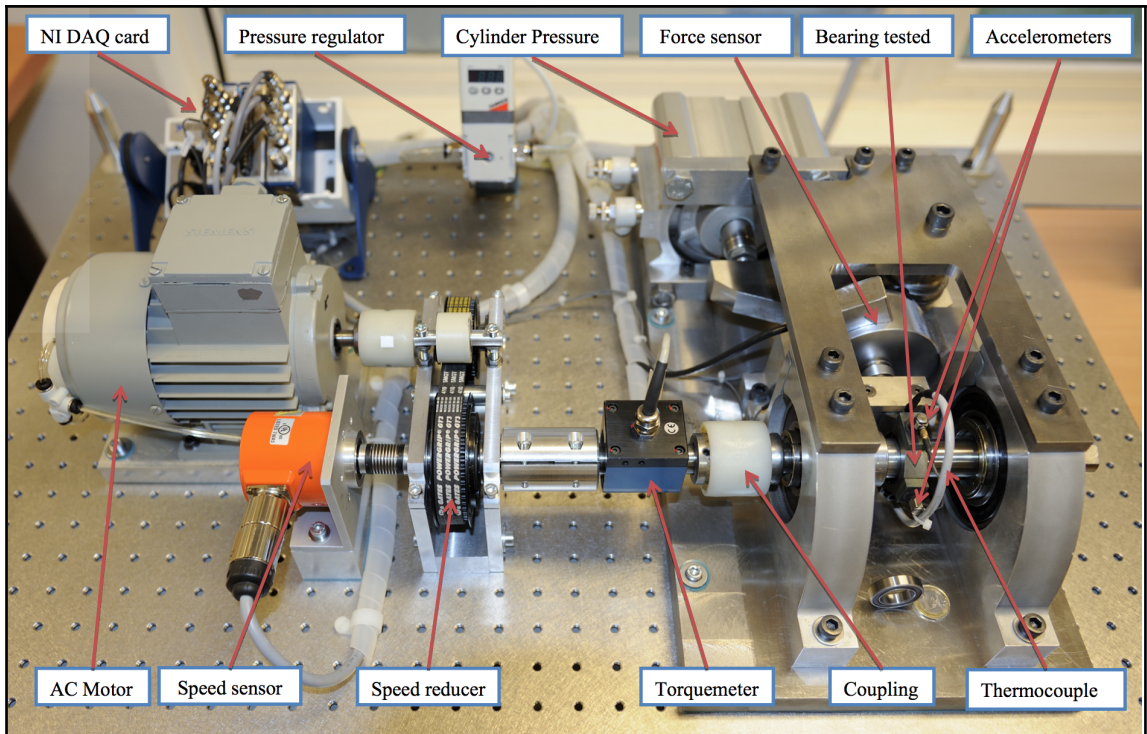
```
scala> result.select("id", "component").orderBy("component").show()
17/06/18 07:24:25 WARN Executor: 1 block locks were not released by TID = 8717:
[rdd_545_0]
+----+-----+
| id|   component|
+----+-----+
|  7| 25769803776|
|  1| 154618822656|
|  2| 154618822656|
|  3| 154618822656|
|  5| 154618822656|
|  6| 644245094400|
|  4|1425929142272|
+----+-----+
```

```
scala> result.select("id", "component").orderBy("component").show()
+----+-----+
| id|   component|
+----+-----+
|  7| 25769803776|
|  1|154618822656|
|  3|154618822656|
|  2|154618822656|
|  4|154618822656|
|  5|154618822656|
|  6|154618822656|
+----+-----+
```

# Apache Spark with Jupyter Notebooks on IBM DataScience Experience



The screenshot shows the 'Create Notebook' page in the IBM Data Science Experience. The top navigation bar includes a hamburger menu, the IBM Data Science Experience logo, and a user profile icon labeled 'RK'. Below the navigation bar, the breadcrumb 'My Projects > New Notebook' is visible. The main heading is 'Create Notebook', with three tabs: 'Blank' (selected), 'From File', and 'From URL'. The 'Name\*' field contains 'packt\_hello\_world' and shows '33 Characters Remaining'. The 'Description' field is empty with the placeholder text 'Type your Description here'. The 'Language\*' section has radio buttons for 'Python 2', 'R', 'Scala' (selected), and 'Python 3.5 Experimental'. The 'Spark version\*' section has radio buttons for '2.1' (selected), '2.0', and '1.6'. The 'Project' dropdown is set to 'Default Project' with a downward arrow. Below it, the text 'Add the notebook to an existing project.' is displayed. The 'Spark Service\*' dropdown is set to 'DSX-Spark' with a downward arrow. Below it, the text 'Associate this notebook with the Spark Service of your choice.' is displayed. At the bottom right, there are 'Cancel' and 'Create Notebook' buttons. A chat icon is located in the bottom right corner of the page.





We download the zip file containing the data to the local stagin area

```
In [*]: import sys.process._
        "wget http://www.femto-st.fr/f/d/Training_set.zip" !
```

```
--2017-06-27 17:31:57-- http://www.femto-st.fr/f/d/Training_set.zip
Resolving www.femto-st.fr (www.femto-st.fr)... 195.83.19.10
Connecting to www.femto-st.fr (www.femto-st.fr)|195.83.19.10|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 140424205 (134M) [application/zip]
Saving to: 'Training_set.zip.1'

 0K ..... 0% 96.4K 23m42s
 50K ..... 0% 384K 14m49s
100K ..... 0% 385K 11m51s
150K ..... 0% 385K 10m22s
200K ..... 0% 81.9M 8m18s
250K ..... 0% 7.70M 6m58s
300K ..... 0% 405K 6m46s
350K ..... 0% 7.23M 5m57s
400K ..... 0% 386K 5m57s
450K ..... 0% 77.9M 5m21s
500K ..... 0% 40.9M 4m52s
```

```
In [3]: "df -h" !
```

