Промышленное машинное обучение на Spark

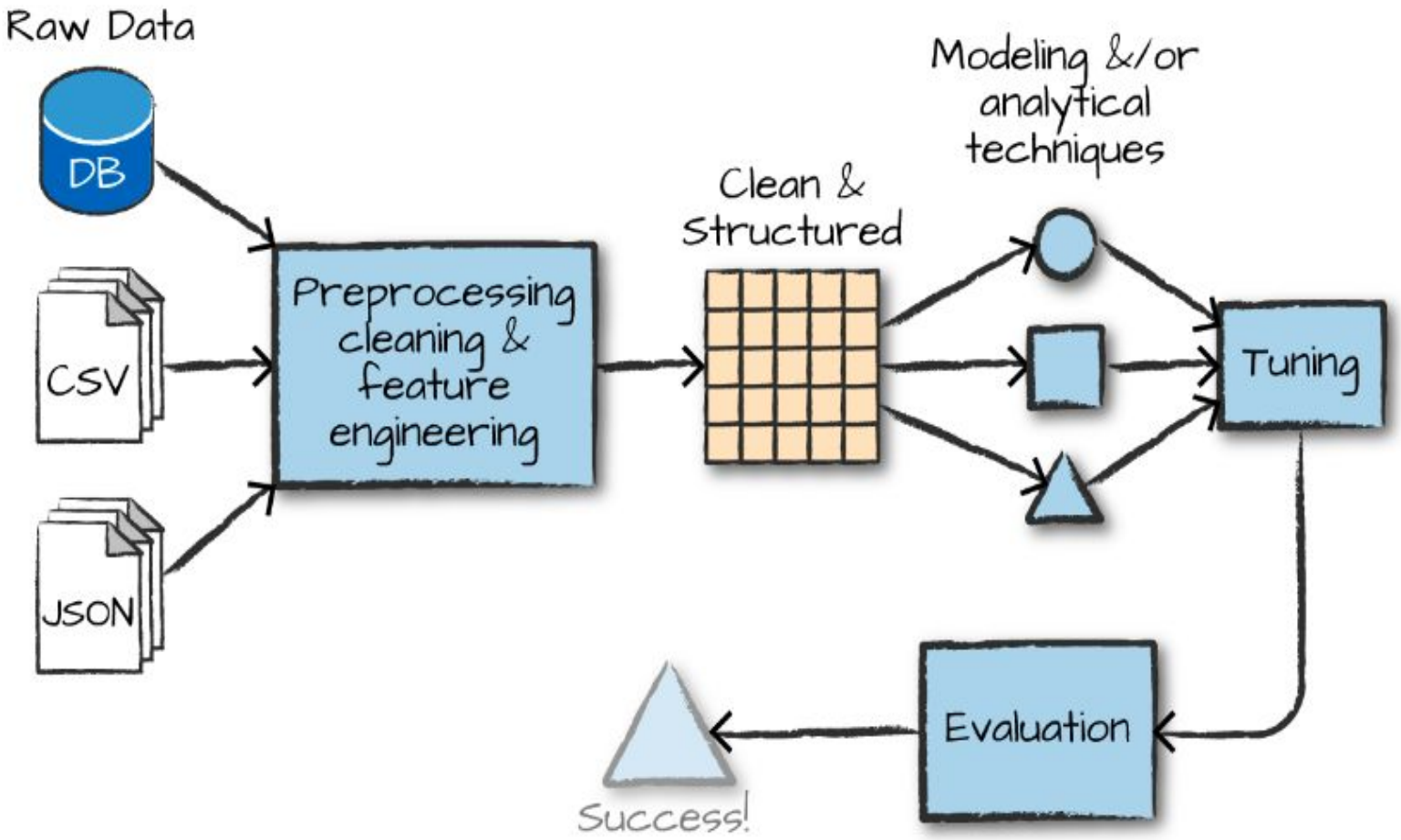
