Strata+Hadoop WORLD

PRESENTED BY

O'REILLY cloudera

strataconf.com
#StrataHadoop

Building machine-learning apps with Spark

Spark ML and GraphX

Jayant / Vartika / Krishna

cloudera

VOLVO

Agenda

Overview	5 min (9:00-9:05)	
Lab Environment Setup	10 min (9:05-9:15)	Scala IDE for Eclipse
Spark ML	75 min (9:15-10:15)	Spam Detection, Movie Recommendations, Streaming K-Means, Churn Prediction
Break	30 min (10:30-11:00)	
Spark ML (cont…)	35 min (11:00-11:35)	
GraphX	50 min (11:35-12:25)	Overview, Exploring Structures, Community-Affiliation, Algorithms, The AlphaGo Community, Wikepedia Page Rank
Closing	5 min (12:25-12:30)	

Source Code

- https://github.com/jayantshekhar/strata-2016

Spark ML

Spark ML

Spam Detection	15 min	
Movie Lens Recommendations	15 min	
Streaming K-Means	15 min	
Churn Prediction	15 min	

- Pipeline
  - DataFrame
  - Transformer
  - Estimator
  - Pipeline

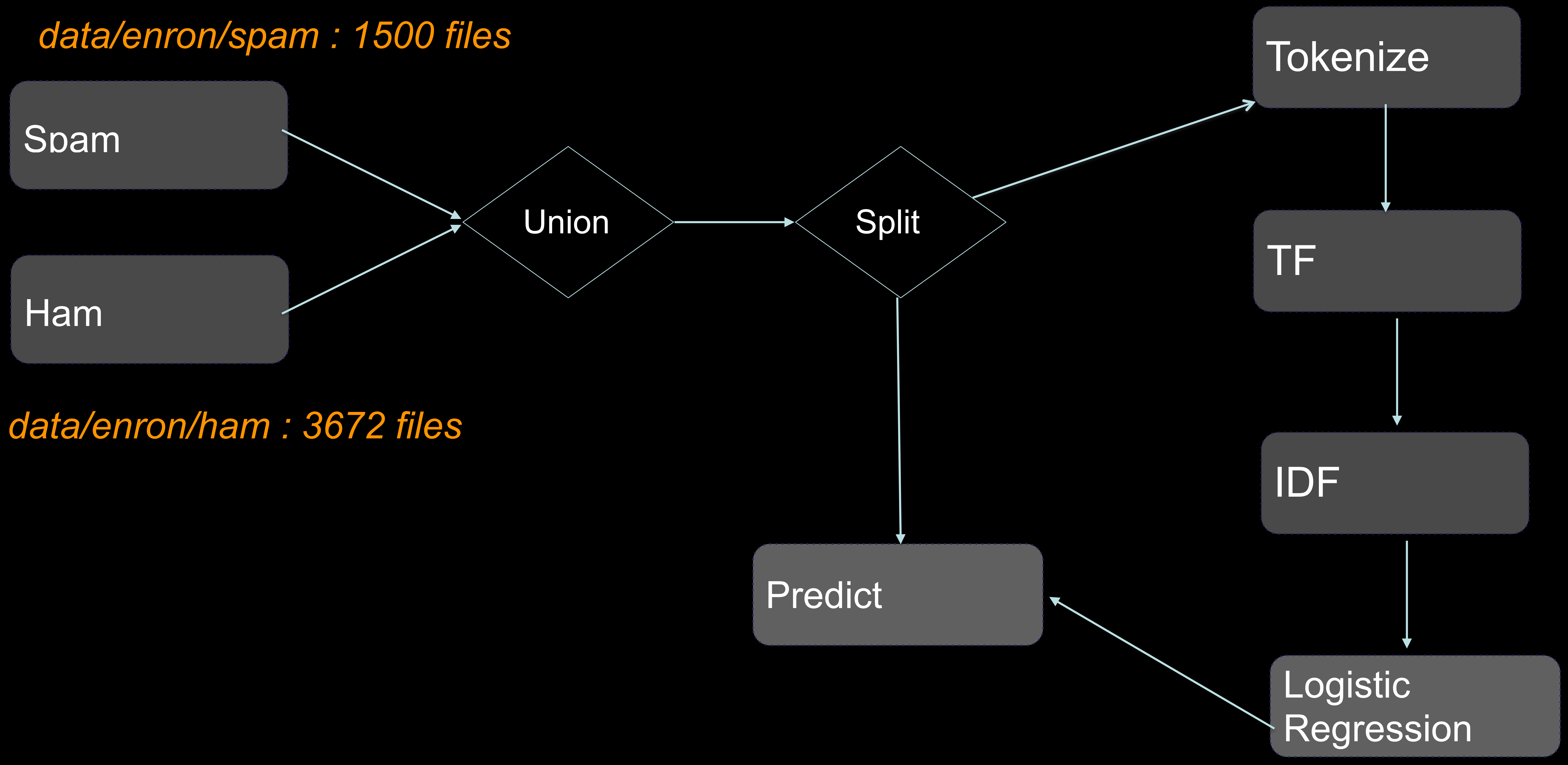
- Feature Extractors & Transformers
  - Tokenizer
  - TF/IDF
  - VectorAssembler
  - StringIndexer

- Classification & Regression
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - Gradient Boosted Tree
- Collaborative Filtering
  - ALS
- Frequent Pattern Mining
- Clustering
  - K-Means
  - LDA