Filesystem	Size	Used	Avail	Use%	Mounted on
/dev/sda3	930G	90G	840G	10%	/
devtmpfs	189G	0	189G	0%	/dev
tmpfs	189G	0	189G	0%	/dev/shm
tmpfs	189G	4.1G	185G	3%	/run
tmpfs	189G	0	189G	0%	/sys/fs/cgroup
/dev/sdb1	3.6T	9.1G	3.4T	1%	/disk1
/dev/sdf1	3.6T	89M	3.4T	1%	/disk5
/dev/sdj1	3.6T	89M	3.4T	1%	/disk9
/dev/sdd1	3.6T	89M	3.4T	1%	/disk3
/dev/sde1	3.6T	89M	3.4T	1%	/disk4
/dev/sdc1	3.6T	89M	3.4T	1%	/disk2
/dev/sdi1	3.6T	89M	3.4T	1%	/disk8
/dev/sdg1	3.6T	89M	3.4T	1%	/disk6
/dev/sdh1	3.6T	89M	3.4T	1%	/disk7
/dev/sda1	253M	163M	91M	65%	/boot
tmpfs	38G	0	38G	0%	/run/user/0
/dev/fs01	246T	173T	73T	71%	/gpfs/global_fs01

```
In [*]: "unzip ./Training_set.zip" !
```

```
Archive:  ./Training_set.zip
  creating: Learning_set/
  creating: Learning_set/Bearing1_1/
 inflating: Learning_set/Bearing1_1/acc_00001.csv
 inflating: Learning_set/Bearing1_1/acc_00002.csv
 inflating: Learning_set/Bearing1_1/acc_00003.csv
 inflating: Learning_set/Bearing1_1/acc_00004.csv
 inflating: Learning_set/Bearing1_1/acc_00005.csv
 inflating: Learning_set/Bearing1_1/acc_00006.csv
 inflating: Learning_set/Bearing1_1/acc_00007.csv
 inflating: Learning_set/Bearing1_1/acc_00008.csv
 inflating: Learning_set/Bearing1_1/acc_00009.csv
 inflating: Learning_set/Bearing1_1/acc_00010.csv
 inflating: Learning_set/Bearing1_1/acc_00011.csv
 inflating: Learning_set/Bearing1_1/acc_00012.csv
 inflating: Learning_set/Bearing1_1/acc_00013.csv
 inflating: Learning_set/Bearing1_1/acc_00014.csv
 inflating: Learning_set/Bearing1_1/acc_00015.csv
 inflating: Learning_set/Bearing1_1/acc_00016.csv
```

```
In [7]: val bearing1_1_acc = spark.read.option("inferSchema", "true").csv("./Learning_set/Bearing1_1/acc*")
```

```
In [8]: bearing1_1_acc.printSchema
```

```
root
 |-- _c0: integer (nullable = true)
 |-- _c1: integer (nullable = true)
 |-- _c2: integer (nullable = true)
 |-- _c3: decimal(5,-1) (nullable = true)
 |-- _c4: double (nullable = true)
 |-- _c5: double (nullable = true)
```

```
In [9]: bearing1_1_acc.show
```

```
+----+-----+-----+-----+-----+
|_c0|_c1|_c2|_c3|_c4|_c5|
+----+-----+-----+-----+-----+
| 9| 38| 46|8.6566E+5|-1.626|-0.086|
| 9| 38| 46|8.6570E+5|-1.538|-0.299|
| 9| 38| 46|8.6574E+5|-0.969|-0.025|
| 9| 38| 46|8.6578E+5|-0.577| 0.008|
| 9| 38| 46|8.6582E+5| 0.143|-0.087|
| 9| 38| 46|8.6586E+5| 0.129|-0.611|
| 9| 38| 46|8.6590E+5| 0.636|-0.496|
| 9| 38| 46|8.6594E+5|-0.129| 0.588|
| 9| 38| 46|8.6598E+5|-0.323| 0.369|
| 9| 38| 46|8.6602E+5|-0.812| 0.019|
| 9| 38| 46|8.6605E+5| -0.8| 0.642|
| 9| 38| 46|8.6609E+5|-0.845|-0.047|
| 9| 38| 46|8.6613E+5|-0.723| 0.117|
| 9| 38| 46|8.6617E+5|-0.527| 0.237|
| 9| 38| 46|8.6621E+5|-0.224| 0.334|
```

```
In [10]: bearing1_1_acc.createOrReplaceTempView("bearing1_1_acc")
val bearing1_1_acc_transformed = spark.sql("""
SELECT concat(_c0,_c1,_c2) as cluster,
(cast(timestamp(concat('1970-01-01 ',_c0,':',_c1,':',_c2,'.123')) as long) *1000000)+_c3 as ts,
_c4 as hacc,
_c5 as vacc
from bearing1_1_acc
""")
```

```
In [ ]: bearing1_1_acc_transformed.write.json("swift://coursera." + name + "/bearing1_1_acc_transformed.json")
```

# Schedule Job

Name \*

ETL Bearing

39 Characters Remaining

Starts on \*

28 June 2017

Description

*Describe what this job is about.*

At time

01:52 AM

Repeats \*

Hourly

Version

A version is saved to your notebook and scheduled.