Лекция 6: Spark ML

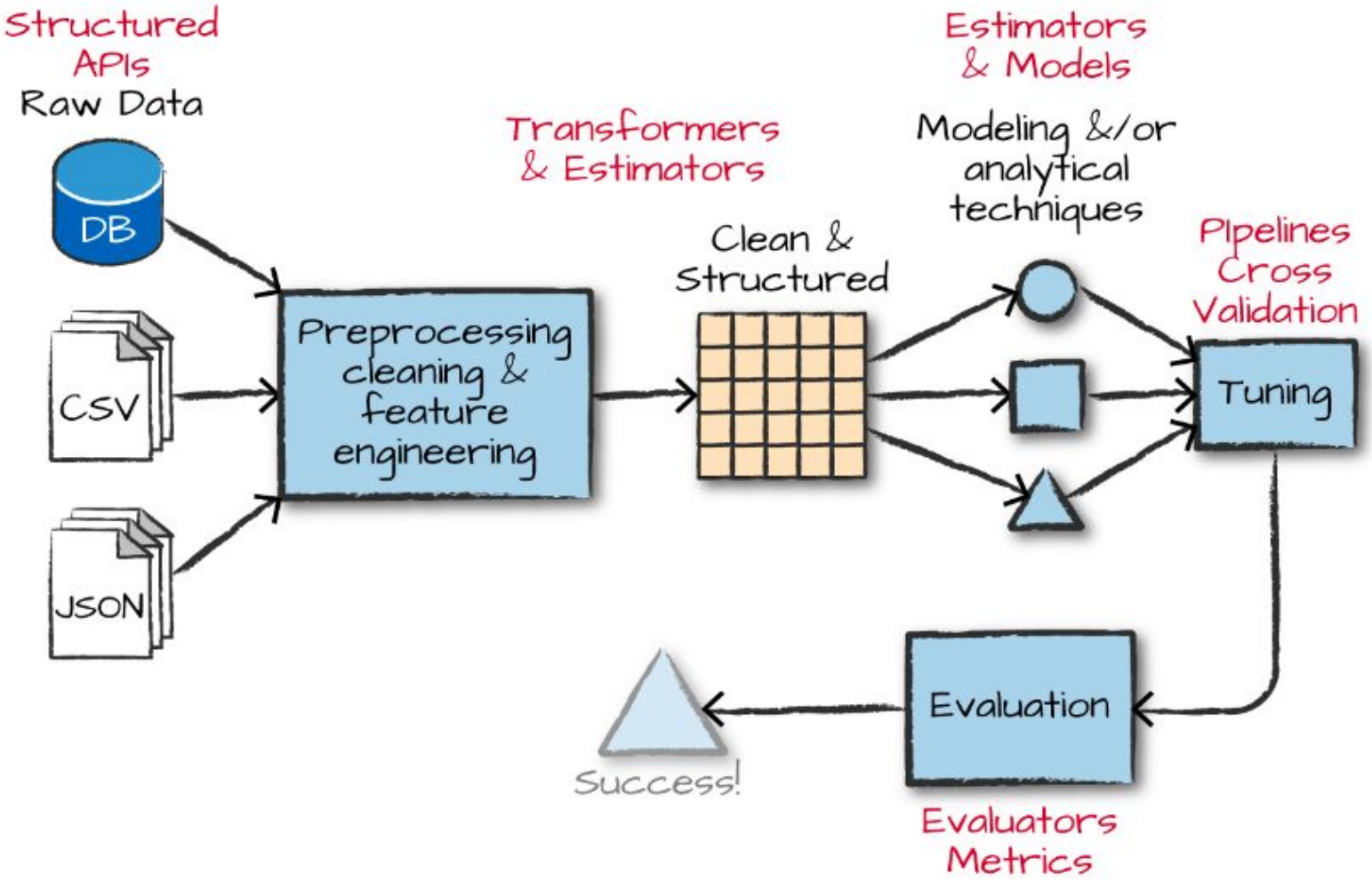
18 ноября 2023 г.

