SPARK MACHINE LEARNING

Background

WHY SPARK?

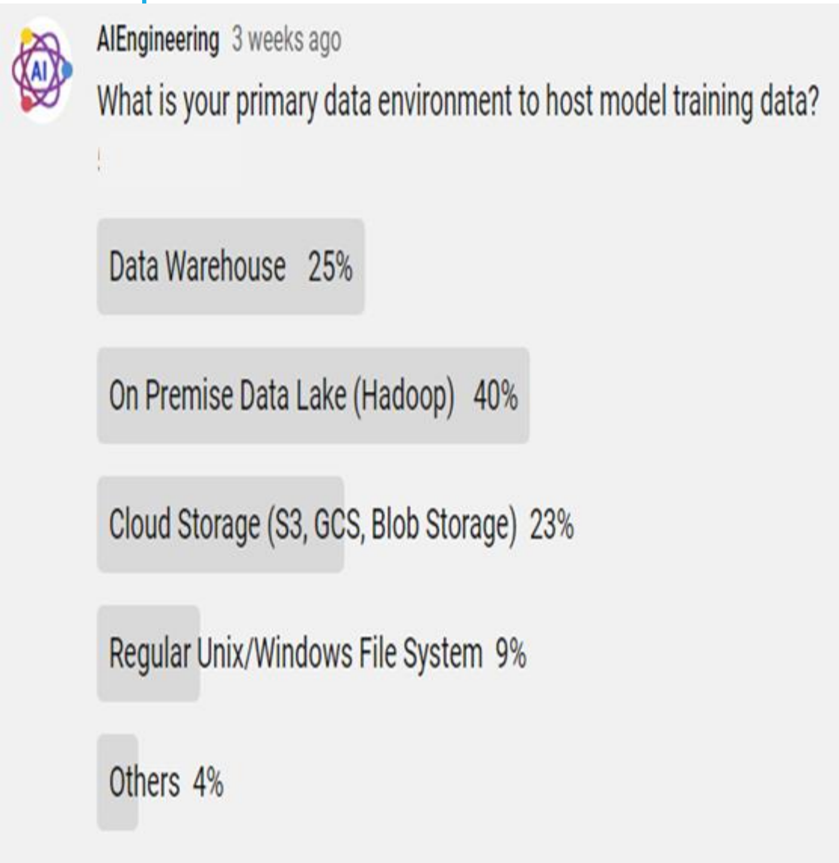
- Open Source distributed cluster computing framework with in memory data processing engine
- Spark can perform ETL, Streaming, Machine Learning, Graph processing on data at rest or in motion
- Support for Python, Scala, Java, R, SQL
- In memory computing compared to MR two staged disk based computing engine
- Created for Big Data workload
- Unified Engine

ADVANTAGES

- Enterprise has made huge investment in Big data and Spark today has become primary data processing framework

- Spark helps you create unified data pipeline from engineering till model training

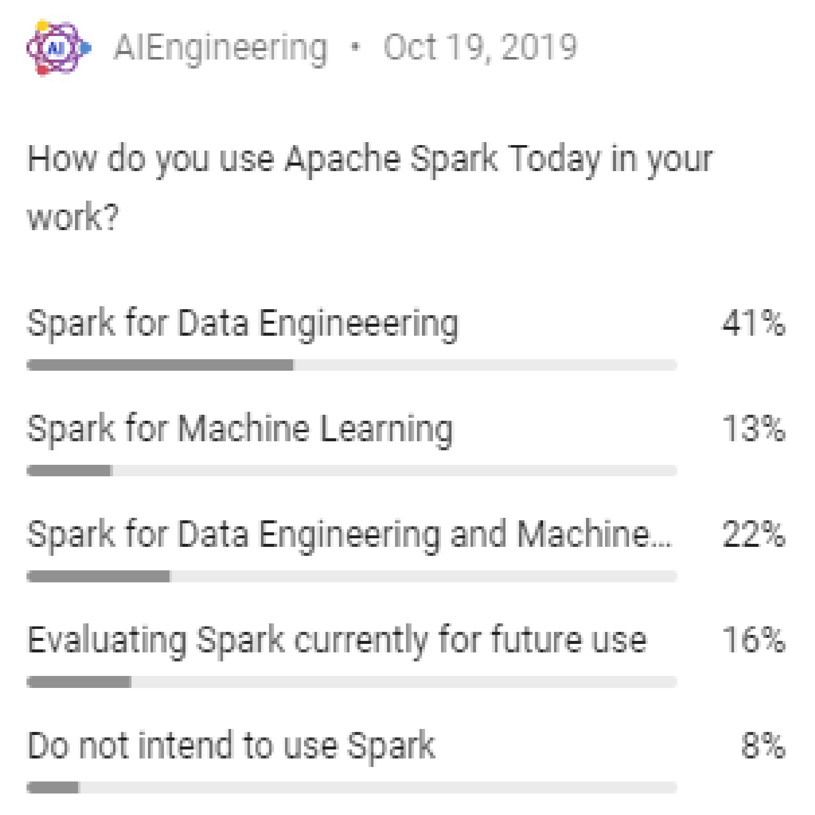
- Easy to migrate to cloud and hybrid cloud



