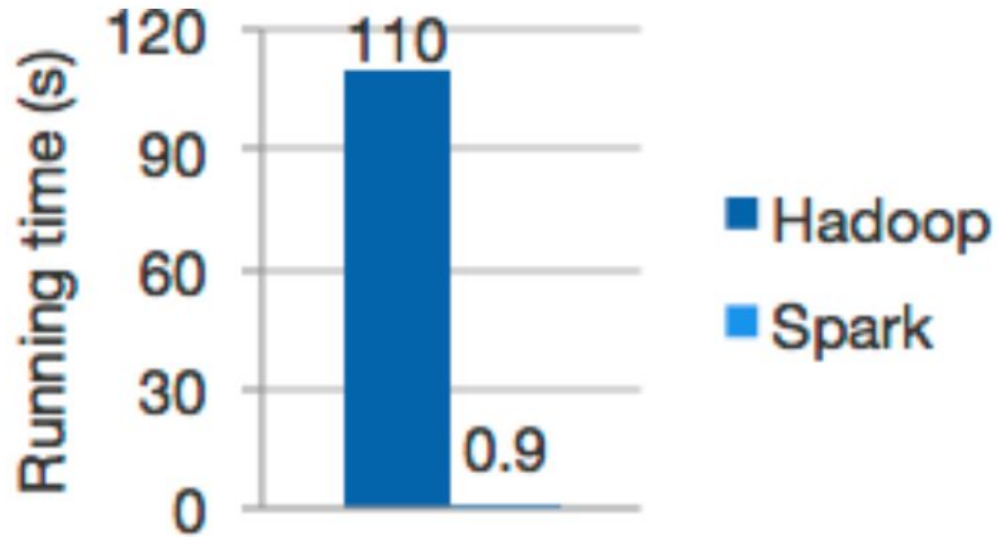


SparkML basics

Dmitry Bugaychenko

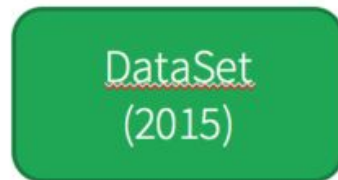
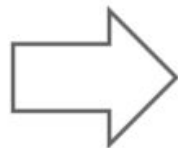
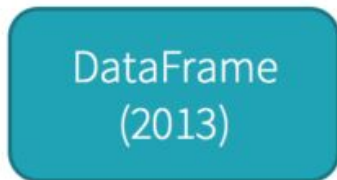
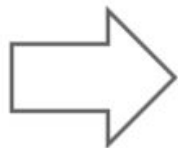




Logistic regression in Hadoop and Spark



History of Spark APIs



Distribute collection
of JVM objects

Distribute collection
of Row objects

Internally rows, externally
JVM objects

Functional Operators (map,
filter, etc.)

