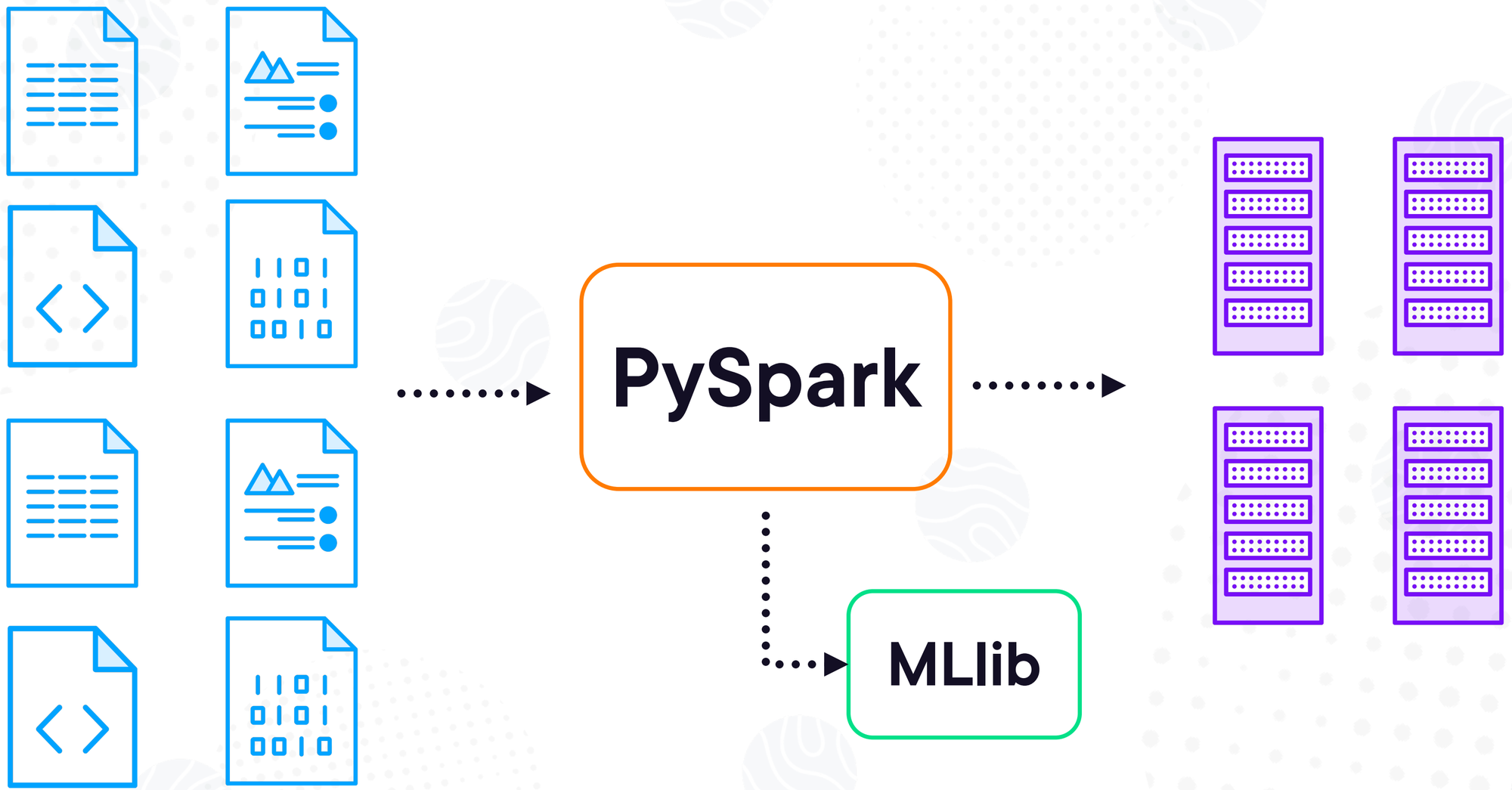
Scalable Machine Learning with PySpark MLlib

Understanding PySpark MLlib

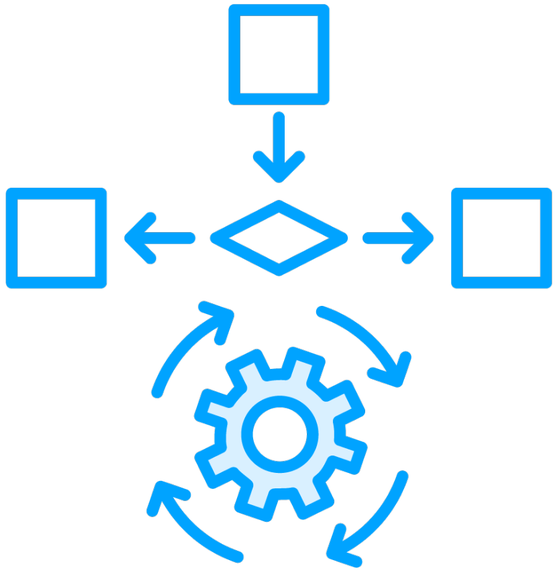
Warner Chaves
Solutions Architect

@warchav | www.createdatapros.com

What Is PySpark MLlib?