Spam Detection

Logistic Regression

Spam Detection on Enron Dataset



data/enron/spam : 1500 files

data/enron/ham : 3672 files

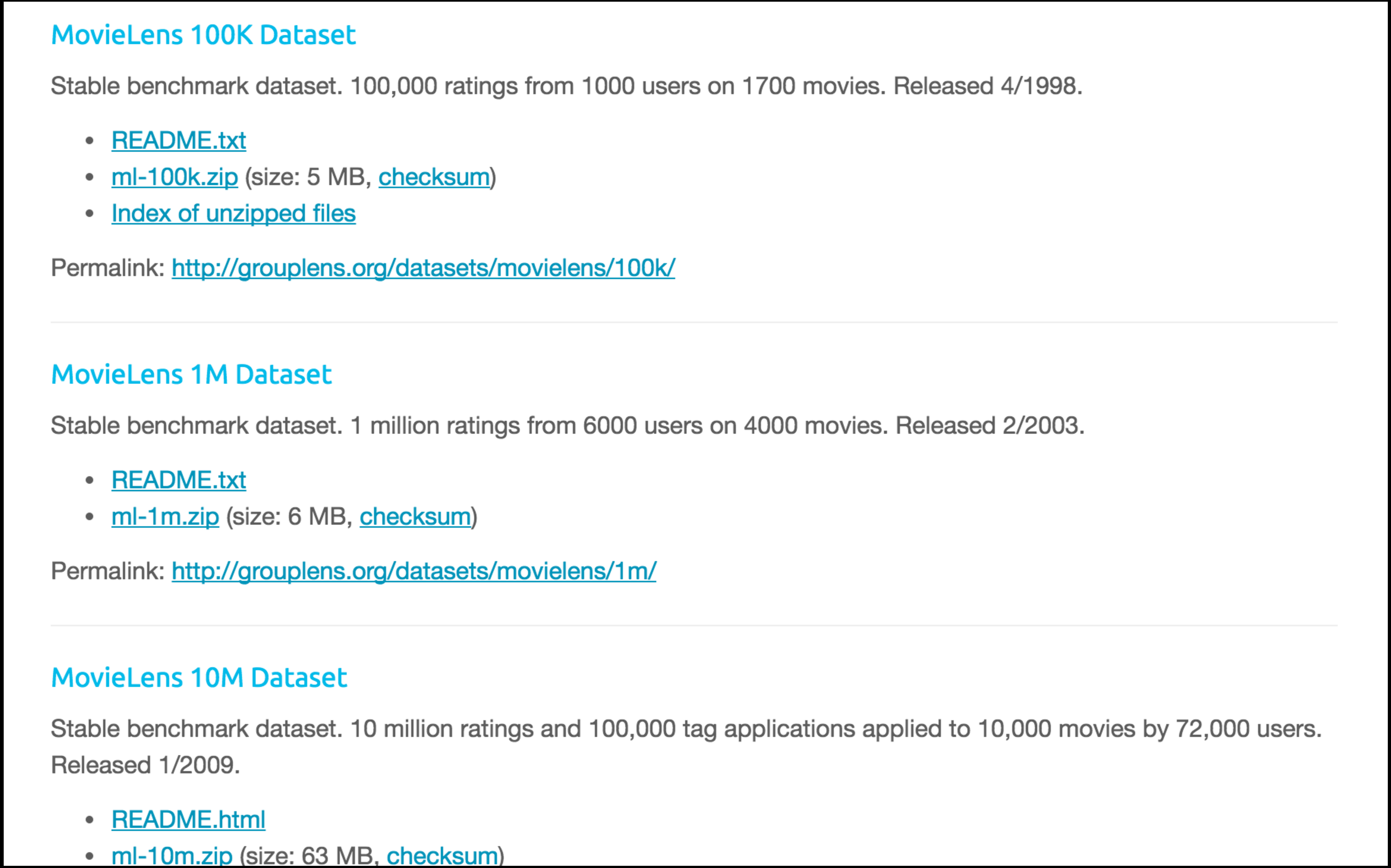
```
+--------------------+--------------------+-----+
|                file|                text|label|
+--------------------+--------------------+-----+
|file:/Users/jayan...|Subject: dobmeos ...|  1.0|
|file:/Users/jayan...|Subject: your pre...|  1.0|
|file:/Users/jayan...|Subject: get that...|  1.0|
|file:/Users/jayan...|Subject: await yo...|  1.0|
|file:/Users/jayan...|Subject: coca col...|  1.0|
```

```
+--------------------+--------------------+-----+
|                file|                text|label|
+--------------------+--------------------+-----+
|file:/Users/jayan...|Subject: christma...|  0.0|
|file:/Users/jayan...|Subject: vastar r...|  0.0|
|file:/Users/jayan...|Subject: calpine ...|  0.0|
|file:/Users/jayan...|Subject: re : iss...|  0.0|
|file:/Users/jayan...|Subject: meter 72...|  0.0|
|file:/Users/jayan...|Subject: mcmullen...|  0.0|
```

```
+--------------------+--------------------+-----+--------------------+----------+
|                file|                text|label|            features|prediction|
+--------------------+--------------------+-----+--------------------+----------+
|file:/Users/jayan...|Subject: dobmeos ...|  1.0|(262144,[0,33,37,...|       1.0|
|file:/Users/jayan...|Subject: await yo...|  1.0|(262144,[0,36,40,...|       0.0|
|file:/Users/jayan...|Subject: real pro...|  1.0|(262144,[0,33,36,...|       1.0|
|file:/Users/jayan...|Subject: re : rdd...|  1.0|(262144,[0,44,58,...|       1.0|
|file:/Users/jayan...|Subject: cut your...|  1.0|(262144,[0,37,39,...|       1.0|
+--------------------+--------------------+-----+--------------------+----------+

|file:/Users/jayan...|Subject: shut - i...|  0.0|(262144,[0,35,40,...|       0.0|
|file:/Users/jayan...|Subject: hpl nomi...|  0.0|(262144,[0,38,44,...|       0.0|
|file:/Users/jayan...|Subject: 98 - 673...|  0.0|(262144,[0,35,38,...|       0.0|
|file:/Users/jayan...|Subject: hl & p m...|  0.0|(262144,[0,33,38,...|       0.0|
|file:/Users/jayan...|Subject: purchasi...|  0.0|(262144,[0,34,39,...|       0.0|
|file:/Users/jayan...|Subject: per nels...|  0.0|(262144,[0,39,40,...|       0.0|
|file:/Users/jayan...|Subject: see atta...|  0.0|(262144,[0,44,58,...|       0.0|
|file:/Users/jayan...|Subject: monthly ...|  0.0|(262144,[0,34,36,...|       0.0|
|file:/Users/jayan...|Subject: koch mid...|  0.0|(262144,[0,44,46,...|       0.0|
|file:/Users/jayan...|Subject: nom chan...|  0.0|(262144,[0,34,40,...|       0.0|
|file:/Users/jayan...|Subject: half day...|  0.0|(262144,[0,46,47,...|       0.0|
```

Recommendations

Movie Lens Ratings



MovieLens 100K Dataset

Stable benchmark dataset. 100,000 ratings from 1000 users on 1700 movies. Released 4/1998.

- README.txt
- ml-100k.zip (size: 5 MB, checksum)
- Index of unzipped files

Permalink: http://grouplens.org/datasets/movielens/100k/

MovieLens 1M Dataset

Stable benchmark dataset. 1 million ratings from 6000 users on 4000 movies. Released 2/2003.

- README.txt
- ml-1m.zip (size: 6 MB, checksum)

Permalink: http://grouplens.org/datasets/movielens/1m/

MovieLens 10M Dataset

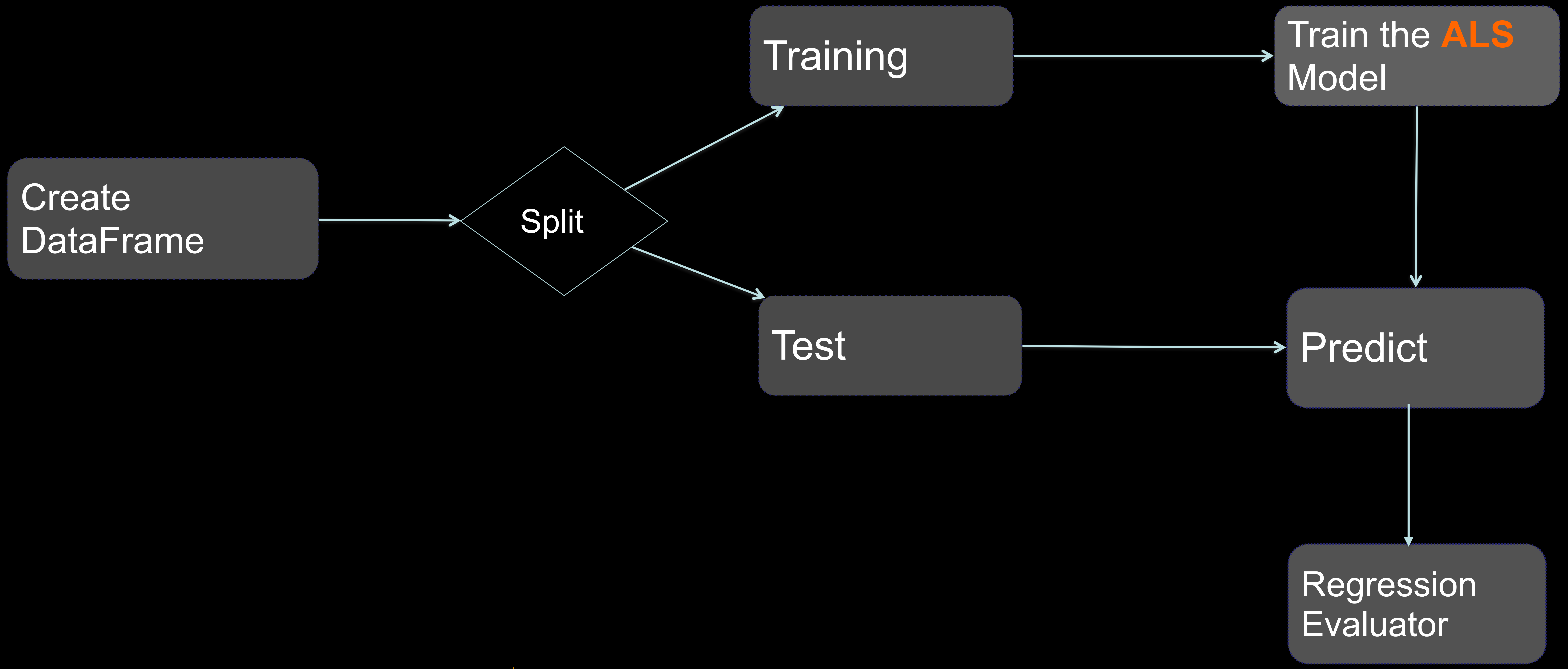
Stable benchmark dataset. 10 million ratings and 100,000 tag applications applied to 10,000 movies by 72,000 users. Released 1/2009.