Ends on

28 June 2017

Summary

"packt\_hello\_world" notebook from scheduled to run hourly starting on Wed, 28 June 2017, 01:52 AM until Wed, 28 June 2017, 11:59 PM.

```
In [28]: df_data_1 = spark.read.json('swift://coursera.' + name + '/bearing1_acc_transformed4.json')
df_data_1.show()
```

SPARK JOB PROGRESS

Hide All ▲

JOB	PROGRESS	DURATION	STATUS
23	1 stage	6.64 sec	▼
24	1 stage	0.59 sec	▼

```
+-----+-----+-----+-----+
|cluster| hacc|      ts|  vacc|
+-----+-----+-----+-----+
| 121149|-0.018|65509065660|-0.077|
| 121149| 0.623|65509065700|-0.189|
| 121149| 0.774|65509065740|-0.424|
| 121149| 0.441|65509065780| 0.749|
| 121149| 0.419|65509065820| 0.08|
| 121149| 0.095|65509065860|-0.183|
| 121149| 0.293|65509065900| 0.282|
| 121149| 0.178|65509065940|-0.059|
| 121149|-0.232|65509065980| 0.594|
| 121149|-0.052|65509066020|-0.338|
| 121149|-0.164|65509066050|-0.092|
| 121149|-0.139|65509066090|-0.111|
| 121149| -0.06|65509066130|-0.854|
| 121149|-0.146|65509066170| 0.705|
| 121149| 0.071|65509066210|-0.005|
| 121149|-0.297|65509066250|-0.115|
| 121149|-0.374|65509066290|-0.146|
| 121149|-0.139|65509066330|-0.099|
| 121149| 0.179|65509066370| 0.518|
| 121149| 0.002|65509066410| 0.053|
+-----+-----+-----+-----+
```

only showing top 20 rows

```
In [29]: import pixiedust
```

```
In [30]: display(df_data_1)
```

Schema

Table

Search table

Showing 100 of 7175680

cluster	hacc	ts	vacc
12149	-0.018	6550906560	-0.077
12149	0.623	65509065700	-0.169
12149	0.774	65509065740	-0.434
12149	0.441	65509065780	0.749
12149	0.419	65509065820	0.08
12149	0.065	65509065860	-0.183
12149	0.283	65509065900	0.282
12149	0.178	65509065940	-0.059
12149	-0.232	65509065980	0.594
12149	-0.062	65509066020	-0.338
12149	-0.164	65509066060	-0.092
12149	-0.139	65509066100	-0.111
12149	-0.06	65509066130	-0.854
12149	-0.146	65509066170	0.706
12149	0.071	65509066210	-0.005
12149	-0.267	65509066250	-0.115
12149	-0.374	65509066290	-0.146
12149	-0.139	65509066330	-0.099
12149	0.179	65509066370	0.518
12149	0.002	65509066410	0.053

## Pixiedust: Line Chart Options

Chart Title:

**Fields:** Show only numeric columns

Search/Filter Fields

- cluster *string*
- hacc *numeric*
- ts *numeric*
- vacc *numeric*

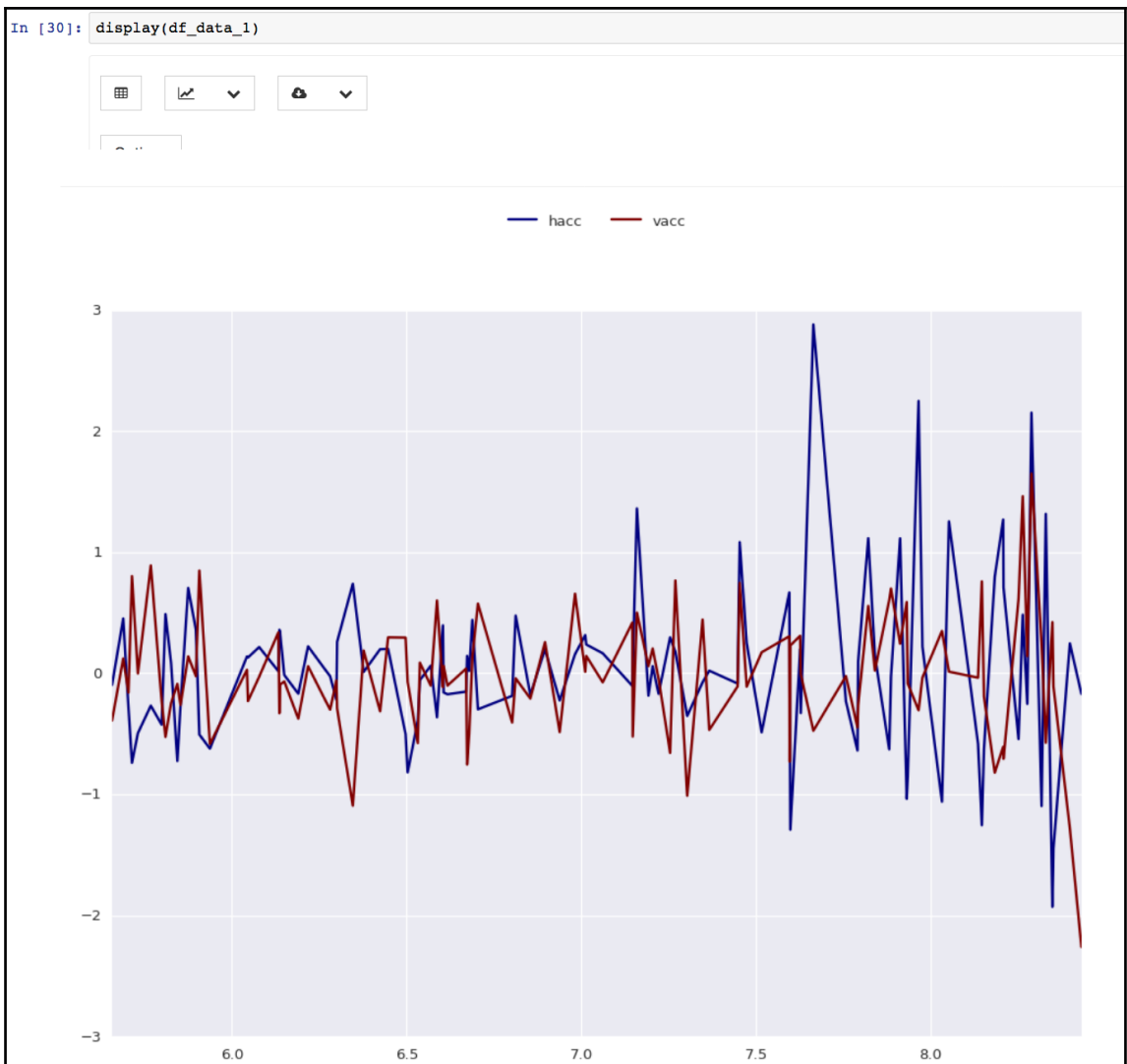
**Keys:**

- ts

**Values:**

- vacc
- hacc

Aggregation:  # of Rows to Display:



```
In [2]: df.data.1 <- read.json(paste("swift://", "coursera", ".", name, "/", "bearing1_acc_transformed4.json", sep=""),
                               source = "org.apache.spark.sql.execution.datasources.csv.CSVFileFormat", header = "true")
head(df.data.1)
```

cluster	hacc	ts	vacc
121149	-0.018	65509065660	-0.077
121149	0.623	65509065700	-0.189
121149	0.774	65509065740	-0.424
121149	0.441	65509065780	0.749
121149	0.419	65509065820	0.080
121149	0.095	65509065860	-0.183

```
In [13]: df_grouped = sql("
        select cluster,
        mean(hacc) as mhacc,
        mean(vacc) as mvacc,
        STDDEV_POP(hacc) as sdhacc,
        STDDEV_POP(vacc) as sdvacc
        from data
        group by cluster
        order by cluster asc")
```

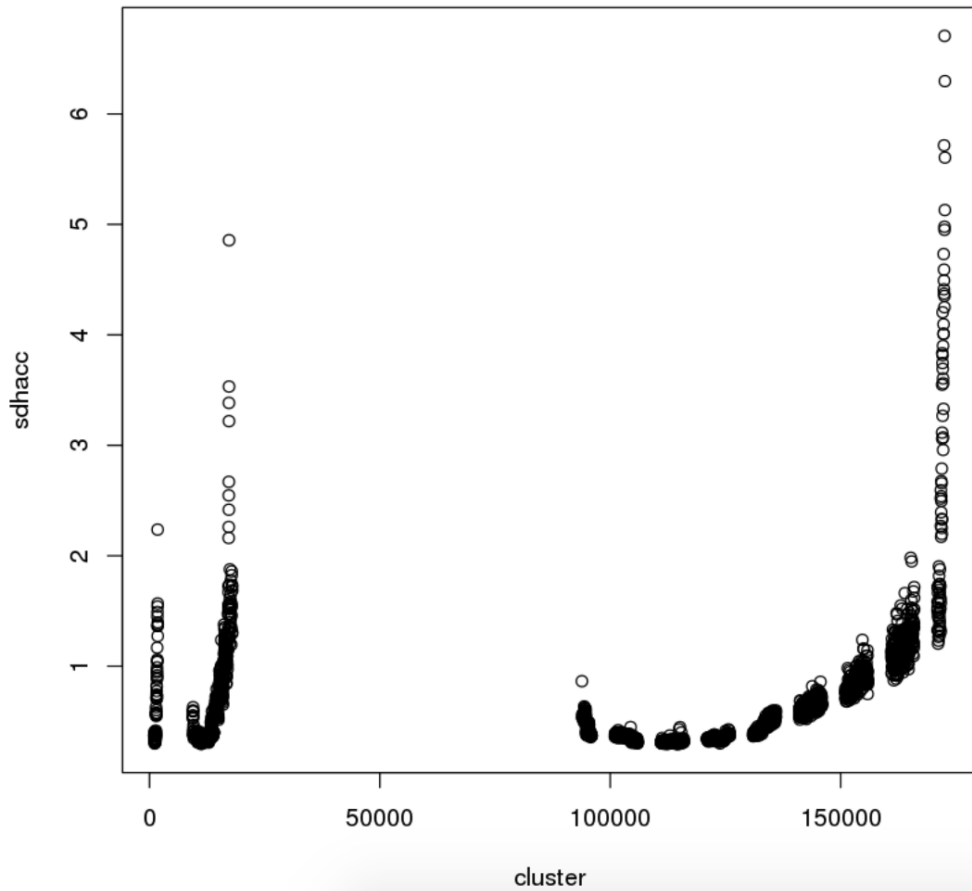
```
In [8]: df_grouped_local = collect(df_grouped)
```



```
In [12]: attach(df_grouped_local)
plot(cluster, sdhacc)
detach(df_grouped_local)
```