PySpark
MLlib

What Is in MLlib?



Common ML algorithms

Pipeline capabilities

ML utilities

Distributed Machine Learning

Structured, semi-structured, and unstructured data

Scale out ML operations to parallel operations on multiple nodes

Scalability and fault tolerance of Spark

Advantages of PySpark MLlib

Python is the most popular programming language

Data ingestion, transformation, and ML in one environment

High-level APIs (DataFrames) for easier development

Built-in ML algorithms for classification, regression, clustering, etc.

MLlib vs. Spark ML

Spark MLlib

- Original RDD library

Spark ML

- DataFrame, pipeline-centric update

PySpark

- Can use both but we will focus on the newer API

MLlib is the umbrella term and the library supports both approaches

New code typically uses the Spark ML (DataFrame-based) APIs

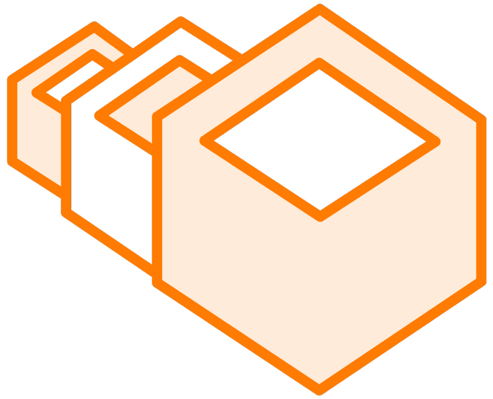
This course references these collectively as PySpark MLlib

Where Can PySpark MLlib Run?

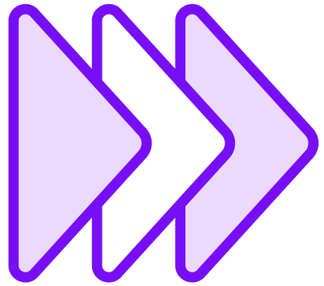
Self-hosted	Vs.	Managed
On-prem or cloud Spark clusters		Databricks (AWS, Azure, GCP)
Physical or virtual machines		AWS EMR, GCP DataProc, MS Fabric
Container-based with Kubernetes		Flexible, no cluster management
Full control of configuration and infrastructure		Pre-configured tools