- README.html
- ml-10m.zip (size: 63 MB, checksum)

MovieLens

Userid, movie id, rating

0,2,3
0,3,1
0,5,2

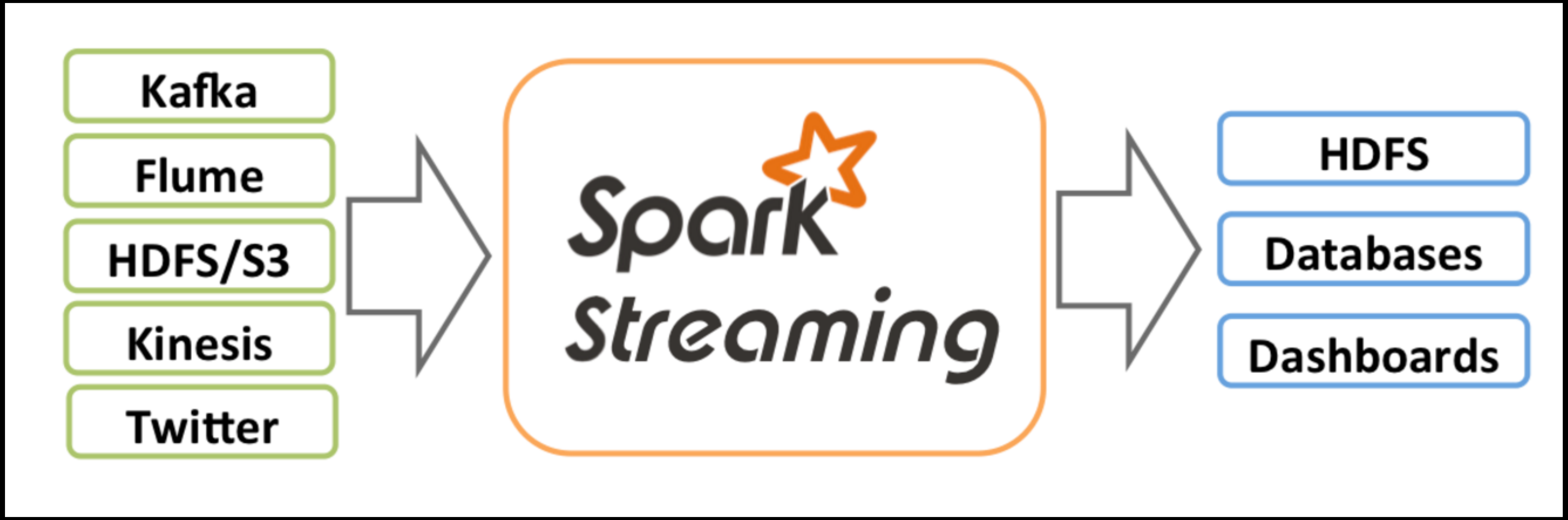


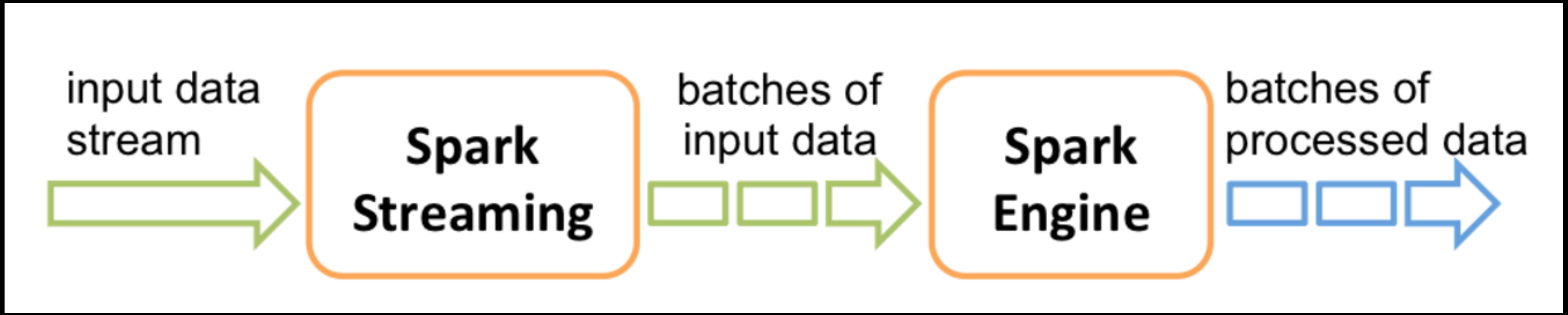
```
root
 |-- C0: string (nullable = true)
 |-- C1: string (nullable = true)
 |-- C2: string (nullable = true)

root
 |-- user: string (nullable = true)
 |-- movie: string (nullable = true)
 |-- rating: string (nullable = true)
```

```
+----+-----+------+
|user|movie|rating|
+----+-----+------+
|   1| 1193|   5.0|
|   1|  661|   3.0|
|   1|  914|   3.0|
|   1| 3408|   4.0|
|   1| 2355|   5.0|
|   1| 1197|   3.0|
|   1| 1287|   5.0|
```

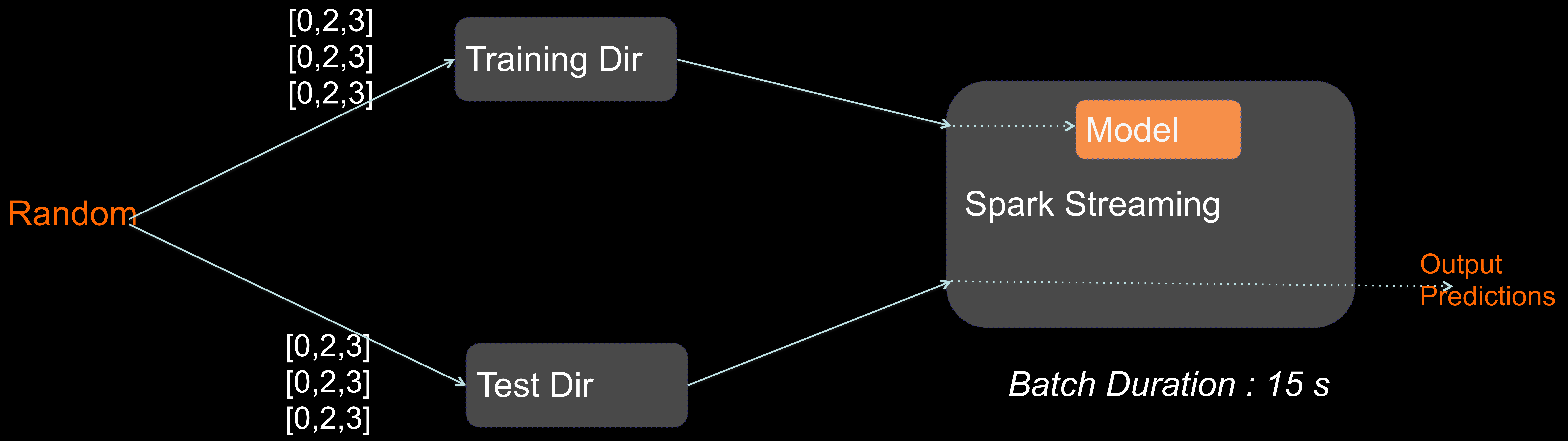
```
|user|movie|rating|prediction|
+----+-----+------+----------+
|5234|   31|   1.0| 2.1774428|
|2242|   31|   5.0| 3.1459289|
|1451|   31|   4.0| 2.3405406|
| 855|   31|   3.0| 2.2023783|
| 855|   31|   3.0| 2.2023783|
|5657|   31|   4.0| 3.7401206|
|5305|   31|   3.0| 2.1689334|
|1306|   31|   3.0| 3.2851512|
 * Alternating Least Squares (ALS) matrix factorization.
 *
 * ALS attempts to estimate the ratings matrix `R` as the product of two lower-rank
matrices,
 * `X` and `Y`, i.e. `X * Yt = R`. Typically these approximations are called 'factor'
matrices.
 * The general approach is iterative. During each iteration, one of the factor matrices is
held constant, while the other is solved for using least squares. The newly-solved factor
matrix is then held constant while solving for the other factor matrix.
```

Streaming K-Means


Kafka
Flume
HDFS/S3
Kinesis
Twitter
Spark Streaming
HDFS
Databases
Dashboards


input data stream
Spark Streaming
batches of input data
Spark Engine
batches of processed data