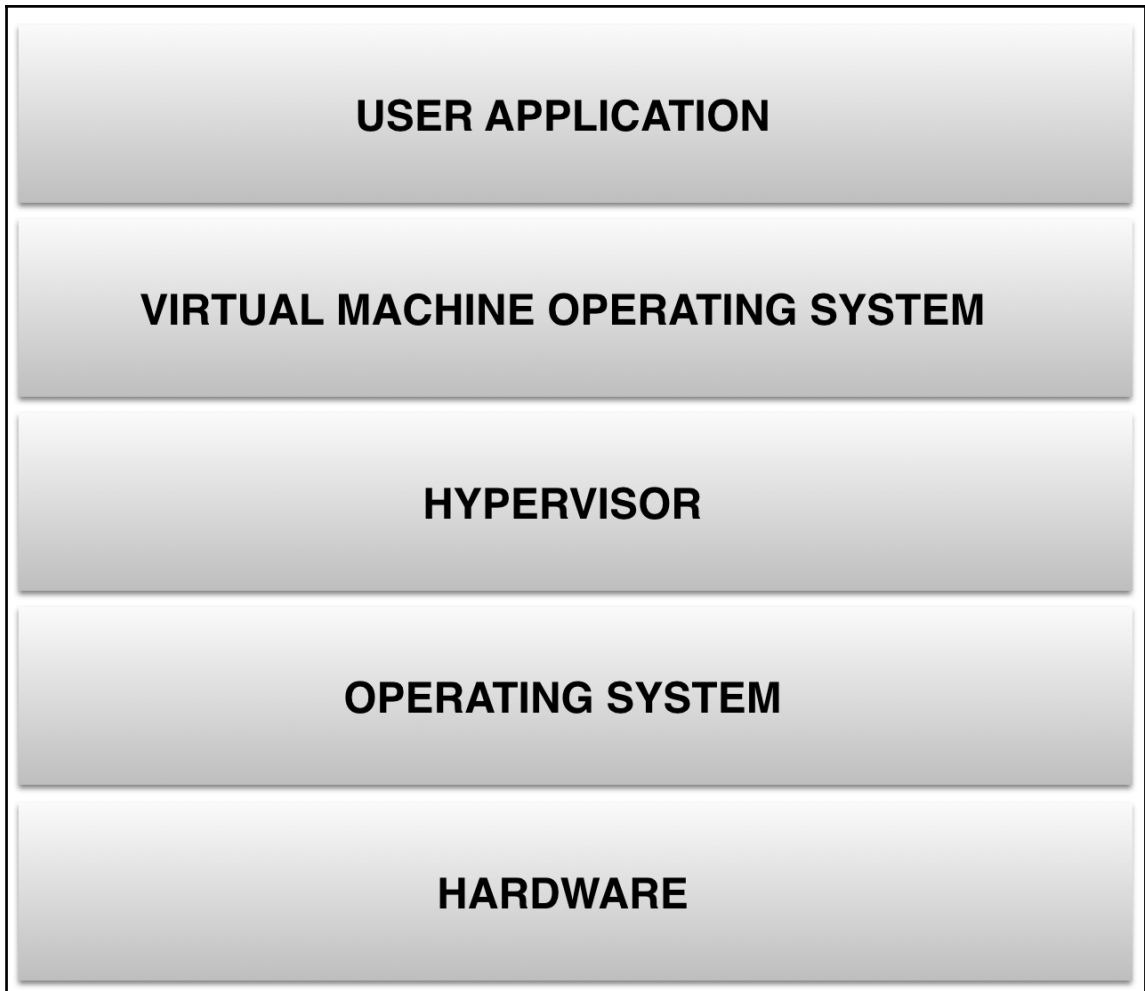
The following objects are masked from df\_grouped\_local (pos = 3):

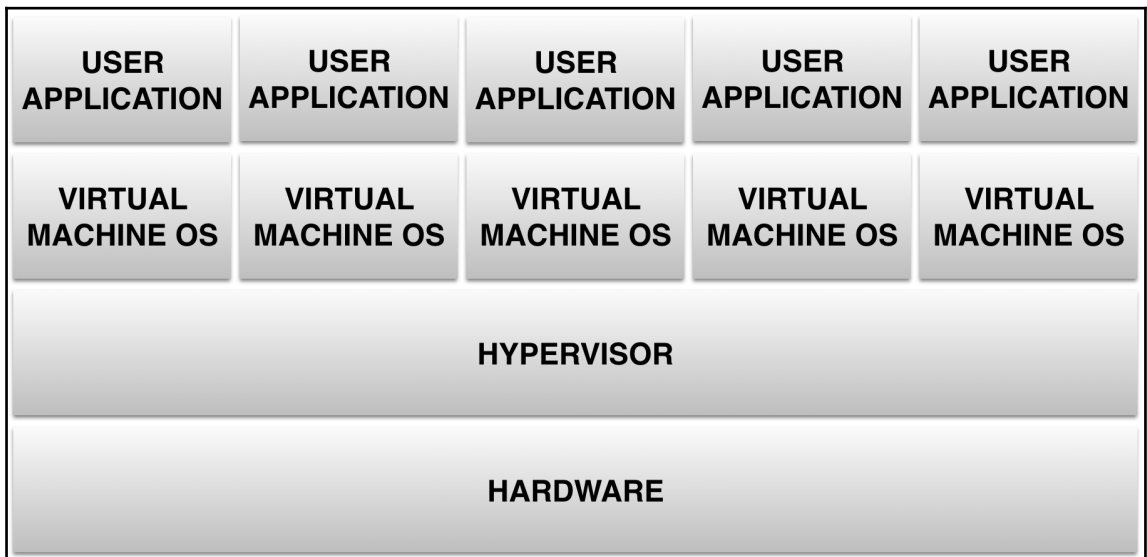
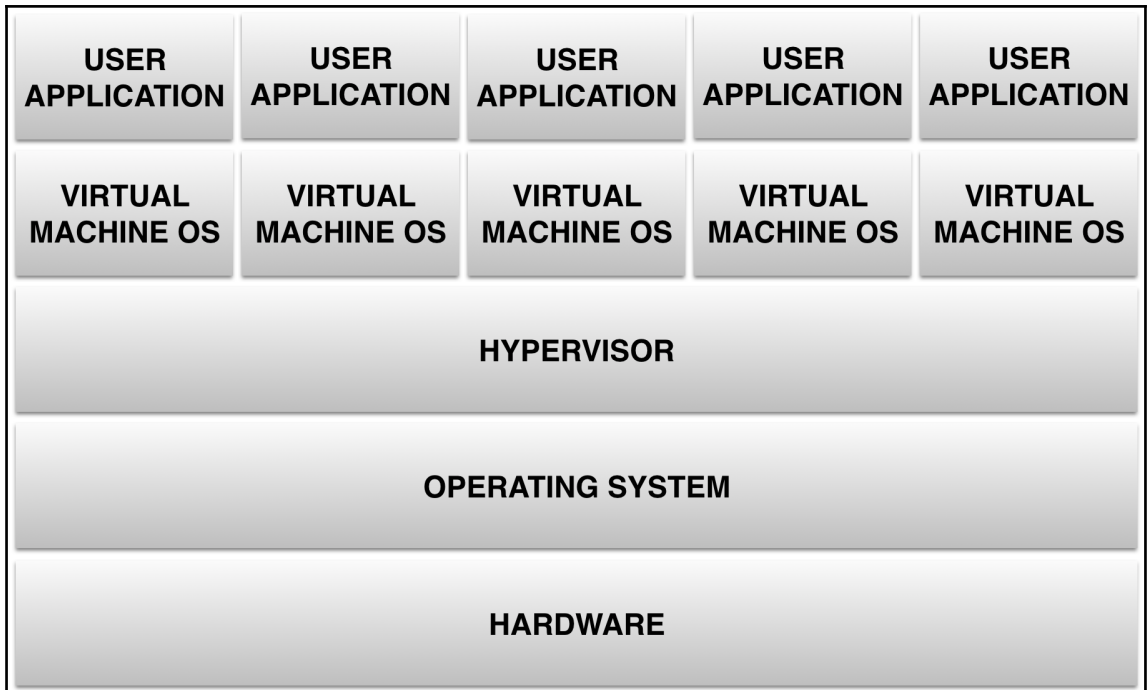
cluster, mhacc, mvacc, sdhacc, svacc