AIEngineering 3 weeks ago

What is your primary data environment to host model training data?

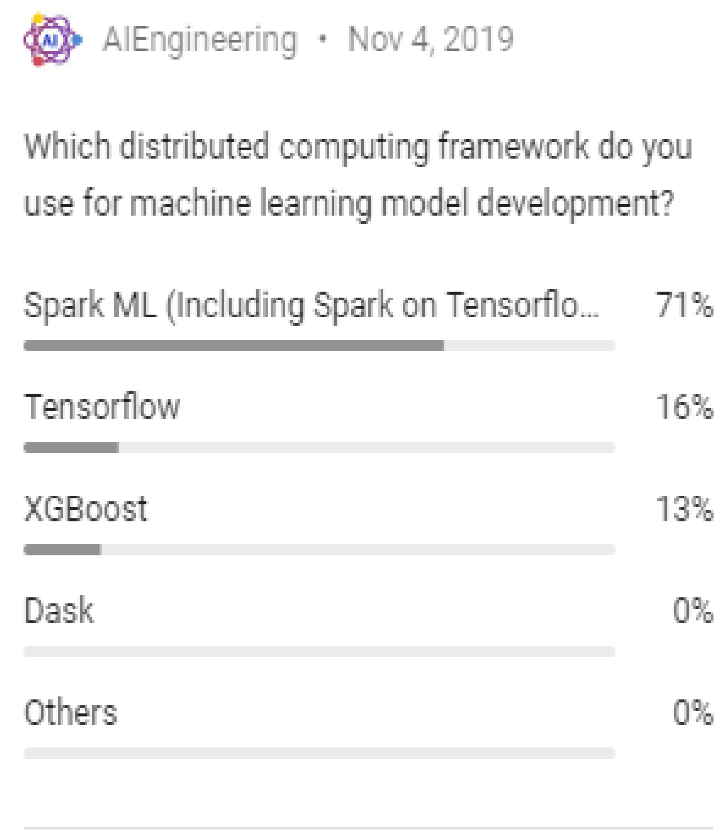
Data Warehouse	25%
On Premise Data Lake (Hadoop)	40%
Cloud Storage (S3, GCS, Blob Storage)	23%
Regular Unix/Windows File System	9%
Others	4%



AIEngineering • Oct 19, 2019

How do you use Apache Spark Today in your work?