Содержание курса

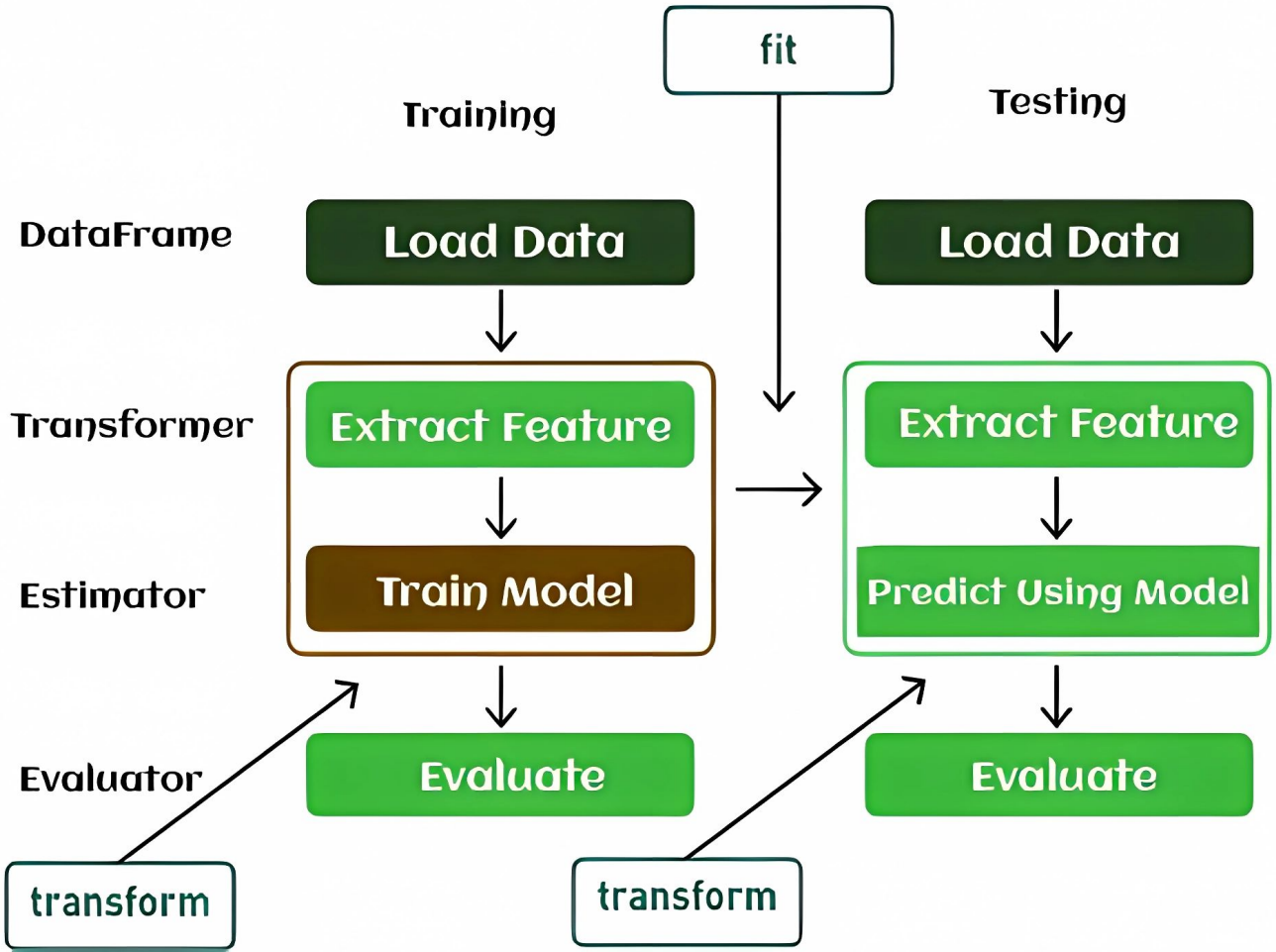
Что позволяет Spark.ML

http://spark.apache.org/docs/latest/ml-features.html

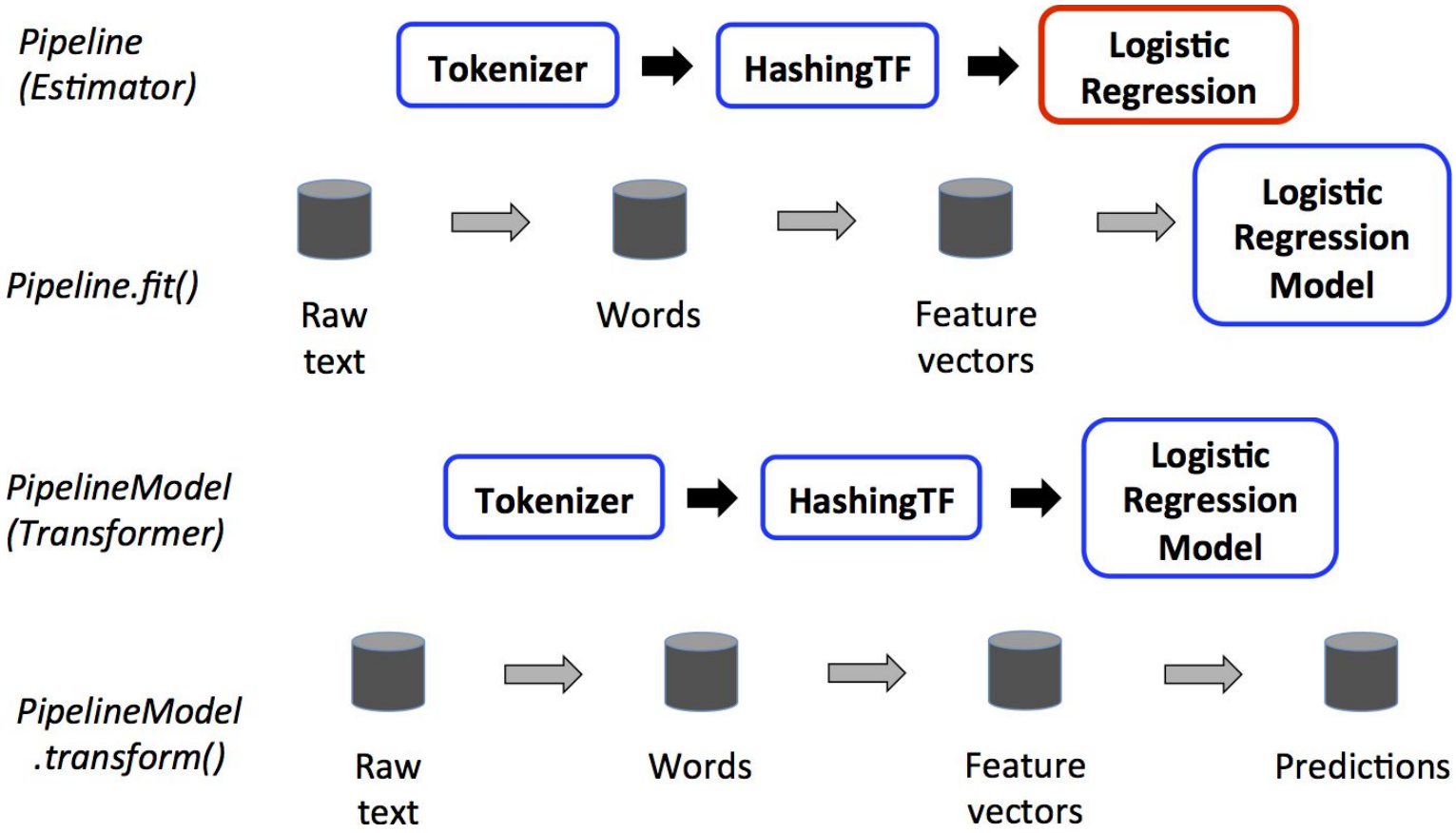

Raw Data
DB
CSV
JSON
Preprocessing cleaning & feature engineering
Clean & Structured
Modeling &/or analytical techniques
Tuning
Evaluation
Success!


Structured APIs
Raw Data
DB
CSV
JSON
Preprocessing cleaning & feature engineering
Transformers & Estimators
Clean & Structured
Estimators & Models
Modeling &/or analytical techniques
Pipelines Cross Validation
Tuning
Evaluation
Success!
Evaluators Metrics

Spark ML Workflow



Pipeline


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

Transformers

- Tokenizer
- StopWordsRemover
- n-gram
- Binarizer
- PCA
- PolynomialExpansion