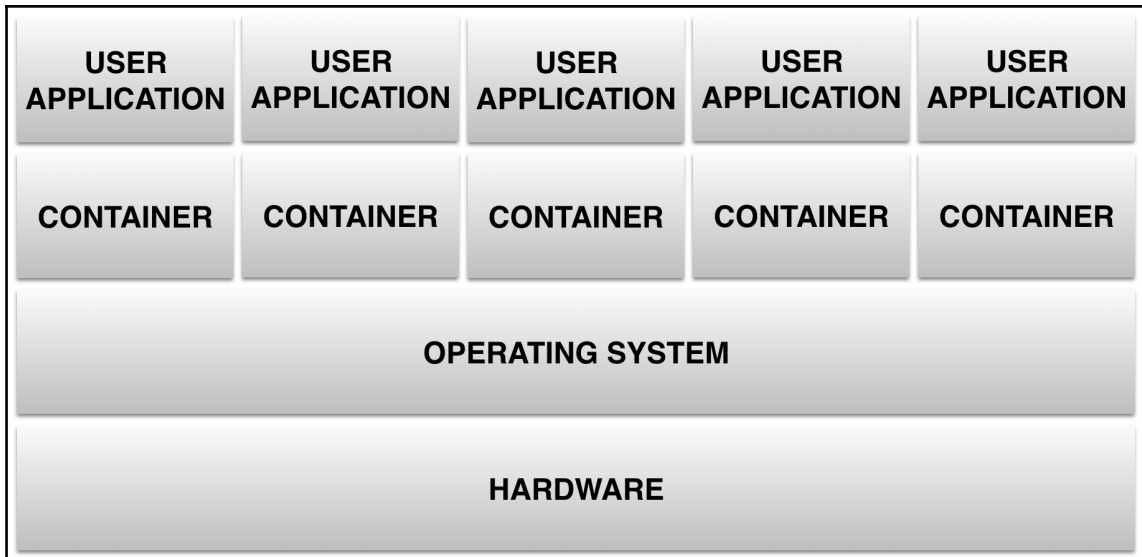


## Apache Spark on Kubernetes





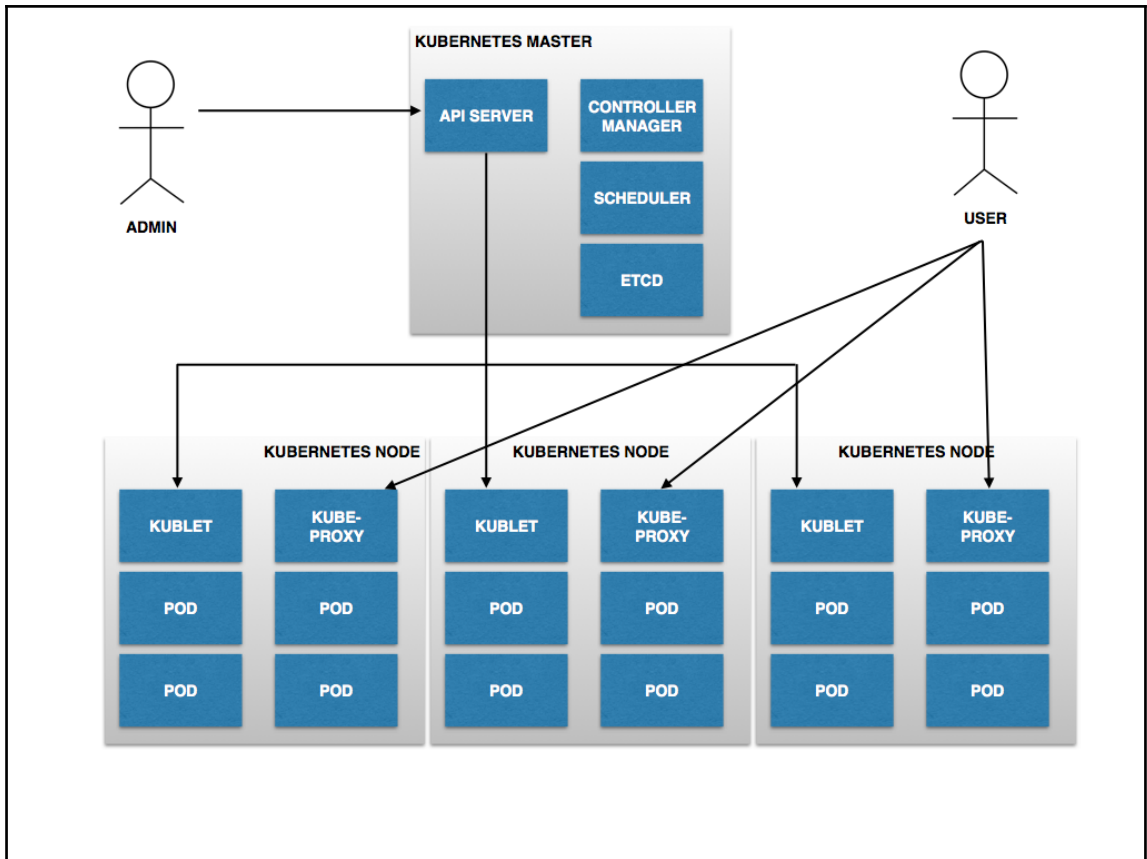




```

root@ubuntu:~# ls -al /proc/2000/ns/
total 0
dr-x--x--x 2 root      root      0 Jun 30 23:22 .
dr-xr-xr-x 9 romeokienzler romeokienzler 0 Jun 20 01:25 ..
lrwxrwxrwx 1 root      root      0 Jun 30 23:22 cgroup -> cgroup:[4026531835]
lrwxrwxrwx 1 root      root      0 Jun 30 23:22 ipc -> ipc:[4026531839]
lrwxrwxrwx 1 root      root      0 Jun 30 23:22 mnt -> mnt:[4026531840]
lrwxrwxrwx 1 root      root      0 Jun 30 23:22 net -> net:[4026531957]
lrwxrwxrwx 1 root      root      0 Jun 30 23:22 pid -> pid:[4026531836]
lrwxrwxrwx 1 root      root      0 Jun 30 23:22 user -> user:[4026531837]
lrwxrwxrwx 1 root      root      0 Jun 30 23:22 uts -> uts:[4026531838]