Spark for Data Engineeering	41%
Spark for Machine Learning	13%
Spark for Data Engineering and Machine...	22%
Evaluating Spark currently for future use	16%
Do not intend to use Spark	8%



AIEngineering • Nov 4, 2019

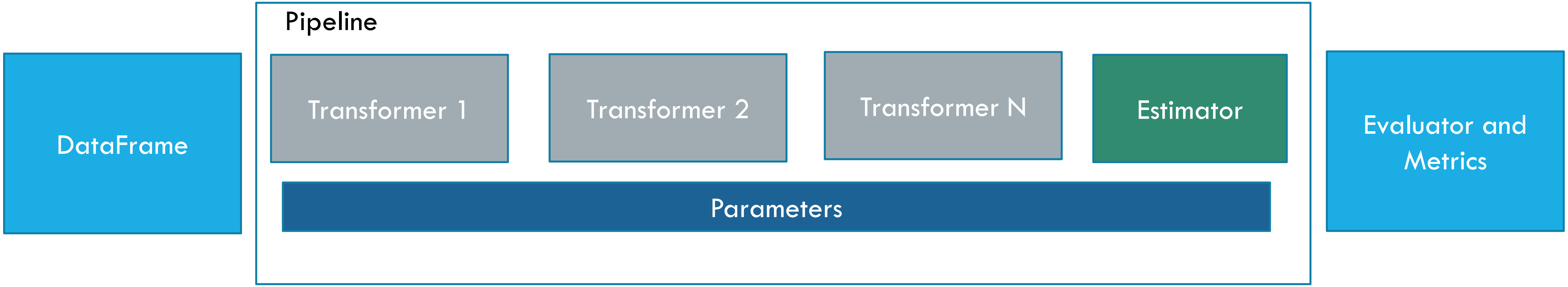
Which distributed computing framework do you use for machine learning model development?

Spark ML (Including Spark on Tensorflo...	71%
Tensorflow	16%
XGBoost	13%
Dask	0%
Others	0%