Use Case for PySpark MLlib

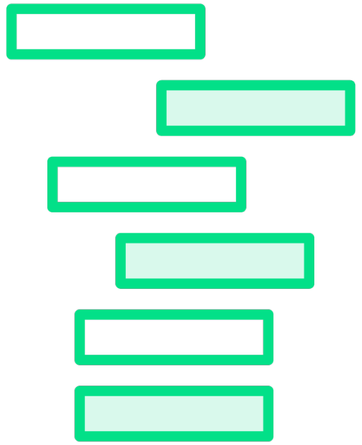
Key Strengths of PySpark MLlib



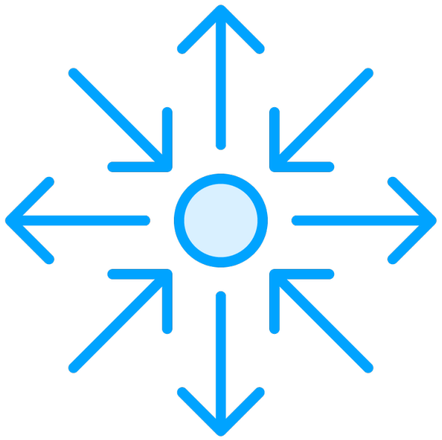
Scalability



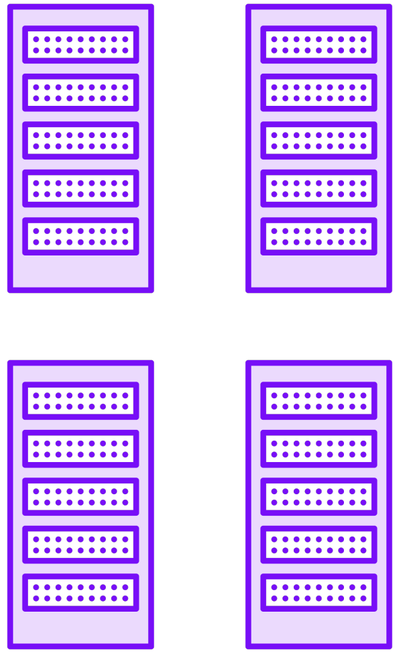
Speed



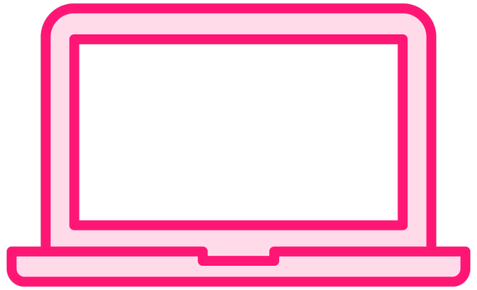
Integration



Fault tolerance

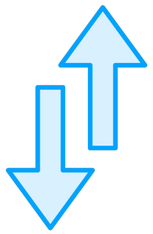


MLlib



Scikit-learn

Limitations



Might be too much overhead for small datasets

Limited algorithm support vs. sci-kit learn

Requires spark install and cluster setup

Choosing the Best Tool

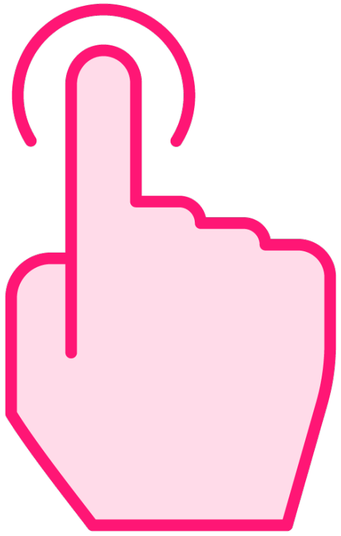
scikit-learn

- Data fits on single node
- No resource constraints
- Specialized ML algorithm

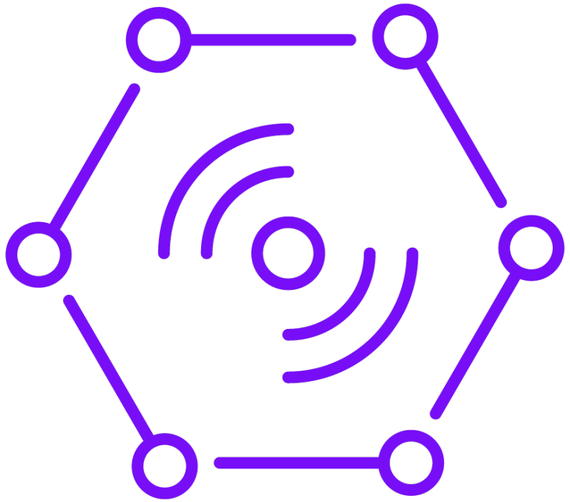
MLlib

- Large amounts of data
- Already using Spark
- Algorithm is supported