```



```
romeos-mbp:~ romeokienzler$ kubectl create -f https://raw.githubusercontent.com/kubernetes/kubernetes/master/examples/spark/spark-master-service.yaml
service "spark-master" created
```

```
romeos-mbp:~ romeokienzler$ kubectl get pods
NAME                                READY   STATUS              RESTARTS   AGE
spark-master-controller-ljvq1       0/1    ContainerCreating   0          6s
```

```
romeos-mbp:~ romeokienzler$ kubectl get pods
NAME                                READY   STATUS    RESTARTS   AGE
spark-master-controller-ljvq1       1/1    Running   0          17m
```

```
romeos-mbp:~ romeokienzler$ kubectl logs spark-master-controller-ljvq1
17/07/02 05:47:14 INFO Master: Registered signal handlers for [TERM, HUP, INT]
17/07/02 05:47:15 INFO SecurityManager: Changing view acls to: root
17/07/02 05:47:15 INFO SecurityManager: Changing modify acls to: root
17/07/02 05:47:15 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(root); users with modify permissions: Set(root)
17/07/02 05:47:15 INFO Slf4jLogger: Slf4jLogger started
17/07/02 05:47:15 INFO Remoting: Starting remoting
17/07/02 05:47:16 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkMaster@spark-master:7077]
17/07/02 05:47:16 INFO Utils: Successfully started service 'sparkMaster' on port 7077.
17/07/02 05:47:16 INFO Master: Starting Spark master at spark://spark-master:7077
17/07/02 05:47:16 INFO Master: Running Spark version 1.5.2
17/07/02 05:47:26 INFO Utils: Successfully started service 'MasterUI' on port 8080.
17/07/02 05:47:26 INFO MasterWebUI: Started MasterWebUI at http://172.17.0.2:8080
17/07/02 05:47:26 INFO Utils: Successfully started service on port 6066.
17/07/02 05:47:26 INFO StandaloneRestServer: Started REST server for submitting applications on port 6066
17/07/02 05:47:26 INFO Master: I have been elected leader! New state: ALIVE
```

```
romeos-mbp:~ romeokienzler$ kubectl create -f https://raw.githubusercontent.com/kubernetes/kubernetes/master/examples/spark/spark-ui-proxy-controller.yaml
replicationcontroller "spark-ui-proxy-controller" created
```

```
kind: Service
apiVersion: v1
metadata:
  name: spark-master
spec:
  ports:
    - port: 7077
      targetPort: 7077
      name: spark
    - port: 8080
      targetPort: 8080
      name: http
  selector:
    component: spark-master
```

```
romeos-mbp:~ romeokienzler$ kubectl get svc spark-ui-proxy -o wide
NAME          CLUSTER-IP   EXTERNAL-IP   PORT(S)          AGE    SELECTOR
spark-ui-proxy 10.0.0.146   <pending>     80:30621/TCP    16s    component=spark-ui-proxy
```

```
romeos-mbp:~ romeokienzler$ minikube service spark-ui-proxy --url
http://192.168.99.100:30621
```

**Spark Master at spark://spark-master:7077**

URL: spark://spark-master:7077  
 REST URL: spark://spark-master:6066 (cluster mode)  
 Alive Workers: 0  
 Cores in use: 0 Total, 0 Used  
 Memory in use: 0.0 B Total, 0.0 B Used  
 Applications: 0 Running, 0 Completed  
 Drivers: 0 Running, 0 Completed  
 Status: ALIVE