Streaming K-Means



Batch Duration : 15 s

Estimate clusters on one stream of data and make predictions on another stream

Streaming K-Means

- Each point should be formatted as [x1, x2, x3]

- Anytime a text file is placed in ../trainingDir the model would update

- Any time a text file is placed in ../testDir they would be processed to produce predictions using the current model

- The decay can be specified using a halfLife parameter, which determines the correct decay factor such that, for data acquired at time t, its contribution by time t + halfLife will have dropped to 0.5.

```
var trainingDir = "streamingTrainDir"
var testDir = "streamingTestDir"
var batchDuration : Long = 15 // in seconds
var numClusters = 3
```

```
Jayant:strata-2016 jayant$ ls -l streamingDataDir/
total 0
drwxr-xr-x  5 jayant  staff  170 Jun  1 00:09 1
drwxr-xr-x  5 jayant  staff  170 Jun  1 00:10 2
drwxr-xr-x  5 jayant  staff  170 Jun  1 00:10 3
drwxr-xr-x  5 jayant  staff  170 Jun  1 00:10 4
drwxr-xr-x  5 jayant  staff  170 Jun  1 00:10 5
drwxr-xr-x  5 jayant  staff  170 Jun  1 00:11 6
```

```
Jayant:strata-2016 jayant$ ls -l streamingTrainDir/
total 64
-rw-r--r--  1 jayant  staff  6109 Jun  1 00:09 1
-rw-r--r--  1 jayant  staff  6101 Jun  1 00:10 3
-rw-r--r--  1 jayant  staff  6074 Jun  1 00:10 5
-rw-r--r--  1 jayant  staff  6082 Jun  1 00:11 7
```

```
Jayant:strata-2016 jayant$ ls -l streamingTestDir/
total 64
-rw-r--r--  1 jayant  staff  8153 Jun  1 00:10 2
-rw-r--r--  1 jayant  staff  8171 Jun  1 00:10 4
-rw-r--r--  1 jayant  staff  8192 Jun  1 00:11 6
-rw-r--r--  1 jayant  staff  8142 Jun  1 00:11 8
```

Training Data

```
[-0.28875921344482436,1.3858904992406753,0.08997605487060531]
[-1.2805130758440209,0.9939584612872737,-0.47655452750026767]
[-0.010443281100194716,1.4390597064832207,0.1060992764324971]
[-0.621080758021953,-1.0856074524083963,-0.6240457792919338]
[0.8147102202208705,0.3347047775444069,0.998239073229219]
```

Test Data

```
(2.0,[1.314674536226186,-0.5939141316825893,-0.2665244694182238])
(2.0,[0.15896607505959656,-1.3248116154352312,1.7005387494315547])
(1.0,[2.092288692338904,-0.42478085016618355,1.1944557205082678])
(2.0,[0.8231075882687068,-1.7338222010770865,-2.274387117344973])
(1.0,[-0.48919972592668304,-1.3353797854575076,1.3845477789028335])
```

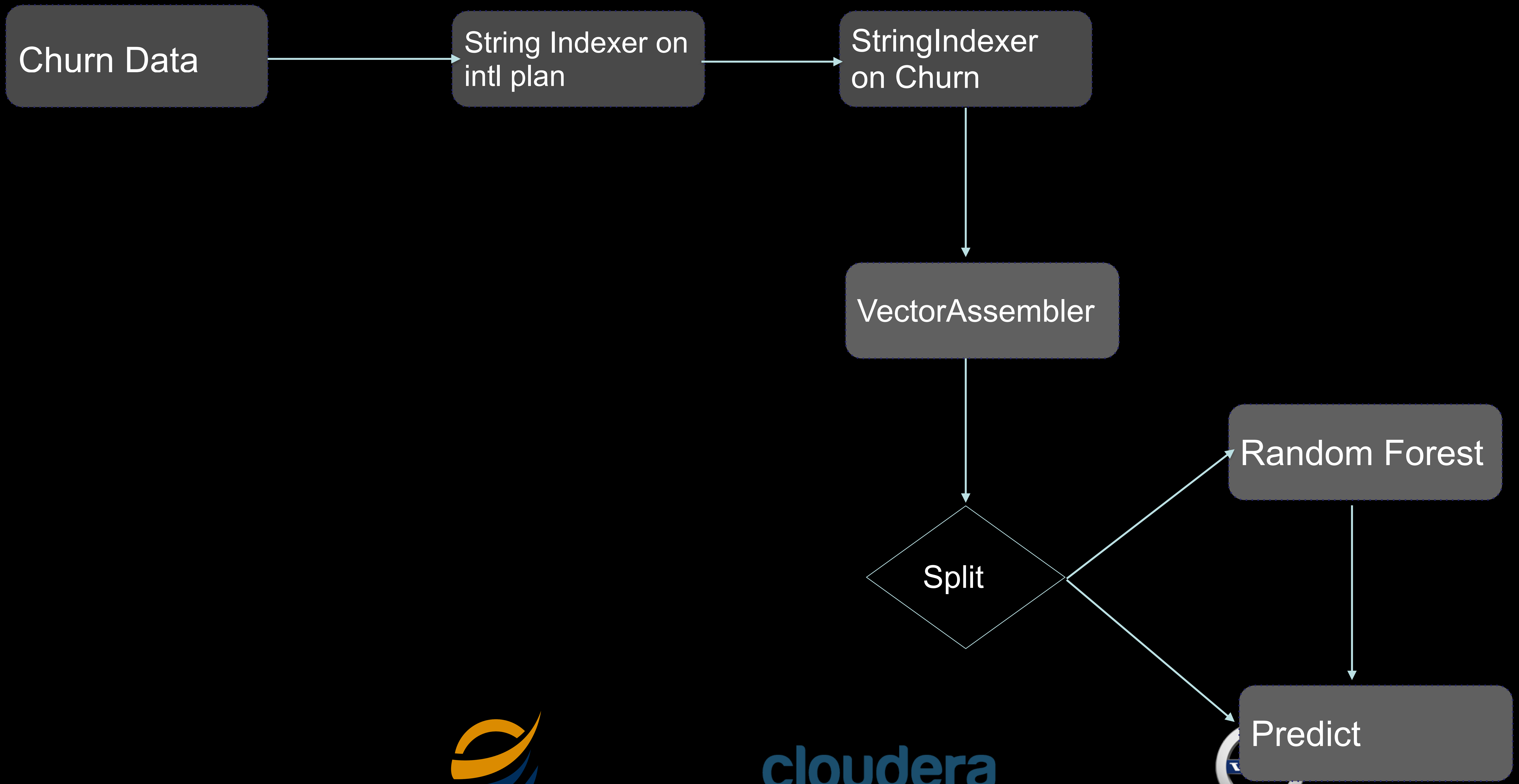
Predictions

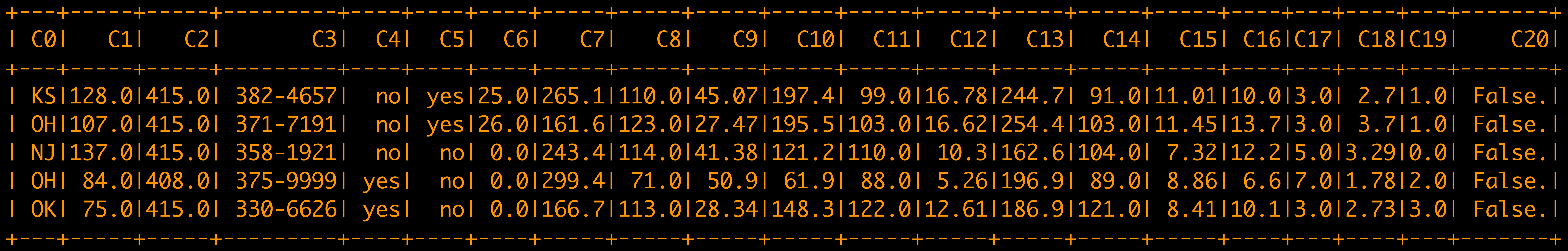
```
(2.0,1)
(2.0,0)
(1.0,1)
(2.0,1)
(1.0,0)
(2.0,2)
```