Expression-based operations
and UDFs

Almost the “Best of both
worlds”: **type safe + fast**

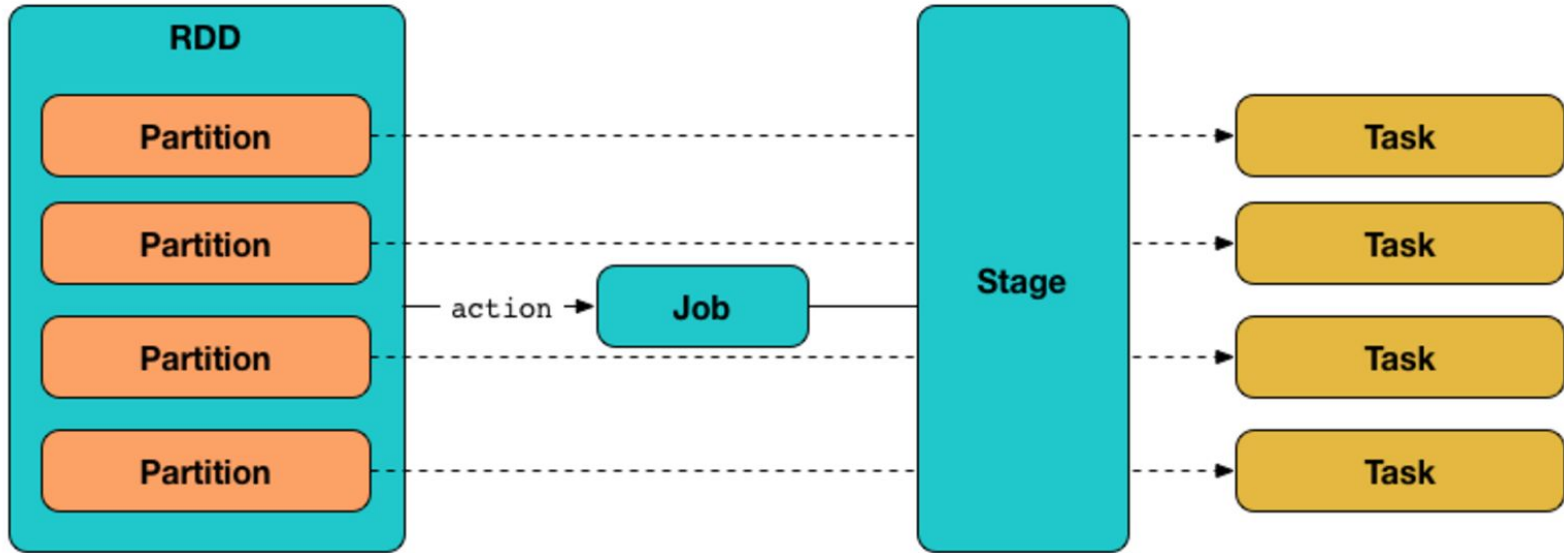
Logical plans and optimizer

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

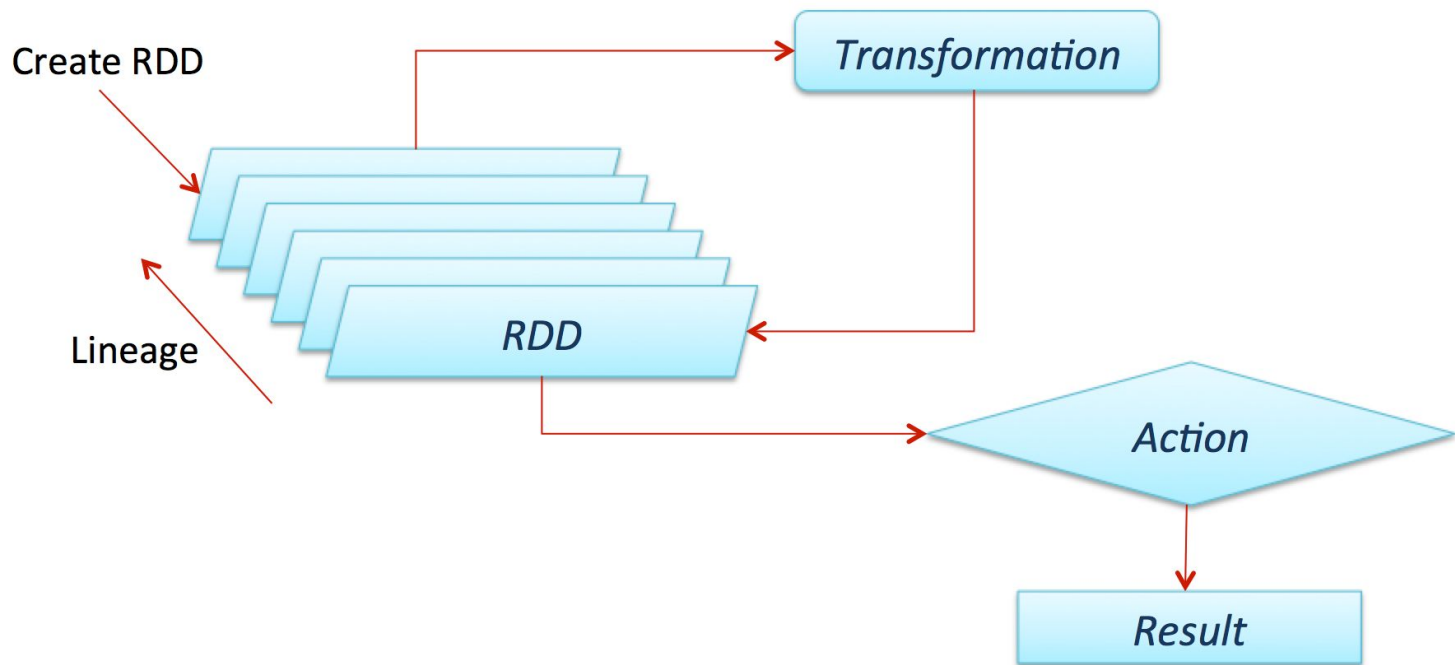
Fast/efficient internal
representations



RDD Basics



RDD Basics

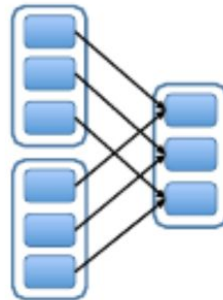
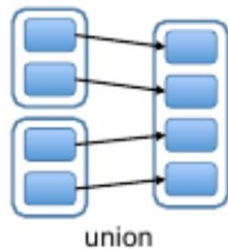
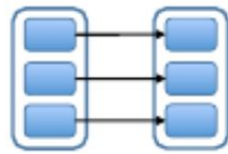


RDD Basics

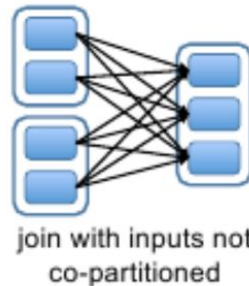
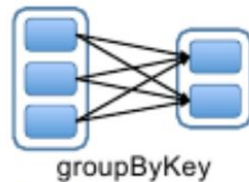
Transformations
<code>map(func)</code>
<code>flatMap(func)</code>
<code>filter(func)</code>
<code>groupByKey()</code>
<code>reduceByKey(func)</code>
<code>mapValues(func)</code>
...

Actions
<code>take(N)</code>
<code>count()</code>
<code>collect()</code>
<code>reduce(func)</code>
<code>takeOrdered(N)</code>
<code>top(N)</code>
...

Narrow Dependencies:



Wide Dependencies:



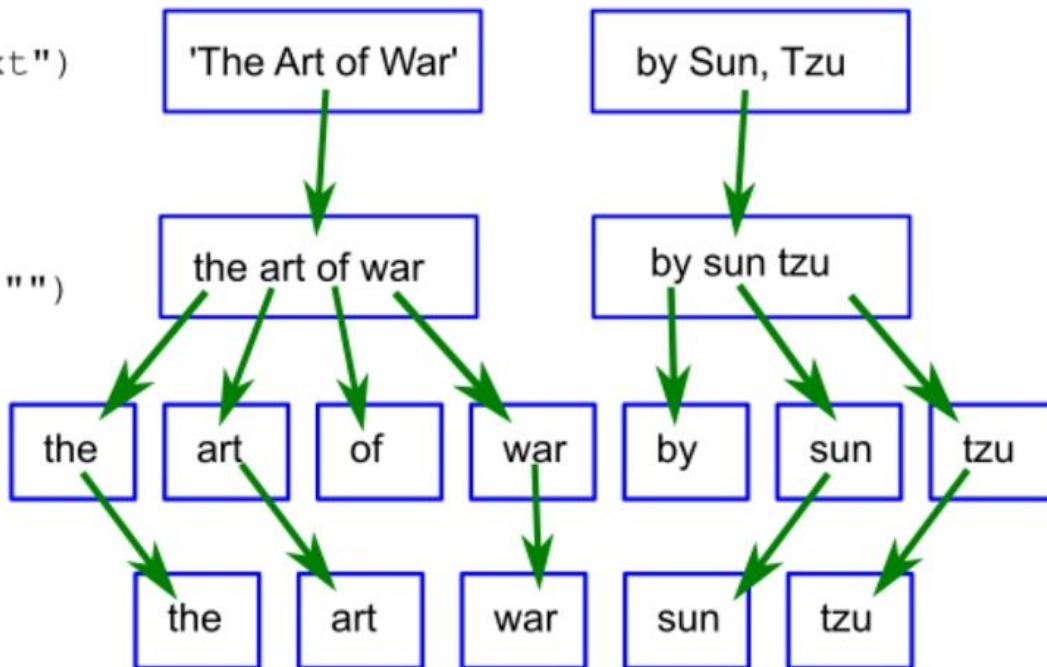
RDD Basics

```
sc.textFile("artofwar.txt")
```