- Discrete Cosine Transform (DCT)
- StringIndexer
- IndexToString
- OneHotEncoder
- VectorIndexer
- Interaction

- Normalizer
- StandardScaler
- RobustScaler
- MinMaxScaler
- MaxAbsScaler
- Bucketizer
- ElementwiseProduct
- SQLTransformer
- VectorAssembler
- VectorSizeHint
- QuantileDiscretizer
- Imputer

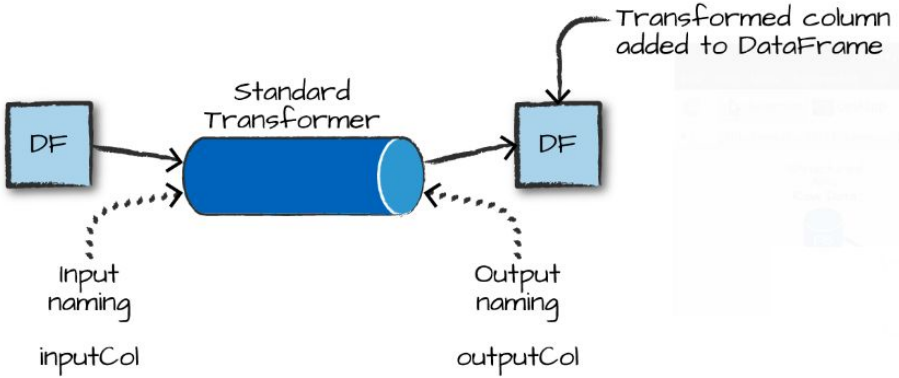


Figure 24-3. A standard transformer

Estimators

- LogisticRegression
- DecisionTreeClassifier
- RandomForestClassifier
- NaiveBayes
- OneVsRest
- MultilayerPerceptronClassifier

- KMeans
- LDA
- GaussianMixture
- ALS
- DecisionTreeRegressor
- LinearRegression
- RandomForestRegressor

Evaluators

- BinaryClassificationEvaluator
- RegressionEvaluator
- MulticlassClassificationEvaluator
- MultilabelClassificationEvaluator
- ClusteringEvaluator