- Churn Prediction

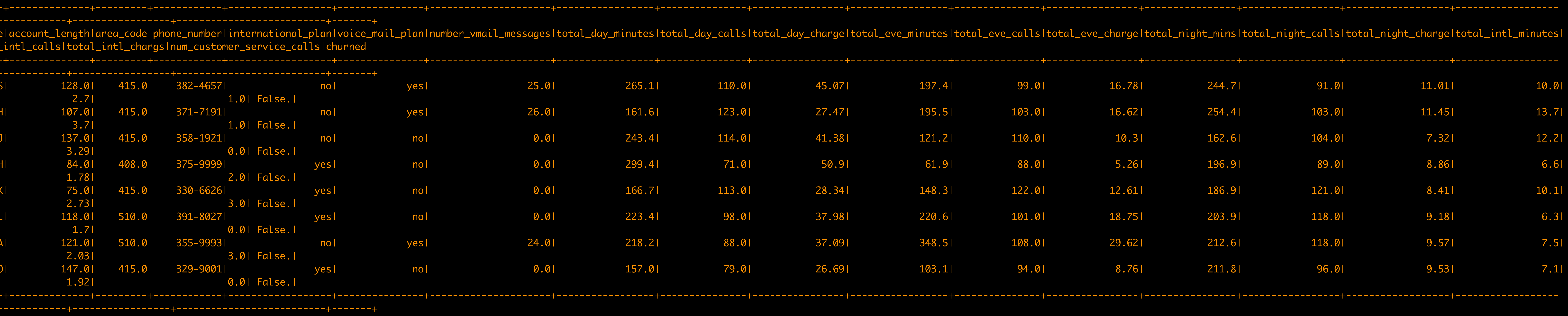
- Random Forest

Churn Prediction on Enron Dataset





```
+---+-----+-----+---------+----+----+----+-----+-----+-----+-----+-----+-----+-----+-----+-----+----+---+----+---+------+
| C0|   C1|   C2|       C3|  C4|  C5|  C6|   C7|   C8|   C9|  C10|  C11|  C12|  C13|  C14|  C15| C16|C17| C18|C19|   C20|
+---+-----+-----+---------+----+----+----+-----+-----+-----+-----+-----+-----+-----+-----+-----+----+---+----+---+------+
| KS|128.0|415.0| 382-4657|  no| yes|25.0|265.1|110.0|45.07|197.4| 99.0|16.78|244.7| 91.0|11.01|10.0|3.0| 2.7|1.0|False.|
| OH|107.0|415.0| 371-7191|  no| yes|26.0|161.6|123.0|27.47|195.5|103.0|16.62|254.4|103.0|11.45|13.7|3.0| 3.7|1.0|False.|
| NJ|137.0|415.0| 358-1921|  no|  no| 0.0|243.4|114.0|41.38|121.2|110.0| 10.3|162.6|104.0| 7.32|12.2|5.0|3.29|0.0|False.|
| OH| 84.0|408.0| 375-9999| yes|  no| 0.0|299.4| 71.0| 50.9| 61.9| 88.0| 5.26|196.9| 89.0| 8.86| 6.6|7.0|1.78|2.0|False.|
| OK| 75.0|415.0| 330-6626| yes|  no| 0.0|166.7|113.0|28.34|148.3|122.0|12.61|186.9|121.0| 8.41|10.1|3.0|2.73|3.0|False.|
+---+-----+-----+---------+----+----+----+-----+-----+-----+-----+-----+-----+-----+-----+-----+----+---+----+---+------+
```



```
-+--------------+---------+------------+------------------+---------------+---------------------+-----------------+---------------+----------------+-----------------+---------------+----------------+----------------+-----------------+------------------+------------------
-----------+-----------------+--------------------------+-------+
e|account_length|area_code|phone_number|international_plan|voice_mail_plan|number_vmail_messages|total_day_minutes|total_day_calls|total_day_charge|total_eve_minutes|total_eve_calls|total_eve_charge|total_night_mins|total_night_calls|total_night_charge|total_intl_minutes|
_intl_calls|total_intl_chargs|num_customer_service_calls|churned|
-+--------------+---------+------------+------------------+---------------+---------------------+-----------------+---------------+----------------+-----------------+---------------+----------------+----------------+-----------------+------------------+------------------
-----------+-----------------+--------------------------+-------+
S|         128.0|    415.0|    382-4657|                no|            yes|                 25.0|            265.1|          110.0|           45.07|            197.4|           99.0|           16.78|           244.7|             91.0|             11.01|              10.0|
            2.7|                    1.0| False.|
H|         107.0|    415.0|    371-7191|                no|            yes|                 26.0|            161.6|          123.0|           27.47|            195.5|          103.0|           16.62|           254.4|            103.0|             11.45|              13.7|
            3.7|                    1.0| False.|
J|         137.0|    415.0|    358-1921|                no|             no|                  0.0|            243.4|          114.0|           41.38|            121.2|          110.0|            10.3|           162.6|            104.0|              7.32|              12.2|
           3.29|                    0.0| False.|
H|          84.0|    408.0|    375-9999|               yes|             no|                  0.0|            299.4|           71.0|            50.9|             61.9|           88.0|            5.26|           196.9|             89.0|              8.86|               6.6|
           1.78|                    2.0| False.|
K|          75.0|    415.0|    330-6626|               yes|             no|                  0.0|            166.7|          113.0|           28.34|            148.3|          122.0|           12.61|           186.9|            121.0|              8.41|              10.1|
           2.73|                    3.0| False.|
L|         118.0|    510.0|    391-8027|               yes|             no|                  0.0|            223.4|           98.0|           37.98|            220.6|          101.0|           18.75|           203.9|            118.0|              9.18|               6.3|
            1.7|                    0.0| False.|
A|         121.0|    510.0|    355-9993|                no|            yes|                 24.0|            218.2|           88.0|           37.09|            348.5|          108.0|           29.62|           212.6|            118.0|              9.57|               7.5|
           2.03|                    3.0| False.|
D|         147.0|    415.0|    329-9001|               yes|             no|                  0.0|            157.0|           79.0|           26.69|            103.1|           94.0|            8.76|           211.8|             96.0|              9.53|               7.1|
           1.92|                    0.0| False.|
-+--------------+---------+------------+------------------+---------------+---------------------+-----------------+---------------+----------------+-----------------+---------------+----------------+----------------+-----------------+------------------+------------------
-----------+-----------------+--------------------------+-------+
```

Original Schema…

```
root
 |-- state: string (nullable = true)
 |-- account_length: double (nullable = true)
 |-- area_code: double (nullable = true)
 |-- phone_number: string (nullable = true)
 |-- international_plan: string (nullable = true)
 |-- voice_mail_plan: string (nullable = true)
 |-- number_vmail_messages: double (nullable = true)
 |-- total_day_minutes: double (nullable = true)
 |-- total_day_calls: double (nullable = true)
 |-- total_day_charge: double (nullable = true)
 |-- total_eve_minutes: double (nullable = true)
 |-- total_eve_calls: double (nullable = true)
 |-- total_eve_charge: double (nullable = true)
 |-- total_night_mins: double (nullable = true)
 |-- total_night_calls: double (nullable = true)
 |-- total_night_charge: double (nullable = true)
 |-- total_intl_minutes: double (nullable = true)
 |-- total_intl_calls: double (nullable = true)
 |-- total_intl_chargs: double (nullable = true)
 |-- num_customer_service_calls: double (nullable = true)
 |-- churned: string (nullable = true)
```