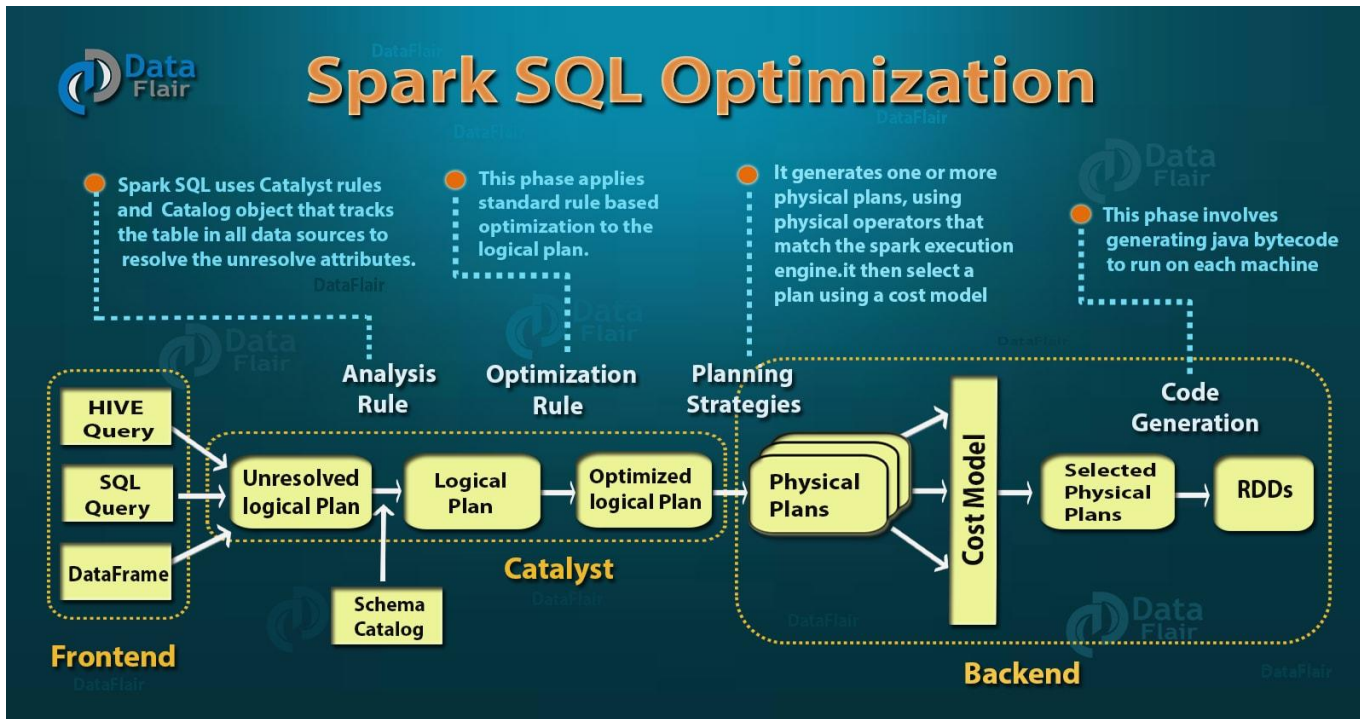
```
.map(  
  _.toLowerCase  
  .replaceAll("[^\\w ]", "")
```

```
.flatMap(_.split(" "))
```

```
.filter(_.size > 2)
```



DataFrames



Datasets

- DataFrame (with **relational operations**) and Dataset (with **lambda functions**) use Catalyst and row-oriented data representation on off-heap

```
case class Pt(x: Int, y: Int)
d = Array(Pt(1, 4), Pt(2, 5))
```