SPARK ML

Provides a set of Unified API for Machine Learning

SPARK ML PIPELINE



DataFrame

Pipeline

Transformer 1

Transformer 2

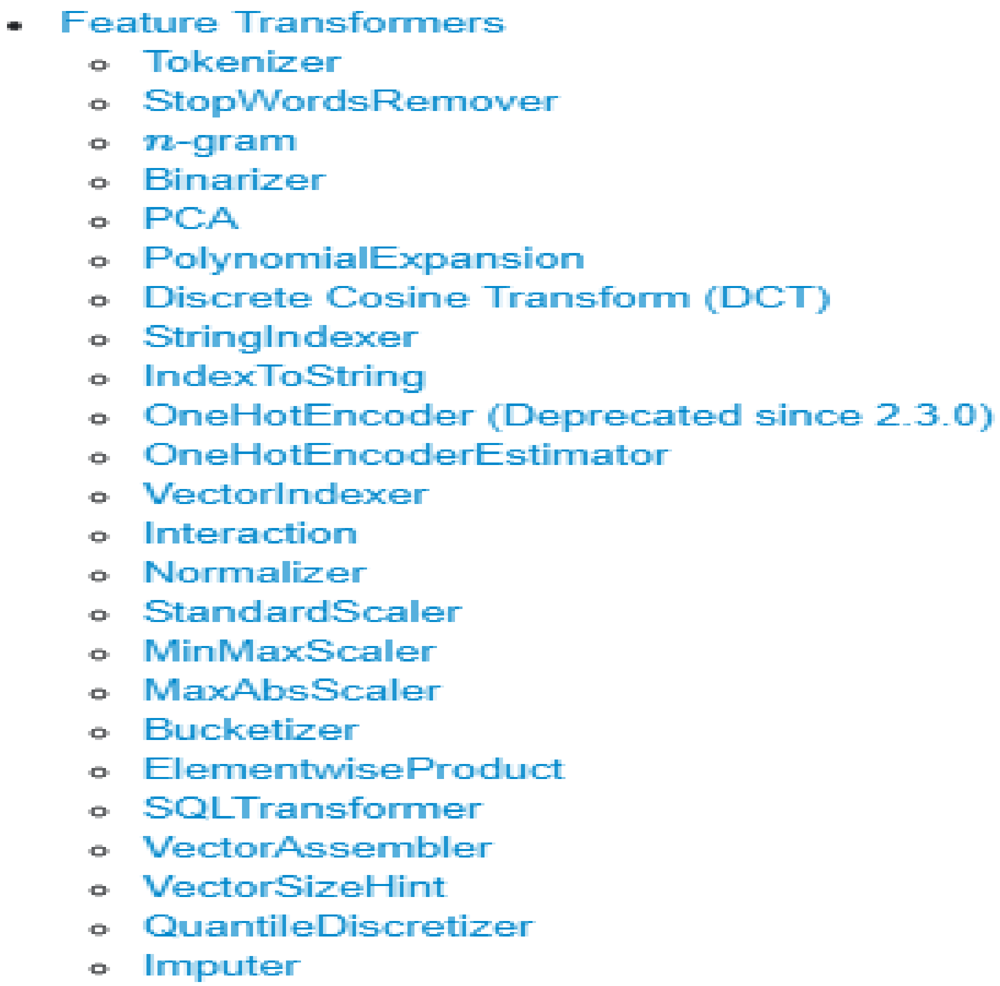
Transformer N

Estimator

Parameters

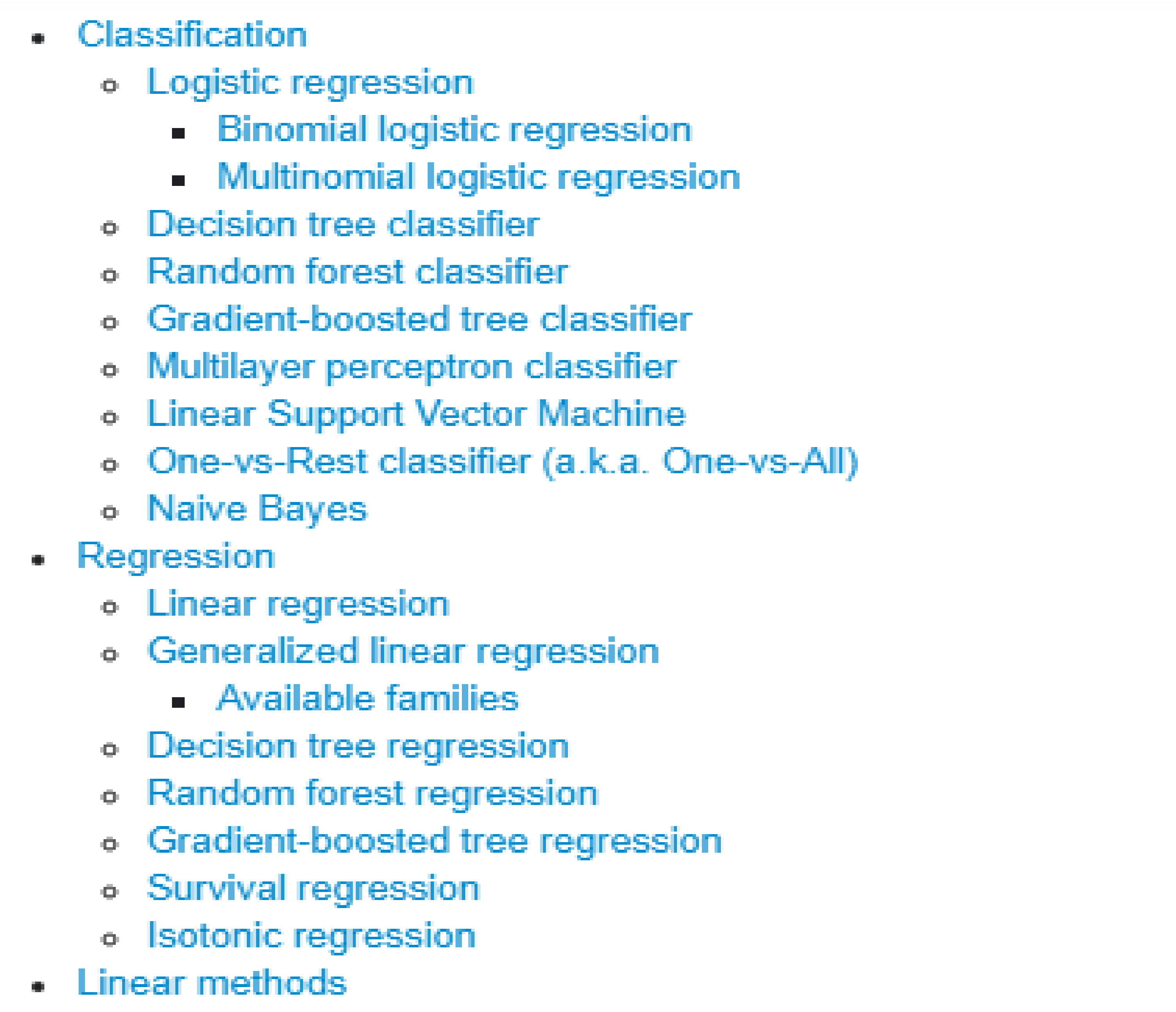
Evaluator and Metrics

TRANSFORMER



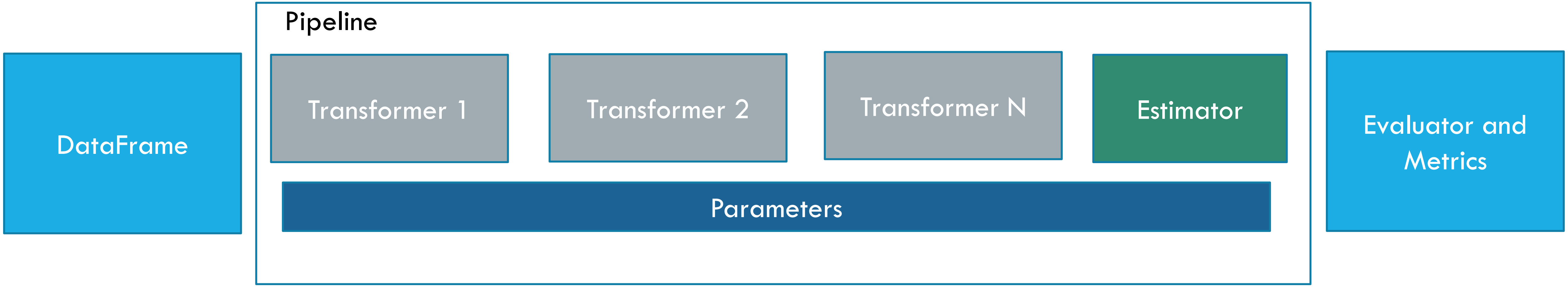
- Feature Transformers
  - Tokenizer
  - StopWordsRemover
  - n-gram
  - Binarizer
  - PCA
  - PolynomialExpansion
  - Discrete Cosine Transform (DCT)
  - StringIndexer
  - IndexToString
  - OneHotEncoder (Deprecated since 2.3.0)
  - OneHotEncoderEstimator
  - VectorIndexer
  - Interaction
  - Normalizer
  - StandardScaler
  - MinMaxScaler
  - MaxAbsScaler
  - Bucketizer
  - ElementwiseProduct
  - SQLTransformer
  - VectorAssembler
  - VectorSizeHint
  - QuantileDiscretizer
  - Imputer

ESTIMATOR



- Classification
  - Logistic regression
    - Binomial logistic regression
    - Multinomial logistic regression
  - Decision tree classifier
  - Random forest classifier
  - Gradient-boosted tree classifier
  - Multilayer perceptron classifier
  - Linear Support Vector Machine
  - One-vs-Rest classifier (a.k.a. One-vs-All)
  - Naive Bayes
- Regression
  - Linear regression
  - Generalized linear regression
    - Available families
  - Decision tree regression
  - Random forest regression
  - Gradient-boosted tree regression
  - Survival regression
  - Isotonic regression
- Linear methods

SPARK ML PIPELINE



```
from pyspark.ml import Pipeline

#data preparation (e.g., VectorAssembler, VectorIndexer, etc.)
transformer1 = …
transformer2 = …
transformer3 = …

#Model algorithm (e.g. DecisionTreeClassifier)
model_algorithm = …

#Pipeline which applys transformation and model building algorithm on dataset
pipeline = Pipeline(stages=[transformer1, transformer2, transformer3, model_algorithm])
model = pipeline.fit(training)
```