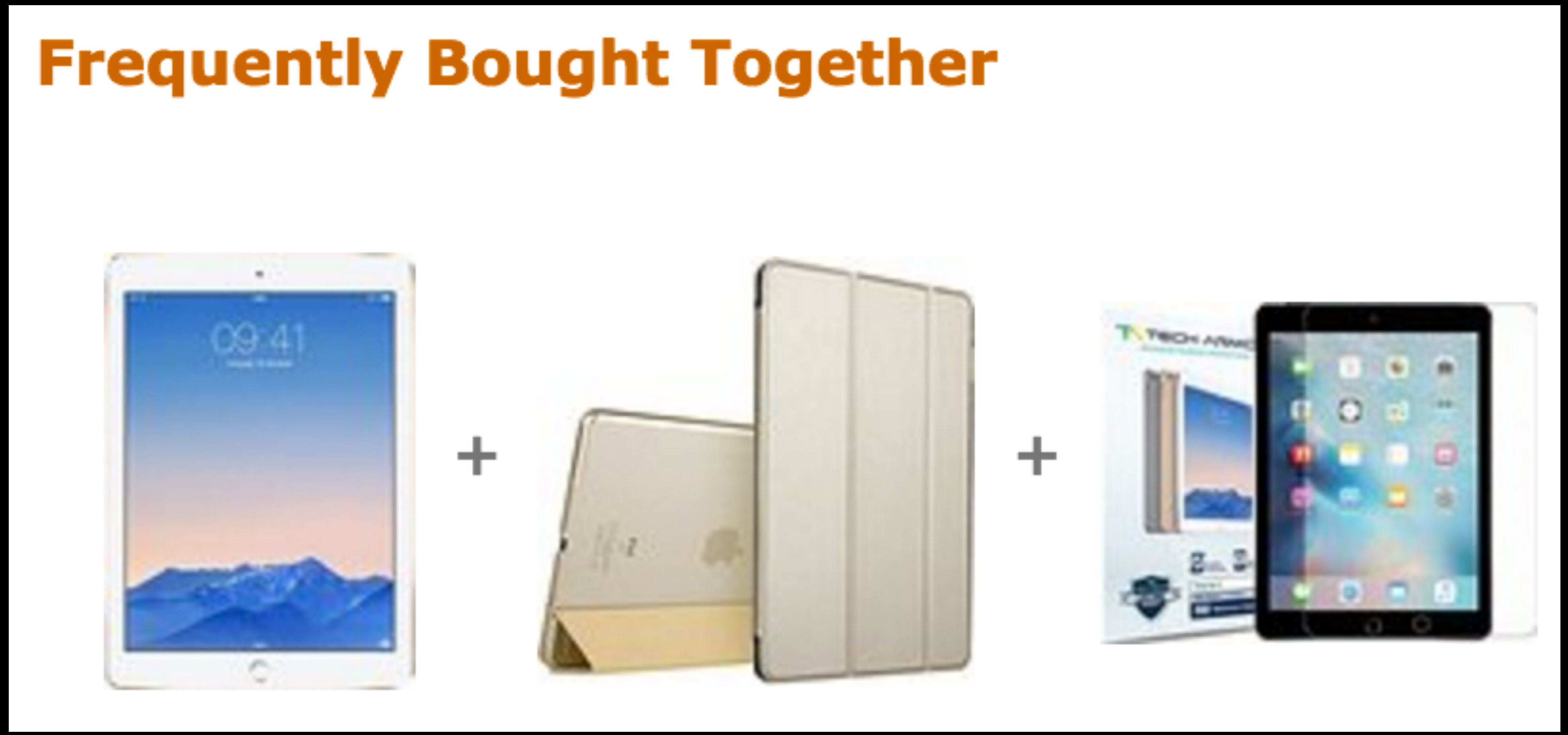
Schema after assembler…

```
root
 |-- state: string (nullable = true)
 |-- account_length: double (nullable = true)
 |-- area_code: double (nullable = true)
 |-- phone_number: string (nullable = true)
 |-- international_plan: string (nullable = true)
 |-- voice_mail_plan: string (nullable = true)
 |-- number_vmail_messages: double (nullable = true)
 |-- total_day_minutes: double (nullable = true)
 |-- total_day_calls: double (nullable = true)
 |-- total_day_charge: double (nullable = true)
 |-- total_eve_minutes: double (nullable = true)
 |-- total_eve_calls: double (nullable = true)
 |-- total_eve_charge: double (nullable = true)
 |-- total_night_mins: double (nullable = true)
 |-- total_night_calls: double (nullable = true)
 |-- total_night_charge: double (nullable = true)
 |-- total_intl_minutes: double (nullable = true)
 |-- total_intl_calls: double (nullable = true)
 |-- total_intl_chargs: double (nullable = true)
 |-- num_customer_service_calls: double (nullable = true)
 |-- churned: string (nullable = true)
 |-- label: double (nullable = true)
 |-- international_plan_indx: double (nullable = true)
 |-- features: vector (nullable = true)
```

```
areaUnderROC = 0.6232287449392713
Learned classification forest model:
RandomForestClassificationModel (uid=rfc_3f29d7cd01e1) with 10 trees
  Tree 0 (weight 1.0):
    If (feature 6 <= 133.4)
     If (feature 9 <= 3.0)
      If (feature 0 <= 76.0)
       Predict: 0.0
      Else (feature 0 > 76.0)
       If (feature 5 <= 29.84)
        If (feature 4 <= 64.0)
         Predict: 1.0
        Else (feature 4 > 64.0)
         Predict: 0.0
       Else (feature 5 > 29.84)
        Predict: 1.0
     Else (feature 9 > 3.0)
      If (feature 2 <= 0.0)
       Predict: 0.0
      Else (feature 2 > 0.0)
       Predict: 1.0
    Else (feature 6 > 133.4)
     If (feature 3 <= 272.6)
      If (feature 5 <= 12.67)
       If (feature 0 <= 64.0)
        Predict: 0.0
       Else (feature 0 > 64.0)
        If (feature 6 <= 180.6)
         Predict: 1.0
        Else (feature 6 > 180.6)
```

```
+----------+-----+--------------------+
|prediction|label|            features|
+----------+-----+--------------------+
|       0.0|  0.0|[93.0,0.0,0.0,271...|
|       0.0|  0.0|[95.0,0.0,0.0,238...|
|       0.0|  0.0|[75.0,0.0,0.0,166...|
|       0.0|  0.0|[116.0,0.0,34.0,2...|
|       1.0|  1.0|[151.0,1.0,0.0,21...|
|       0.0|  0.0|[68.0,0.0,0.0,237...|
|       0.0|  0.0|[107.0,0.0,0.0,13...|
|       0.0|  0.0|[141.0,0.0,32.0,1...|
|       0.0|  1.0|[159.0,1.0,0.0,25...|
```

discover actual shopping behavior



Frequently Bought Together

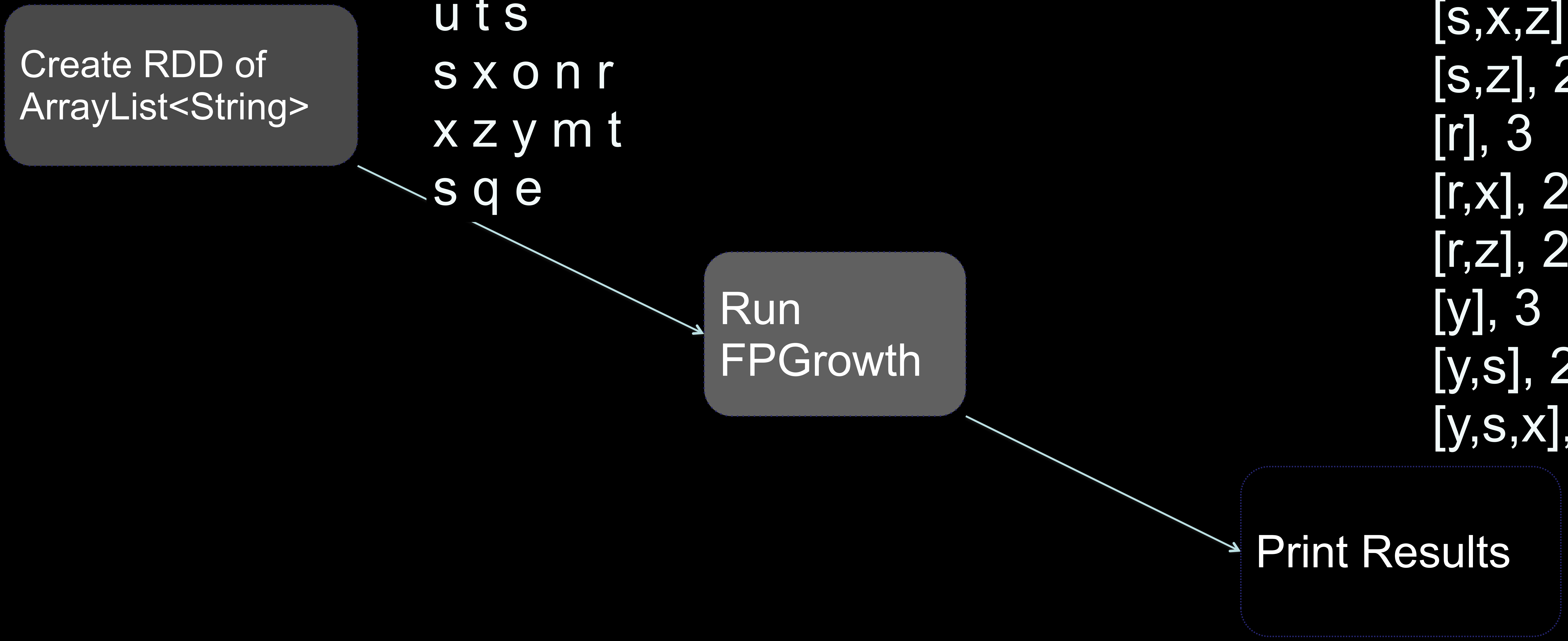
- Frequent Pattern Mining
- FPG

Frequent Pattern Mining

- Mllib has parallel implementation of FP-Growth
  - minSupport: the minimum support for an itemset to be identified as frequent. For example, if an item appears 3 out of 5 transactions, it has a support of 3/5=0.6.
  - numPartitions: the number of partitions used to distribute the work.