DataFrame (v1.3-)

```
df = d.toDF(...)
df.filter("x>1")
  .count()
```

Dataset (v1.6-)

```
ds = d.toDS()
ds.filter(p => p.x>1)
  .count()
```

RDD (v0.5-)

```
rdd = sc.parallelize(d)
rdd.filter(p => p.x>1)
  .count()
```

Frontend
API

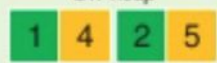
Catalyst

Generated
Java bytecode

Java bytecode in
Spark program and runtime

Backend
computation

Off-heap



Row-oriented

Java heap

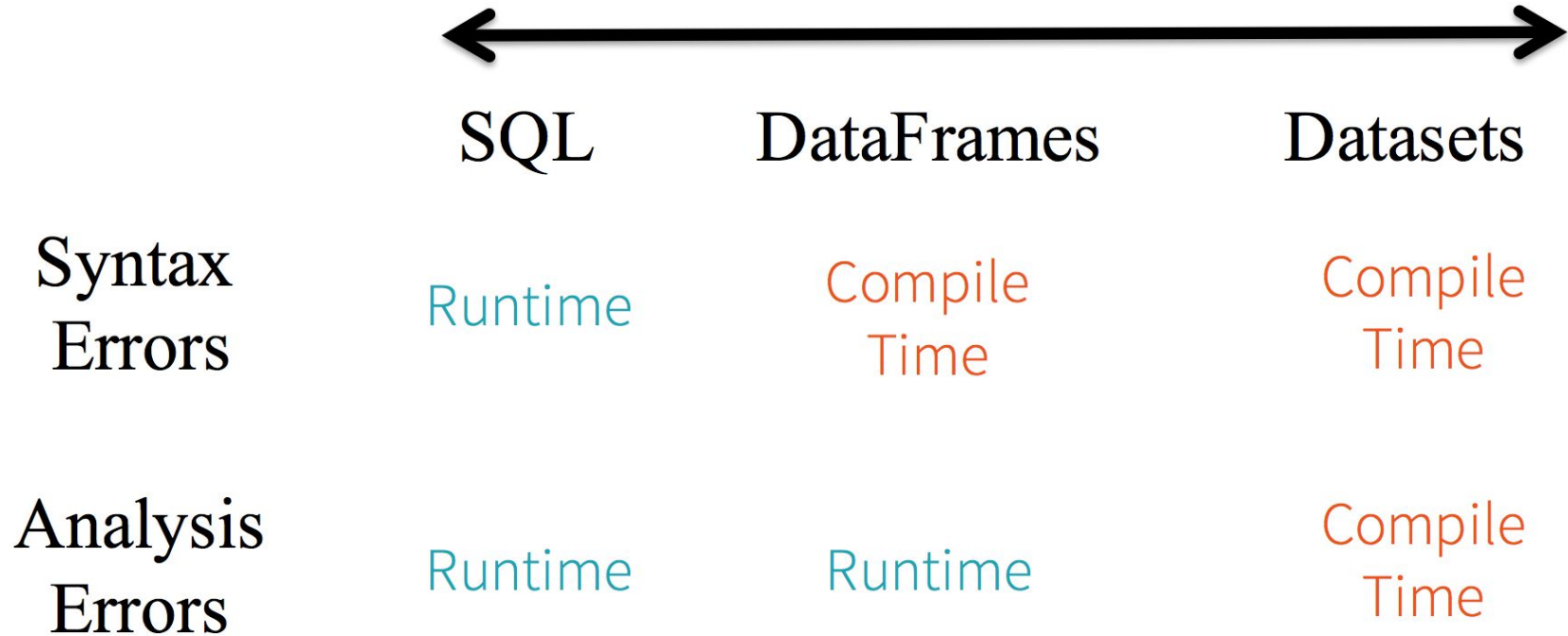


Row-oriented

Data



SQL vs. DataFrame vs. Dataset

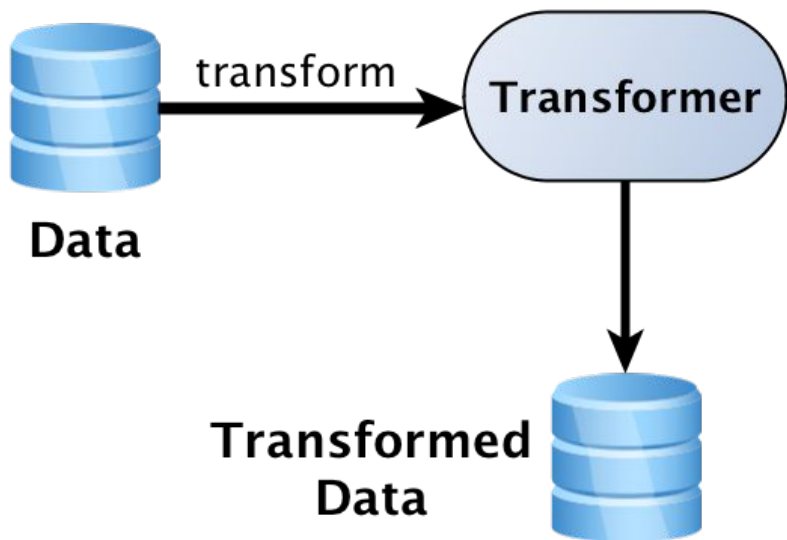


Spark ML Pipelines



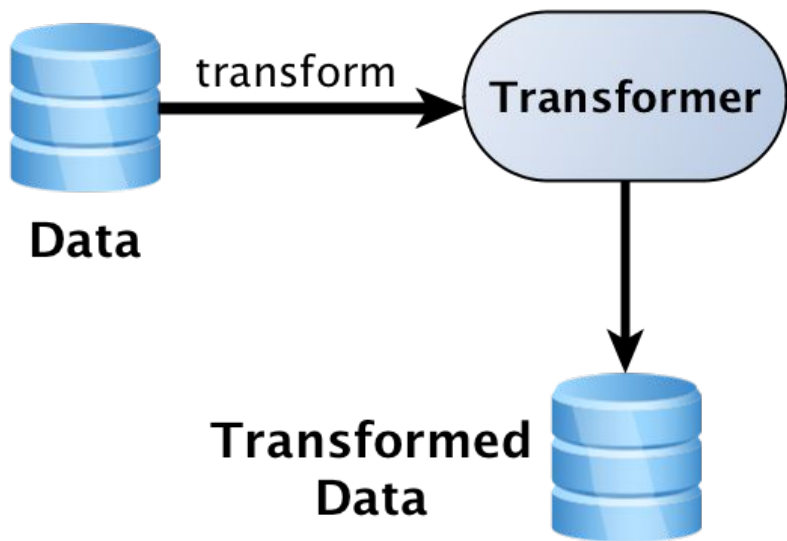
Spark ML Pipelines

Transformer