**Workers**

Worker Id	Address	State	Cores	Memory

**Running Applications**

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration

**Completed Applications**

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration

```
romeos-mbp:~ romeokienzler$ kubectl create -f https://raw.githubusercontent.com/kubernetes/kubernetes/master/examples/spark/spark-worker-controller.yaml
replicationcontroller "spark-worker-controller" created
```

```
romeos-mbp:~ romeokienzler$ kubectl get pod
NAME                                READY   STATUS    RESTARTS   AGE
spark-master-controller-ljvq1       1/1    Running   0           23h
spark-ui-proxy-controller-k3nqs     1/1    Running   23          22h
spark-worker-controller-cz8rx       1/1    Running   0           4s
spark-worker-controller-l121v       1/1    Running   0           4s
```

```
romeos-mbp:~ romeokienzler$ kubectl logs spark-master-controller-ljvq1
17/07/02 05:47:14 INFO Master: Registered signal handlers for [TERM, HUP, INT]
17/07/02 05:47:15 INFO SecurityManager: Changing view acls to: root
17/07/02 05:47:15 INFO SecurityManager: Changing modify acls to: root
17/07/02 05:47:15 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(root); users with modify permissions: Set(root)
17/07/02 05:47:15 INFO Slf4jLogger: Slf4jLogger started
17/07/02 05:47:15 INFO Remoting: Starting remoting
17/07/02 05:47:16 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkMaster@spark-master:7077]
17/07/02 05:47:16 INFO Utils: Successfully started service 'sparkMaster' on port 7077.
17/07/02 05:47:16 INFO Master: Starting Spark master at spark://spark-master:7077
17/07/02 05:47:16 INFO Master: Running Spark version 1.5.2
17/07/02 05:47:26 INFO Utils: Successfully started service 'MasterUI' on port 8080.
17/07/02 05:47:26 INFO MasterWebUI: Started MasterWebUI at http://172.17.0.2:8080
17/07/02 05:47:26 INFO Utils: Successfully started service on port 6066.
17/07/02 05:47:26 INFO Master: Running Spark version 1.5.2
17/07/02 05:47:26 INFO StandaloneRestServer: Started REST server for submitting applications on port 6066
17/07/02 05:47:26 INFO Master: I have been elected leader! New state: ALIVE
17/07/03 04:42:53 INFO Master: Registering worker 172.17.0.6:35693 with 2 cores, 1024.0 MB RAM
17/07/03 04:42:53 INFO Master: Registering worker 172.17.0.7:36563 with 2 cores, 1024.0 MB RAM
```



**Spark Master at spark://spark-master:7077**

URL: spark://spark-master:7077  
 REST URL: spark://spark-master:6066 (cluster mode)  
 Alive Workers: 2  
 Cores in use: 4 Total, 0 Used  
 Memory in use: 2.0 GB Total, 0.0 B Used  
 Applications: 0 Running, 0 Completed  
 Drivers: 0 Running, 0 Completed  
 Status: ALIVE

**Workers**

Worker Id	Address	State	Cores	Memory
worker-20170703044250-172.17.0.6-35693	172.17.0.6:35693	ALIVE	2 (0 Used)	1024.0 MB (0.0 B Used)
worker-20170703044250-172.17.0.7-36563	172.17.0.7:36563	ALIVE	2 (0 Used)	1024.0 MB (0.0 B Used)

**Running Applications**

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
No running applications.							

**Completed Applications**

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
No completed applications.							

```
romeos-mbp:~ romeokienzler$ kubectl create -f https://raw.githubusercontent.com/kubernetes/kubernetes/master/examples/spark/zeppelin-controller.yaml
replicationcontroller "zeppelin-controller" created
```

```
romeos-mbp:~ romeokienzler$ kubectl get pod
NAME                                READY    STATUS    RESTARTS   AGE
spark-master-controller-ljqv1       1/1     Running   0           23h
spark-ui-proxy-controller-k3nqs     1/1     Running   23          22h
spark-worker-controller-cz8rx       1/1     Running   0           3m
spark-worker-controller-l121v       1/1     Running   0           3m
zeppelin-controller-csmvr           0/1     ContainerCreating 0           3s
```

```
romeos-mbp:~ romeokienzler$ kubectl get pod
NAME                                READY    STATUS    RESTARTS   AGE
spark-master-controller-ljqv1       1/1     Running   0           23h
spark-ui-proxy-controller-k3nqs     1/1     Running   23          23h
spark-worker-controller-cz8rx       1/1     Running   0           33m
spark-worker-controller-l121v       1/1     Running   0           33m
zeppelin-controller-csmvr           1/1     Running   0           29m
```

```
romeos-mbp:~ romeokienzler$ kubectl create -f https://raw.githubusercontent.com/kubernetes/kubernetes/master/examples/spark/zeppelin-service.yaml
service "zeppelin" created
```

```
romeos-mbp:~ romeokienzler$ minikube service zeppelin --url
http://192.168.99.100:30510
```

The screenshot shows the Zeppelin web interface at the URL `192.168.99.100:30510/#/`. The page features a blue header with the Zeppelin logo, navigation tabs for 'Notebook' and 'Interpreter', and a search bar. The main content area is titled 'Welcome to Zeppelin!' and includes a brief description of the tool as a web-based notebook for interactive data analytics. It provides links for 'Notebook' (Import note, Create new note), 'Help' (Zeppelin documentation), and 'Community' (Mailing list, Issues tracking, GitHub). A large blue illustration of a blimp is positioned on the right side of the page.

The screenshot shows the Zeppelin web interface at the URL `192.168.99.100:30510/#/notebook/2CNQVDPTG`. The page displays a notebook titled 'Note 9U3FC9ETS'. The code input area contains the following Scala code:

```
val nums = sc.parallelize(Array(1,2,3))
nums.count
```

The output area shows the execution results:

```
nums: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[4] at parallelize at <console>:23
res2: Long = 3
```

The execution status is indicated as 'RUNNING 0%' with a progress bar. The page also includes a toolbar with various icons for code execution and management.