FPGrowth

r z h k p
z y x w v
u t s
s x o n r
x z y m t
s q e

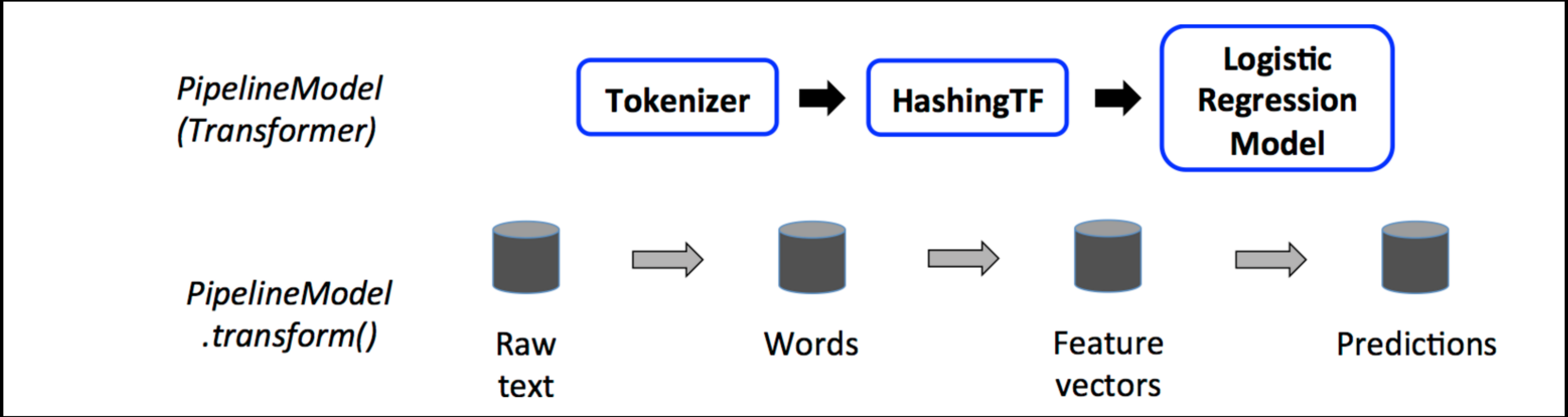


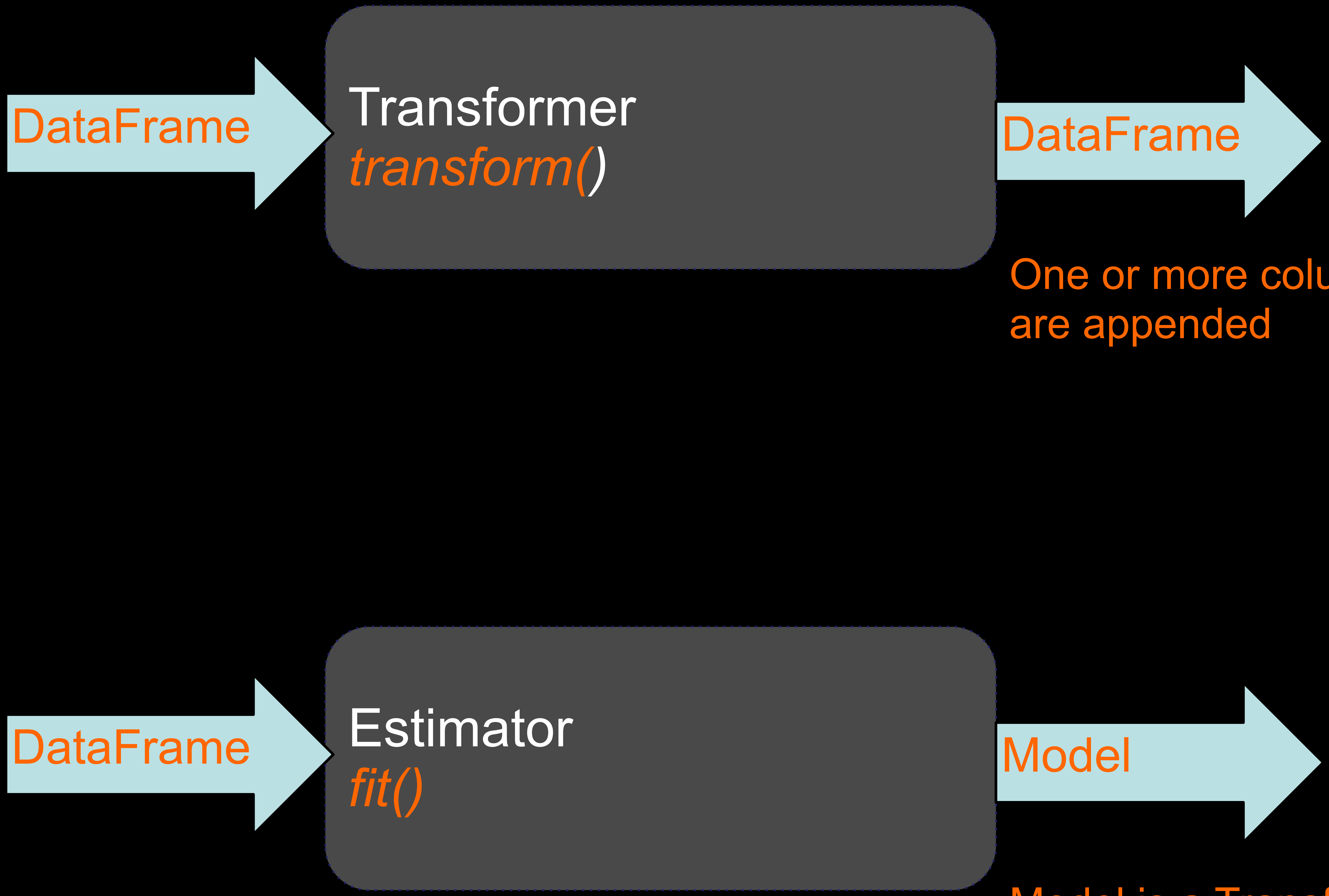
[s], 3
[s,x], 3
[s,x,z], 2
[s,z], 2
[r], 3
[r,x], 2
[r,z], 2
[y], 3
[y,s], 2
[y,s,x], 2

ML Pipelines

Spark ML

- DataFrames
- Transformer
- Estimator
- Pipeline


PipelineModel (Transformer)
Tokenizer
HashingTF
Logistic Regression Model
PipelineModel .transform()
Raw text
Words
Feature vectors
Predictions



DataFrame

Transformer

transform()

DataFrame

One or more columns
are appended

DataFrame

Estimator

fit()

Model

Model is a Transformer

Titanic Survival Prediction

Random Forest

Titanic

Remove the header row when reading

- **Data**

```
PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,,S
```

- Target Variable
  - Survived
- Predictor Variables
  - Pclass, Sex, Age, Fare

Titanic DataSet

VARIABLE DESCRIPTIONS:

survival	Survival (0 = No; 1 = Yes)
pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin
embarked	Port of Embarkation