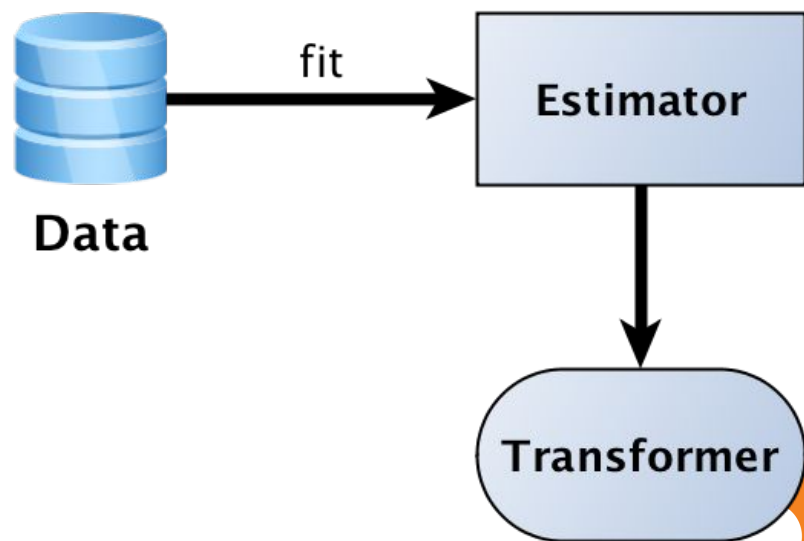


Spark ML Pipelines

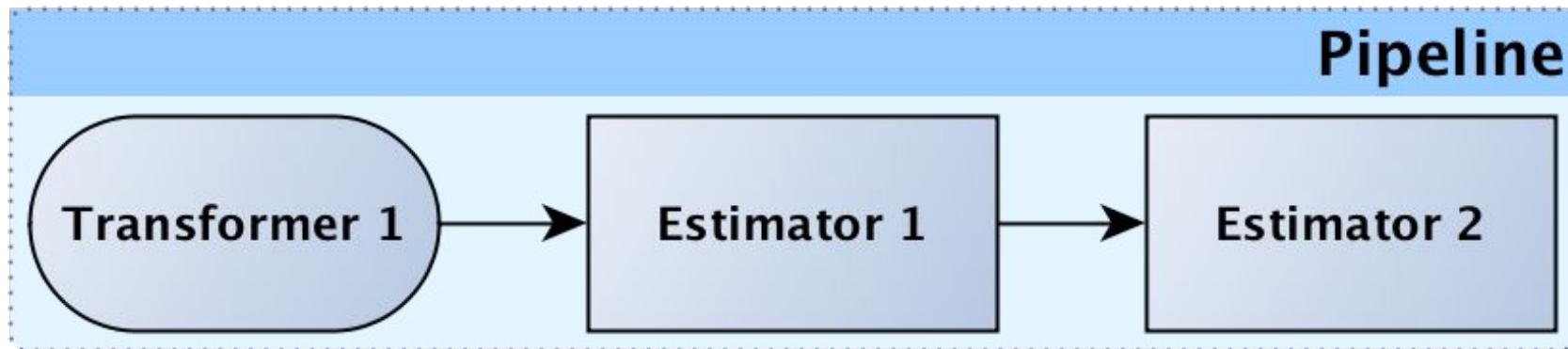
Transformer



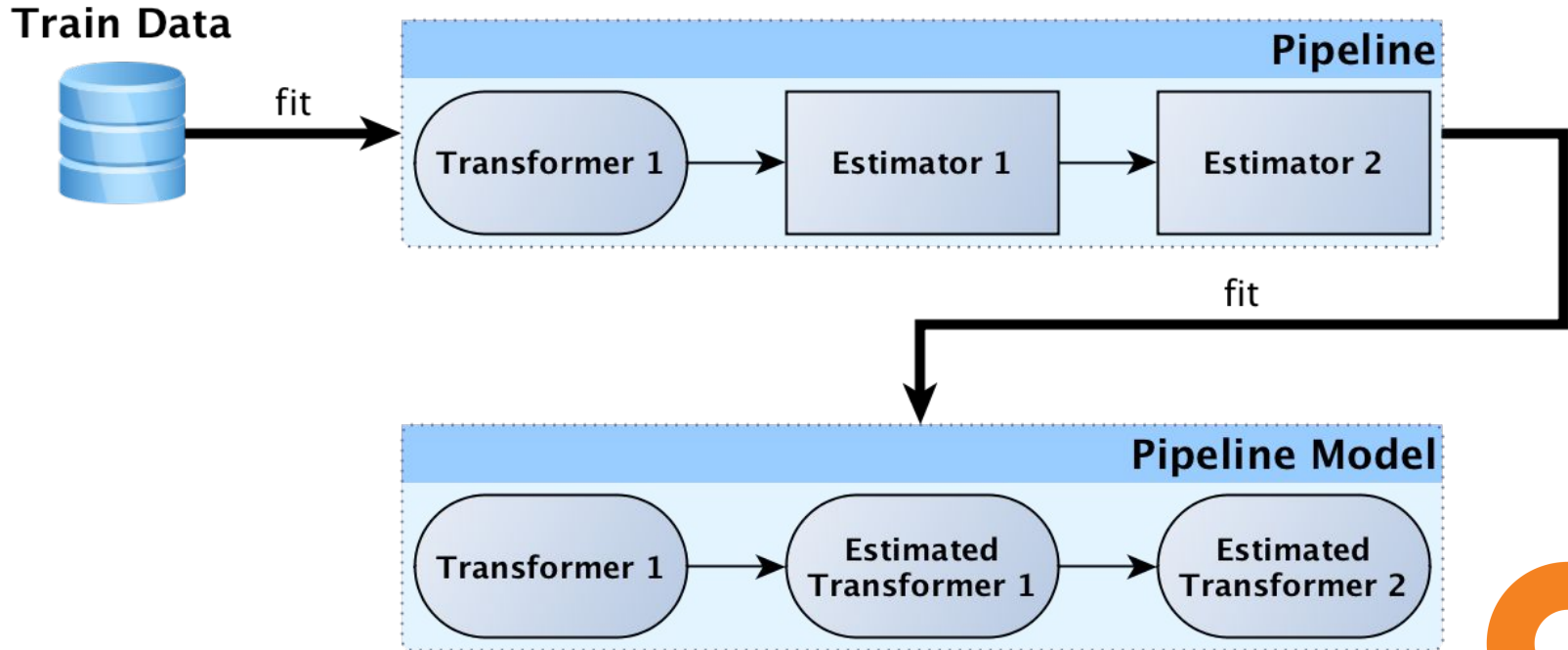
Estimator



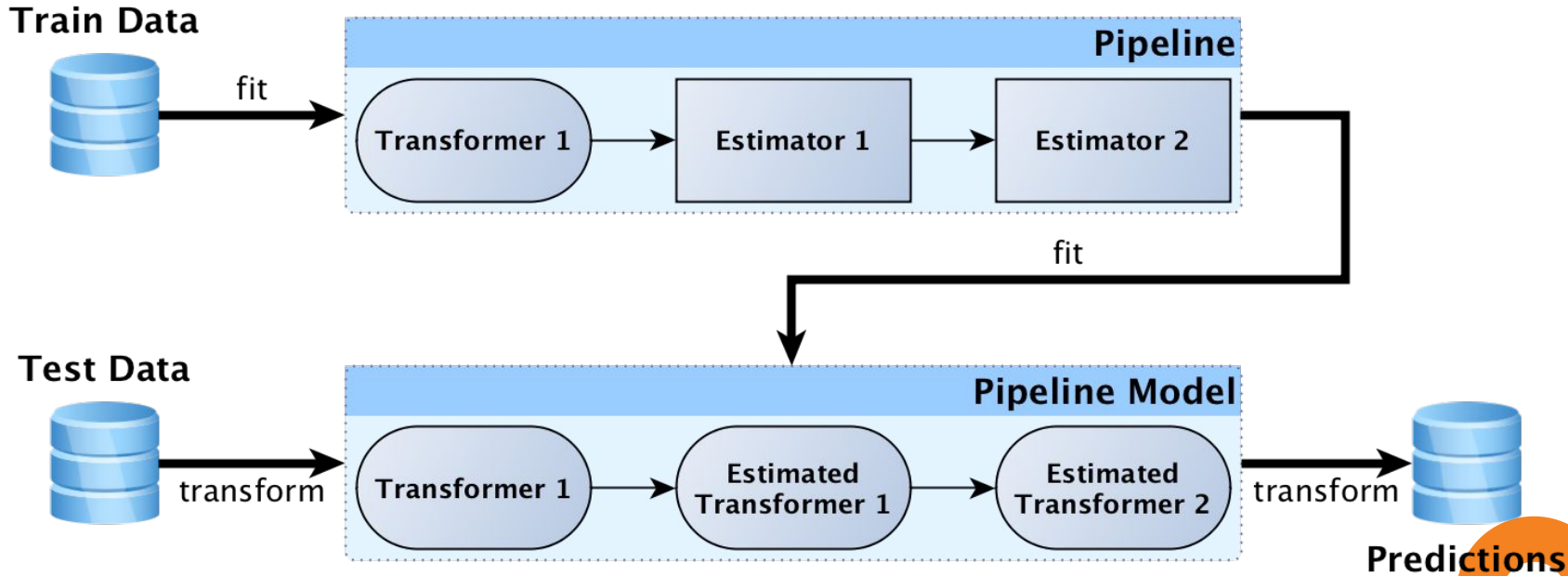
Spark ML Pipelines



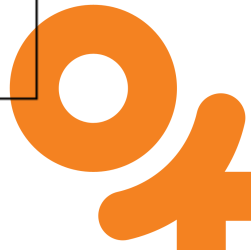
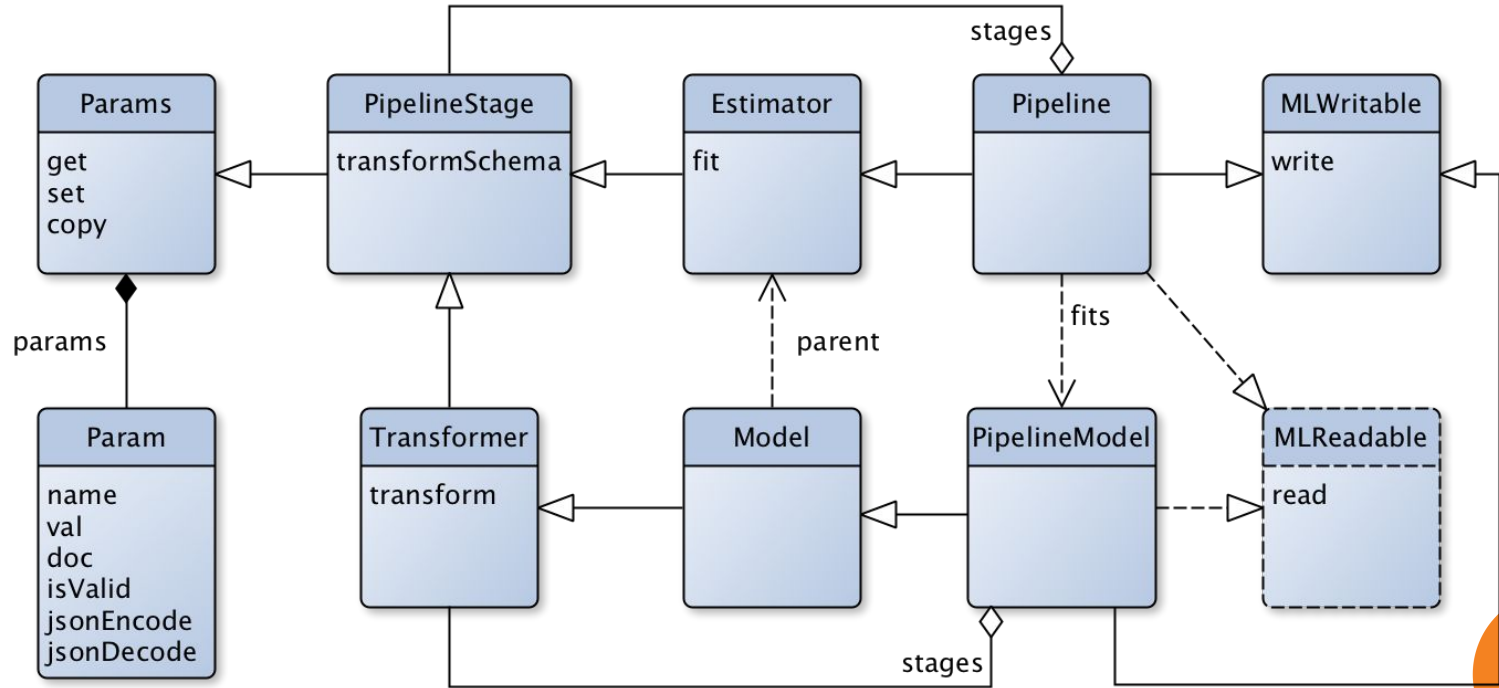
Spark ML Pipelines



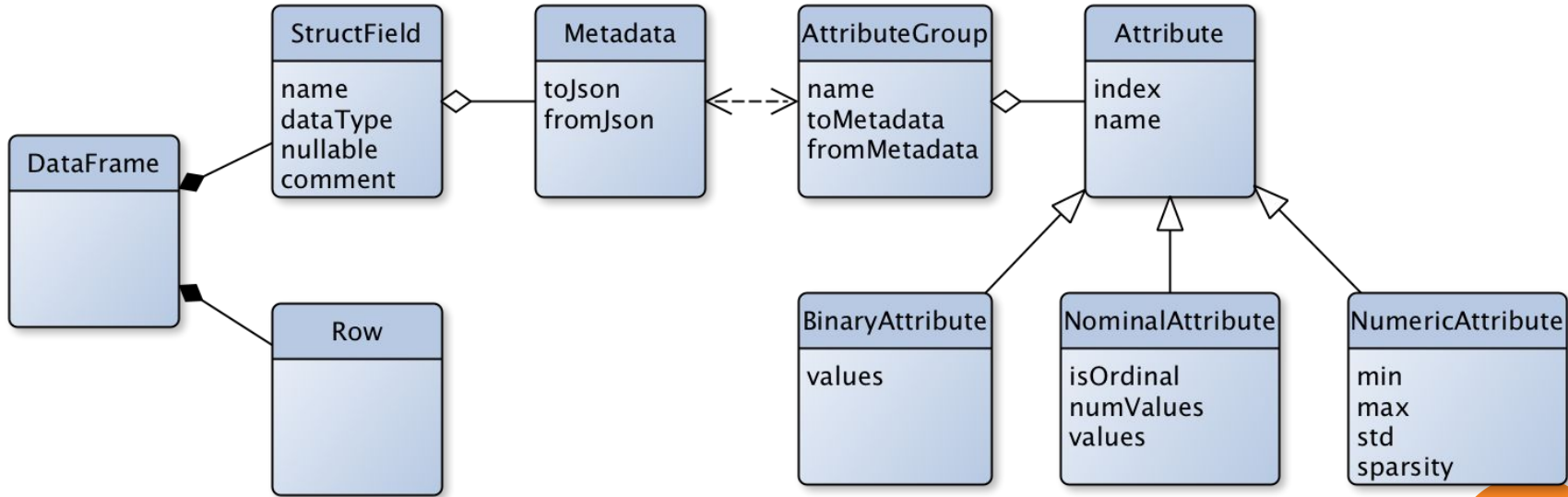
Spark ML Pipelines



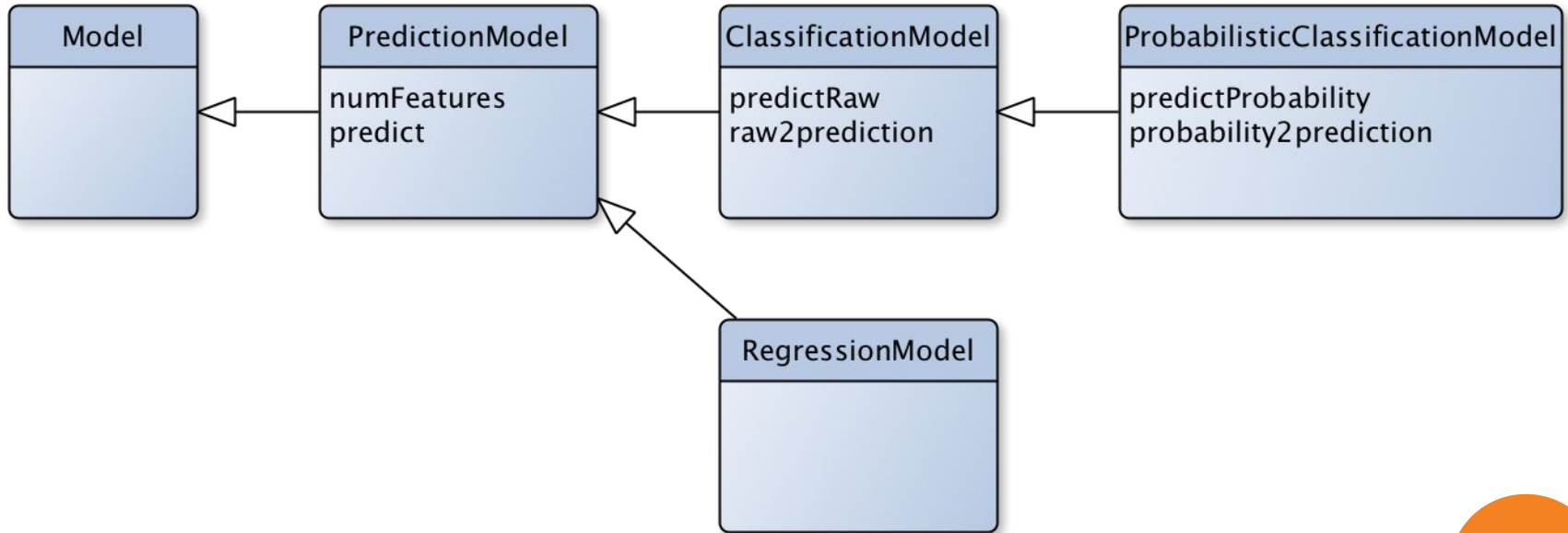
Spark ML Core



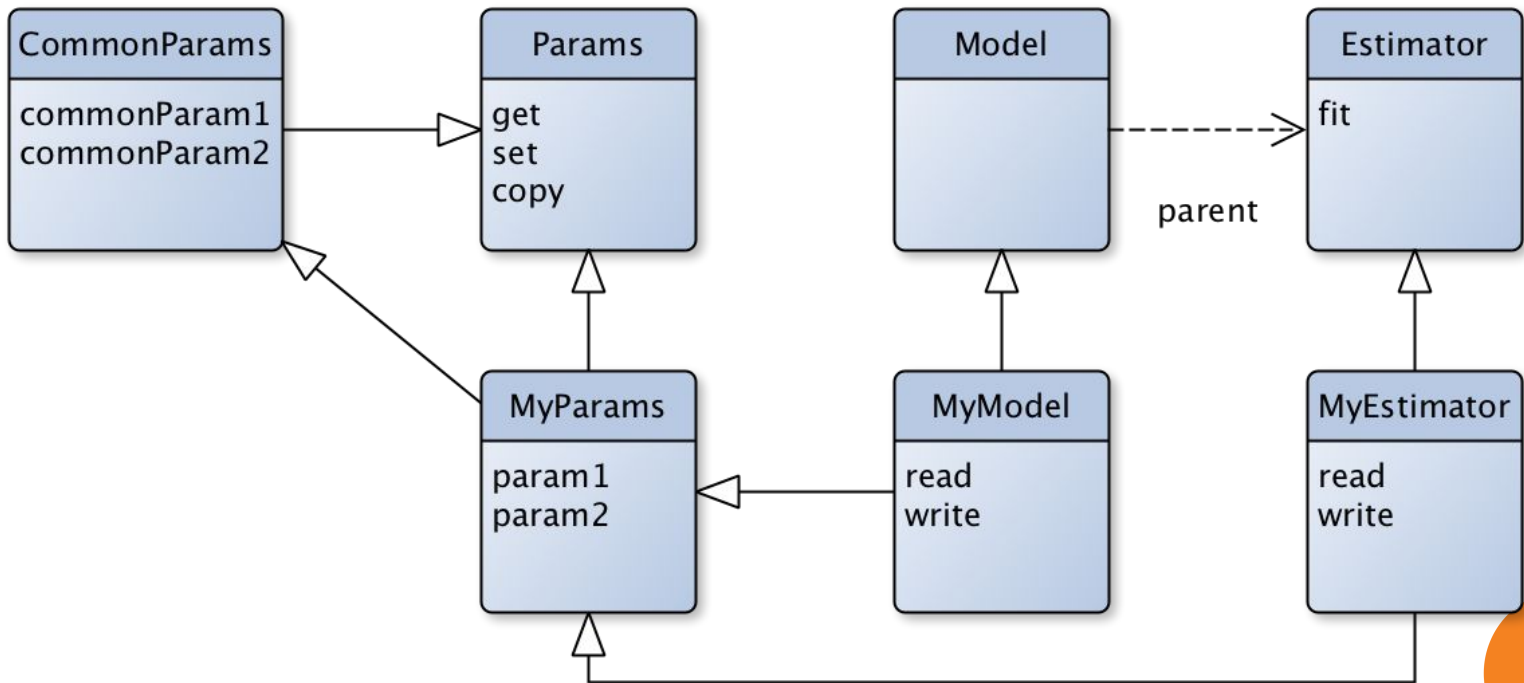
Field Metadata and Attributes



Prediction Model



“My Spark ML Model”



Spark ML Features

- ETL
 - SQLTransformer
 - *SqlFilter, ColumnsExtractor*
- Numerization
 - OneHotEncoder
 - StringIndexer
 - *MultinomialExtractor*
- Vectorization
 - VectorAssembler
 - FeatureHasher
 - *AutoAssembler*
- Feature Normalization
 - MaxAbsScaler
 - MinMaxScaler
 - Normalizer
 - QuantileDiscretizer
 - StandardScaler
- Missing values
 - Imputer
 - *NullToDefaultReplacer*
 - *NaNToMeanReplacer*



Spark ML Features

- Feature Engineering
 - DCT
 - ElementwiseProduct
 - Interaction
 - VectorIndexer
 - PolynomialExpansion
- Feature Selection
 - ChiSqSelector
 - *FoldedFeaturesSelector*
- Dimension reduction
 - PCA
 - MinHashLSHModel
 - BucketedRandomProjectionLSH
 - *RandomProjectionsHasher*



Spark ML Features

- Texts extraction
 - Tokenizer
 - RegexTokenizer
 - Ngram
 - StopWordsRemover
- NLP in Pravada-ML
 - *LanguageDetectorTransformer*
 - *LanguageAwareAnalyzer*