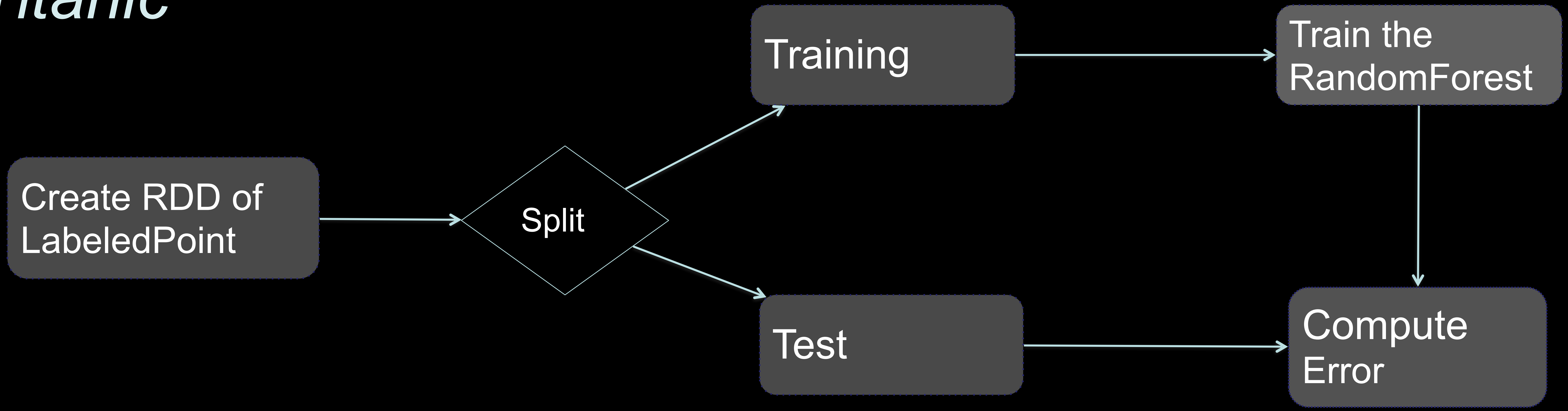
(C = Cherbourg; Q = Queenstown; S = Southampton)

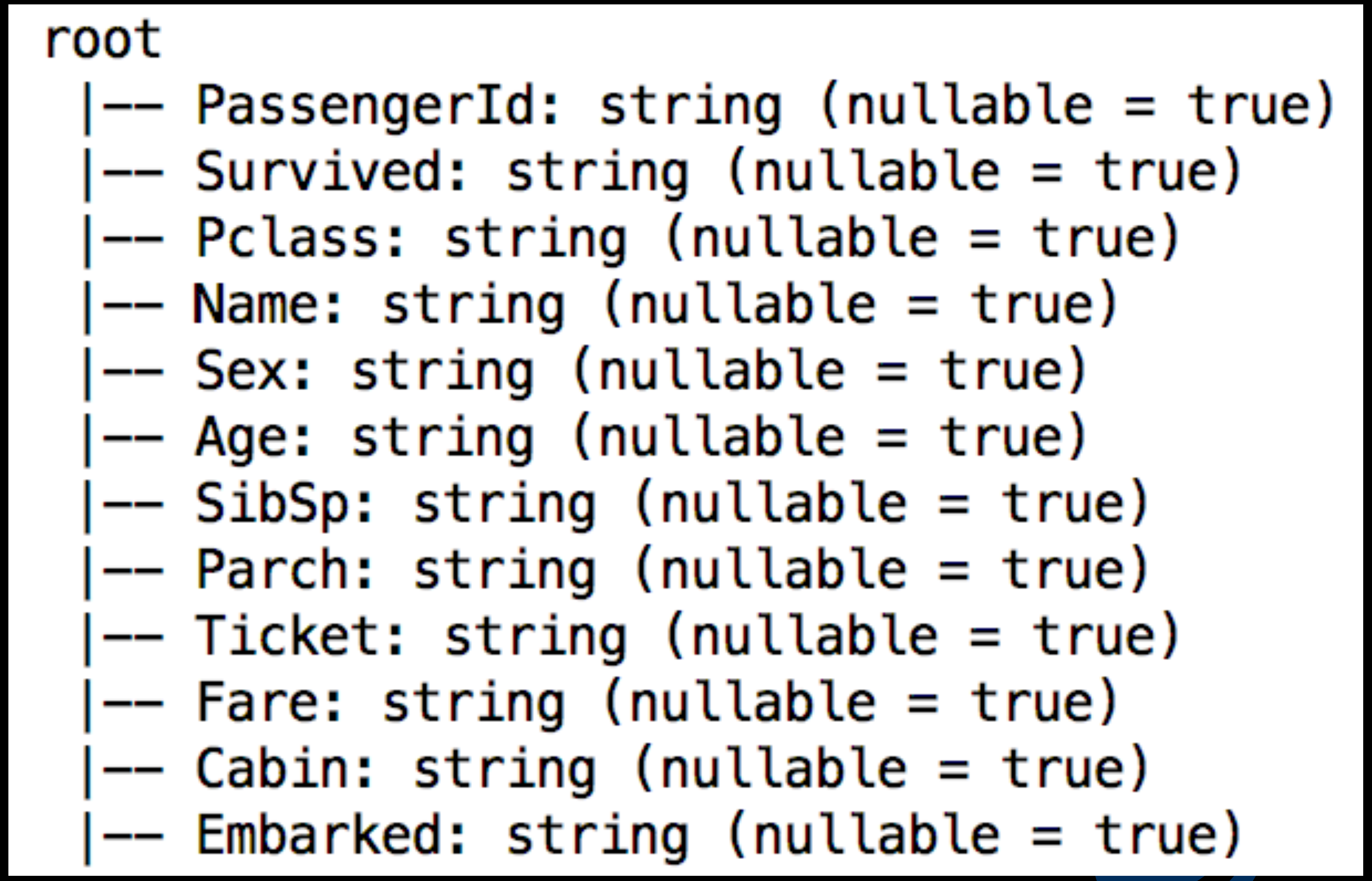
NOTES:

Pclass is a proxy for socio-economic status (SES)
1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)
If the Age is Estimated, it is in the form xx.5

Titanic





```
root
 |-- PassengerId: string (nullable = true)
 |-- Survived: string (nullable = true)
 |-- Pclass: string (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: string (nullable = true)
 |-- SibSp: string (nullable = true)
 |-- Parch: string (nullable = true)
 |-- Ticket: string (nullable = true)
 |-- Fare: string (nullable = true)
 |-- Cabin: string (nullable = true)
 |-- Embarked: string (nullable = true)
```

Random Forest

- **numTrees**: Number of trees in the forest.
- **maxDepth**: Maximum depth of each tree in the forest.
- **categoricalFeaturesInfo**: Specifies which features are categorical and how many categorical values each of those features can take. This is given as a map from feature indices to feature arity (number of categories). Any features not in this map are treated as continuous.
  - E.g., Map(0 -> 2, 4 -> 10) specifies that feature 0 is binary (taking values 0 or 1) and that feature 4 has 10 categories (values {0, 1, ..., 9}). Feature indices are 0-based: features 0 and 4 are the 1st and 5th elements of an instance's feature vector.

```
Tree 0:
 If (feature 0 in {0.0})
  If (feature 4 <= 8.7125)
   If (feature 3 <= 0.0)
    If (feature 2 <= 0.0)
     Predict: 0.0
    Else (feature 2 > 0.0)
     Predict: 0.0
   Else (feature 3 > 0.0)
    If (feature 1 <= 0.42)
     Predict: 1.0
    Else (feature 1 > 0.42)
     Predict: 0.0
  Else (feature 4 > 8.7125)
   If (feature 1 <= 14.0)
    If (feature 2 <= 2.0)
     Predict: 1.0
    Else (feature 2 > 2.0)
     Predict: 0.0
   Else (feature 1 > 14.0)
```

```
Tree 1:
 If (feature 0 in {0.0})
  If (feature 4 <= 9.8375)
   If (feature 4 <= 7.8958)
    If (feature 4 <= 7.8292)
     Predict: 0.0
    Else (feature 4 > 7.8292)
     Predict: 0.0
   Else (feature 4 > 7.8958)
    If (feature 2 <= 0.0)
     Predict: 0.0
    Else (feature 2 > 0.0)
     Predict: 1.0
  Else (feature 4 > 9.8375)
   If (feature 3 <= 0.0)
    If (feature 4 <= 26.0)
     Predict: 0.0
    Else (feature 4 > 26.0)
     Predict: 0.0
   Else (feature 3 > 0.0)
```