 - *NGramExtractor*
 - *URLEliminator*
- Texts vecotization
 - CountVectorizer
 - HashingTF
 - IDF
- Text embedding
 - Word2Vec
- Clustering
 - LDA
 - KMeans/BisectingKMeans
 - GaussianMixture



Spark ML Features

Regression

-  AFTSurvivalRegression.scala
-  DecisionTreeRegressor.scala
-  GBTRegressor.scala
-  GeneralizedLinearRegression.scala
-  IsotonicRegression.scala
-  LinearRegression.scala
-  RandomForestRegressor.scala

Classification

-  DecisionTreeClassifier.scala
-  GBTClassifier.scala
-  LinearSVC.scala
-  LogisticRegression.scala
-  MultilayerPerceptronClassifier.scala
-  NaiveBayes.scala
-  OneVsRest.scala
-  ProbabilisticClassifier.scala
-  RandomForestClassifier.scala



Spark ML Features

- Recommendations
 - ALS
 - FPGrowth
- Evaluation
 - BinaryClassificationEvaluator
 - ClusteringEvaluator
 - MulticlassClassificationEvaluator
 - RegressionEvaluator
- Tuning
 - ParamGridBuilder
 - CrossValidator
- More from Pravda-ML
 - *CombinedModel*
 - *PartitionedRankingEvaluator*
 - *CRRSampler*
 - *XGBoost*
 - *StochasticHyperopt*



Thank you for your attention!

Dmitry.Bugaychenko
@corp.mail.ru