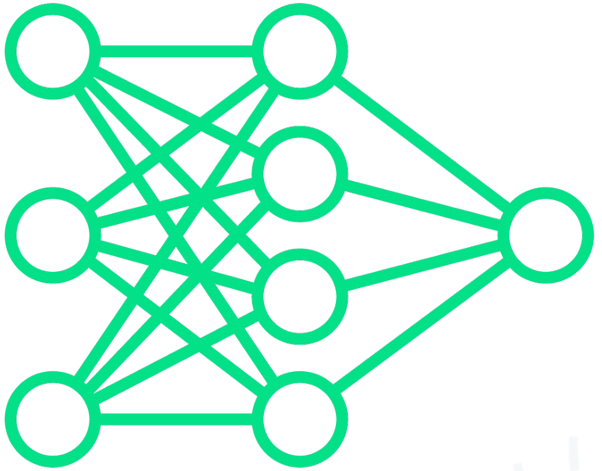
Real-world Scenarios



Clickstream or retail transaction data

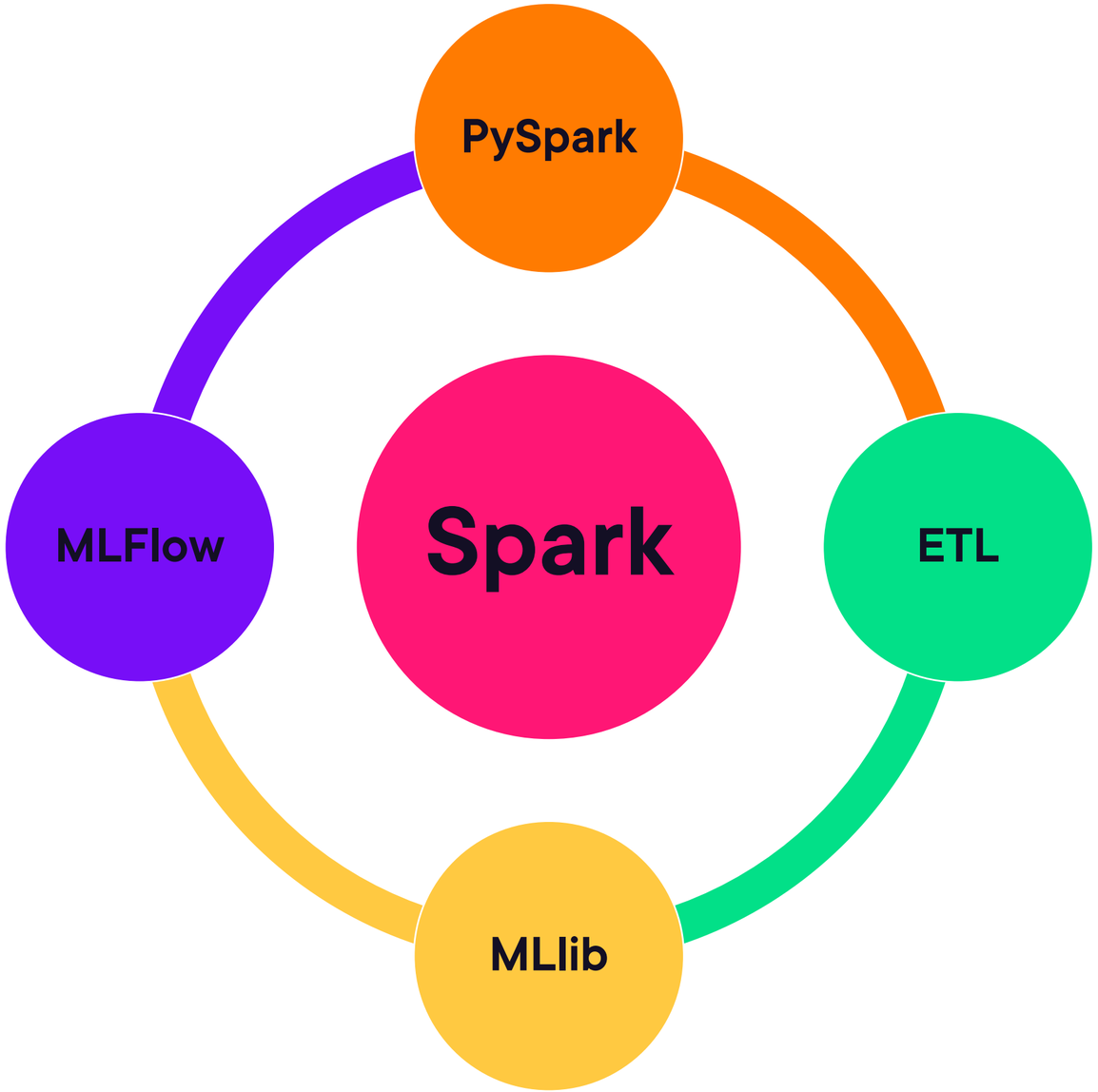


IoT sensor analysis across many devices



Distributed model training to reduce total runtime

Fitting into the Spark Ecosystem

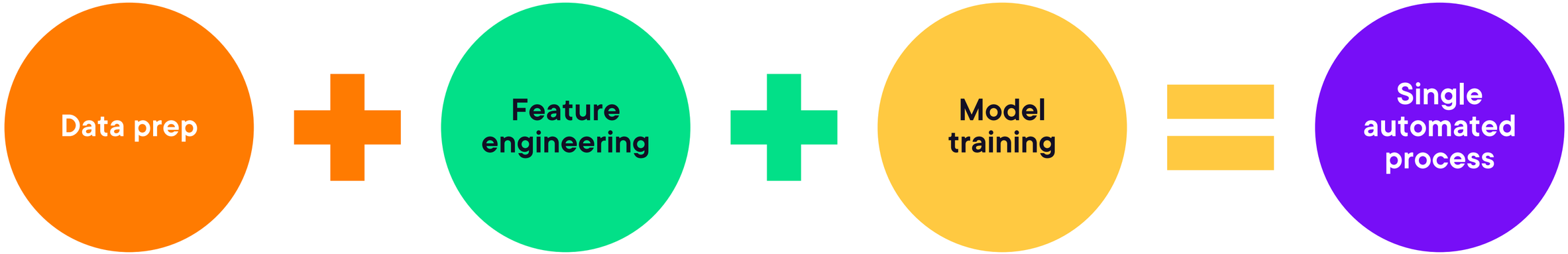


Spark ML Pipelines

ML Pipeline

A sequence of transformers and estimators that produce a trained ML model

Purpose of a Pipeline



Encourages modular and reusable ML code

Main Components of a Pipeline

DataFrames	Backbone for data access
Transformer	Applies a function to a dataframe
Estimator	Learns from data to produce a transformer

Once an estimator is fit, it becomes a trained model, which is a transformer in subsequent pipeline stages

Spark ML Pipelines Terms

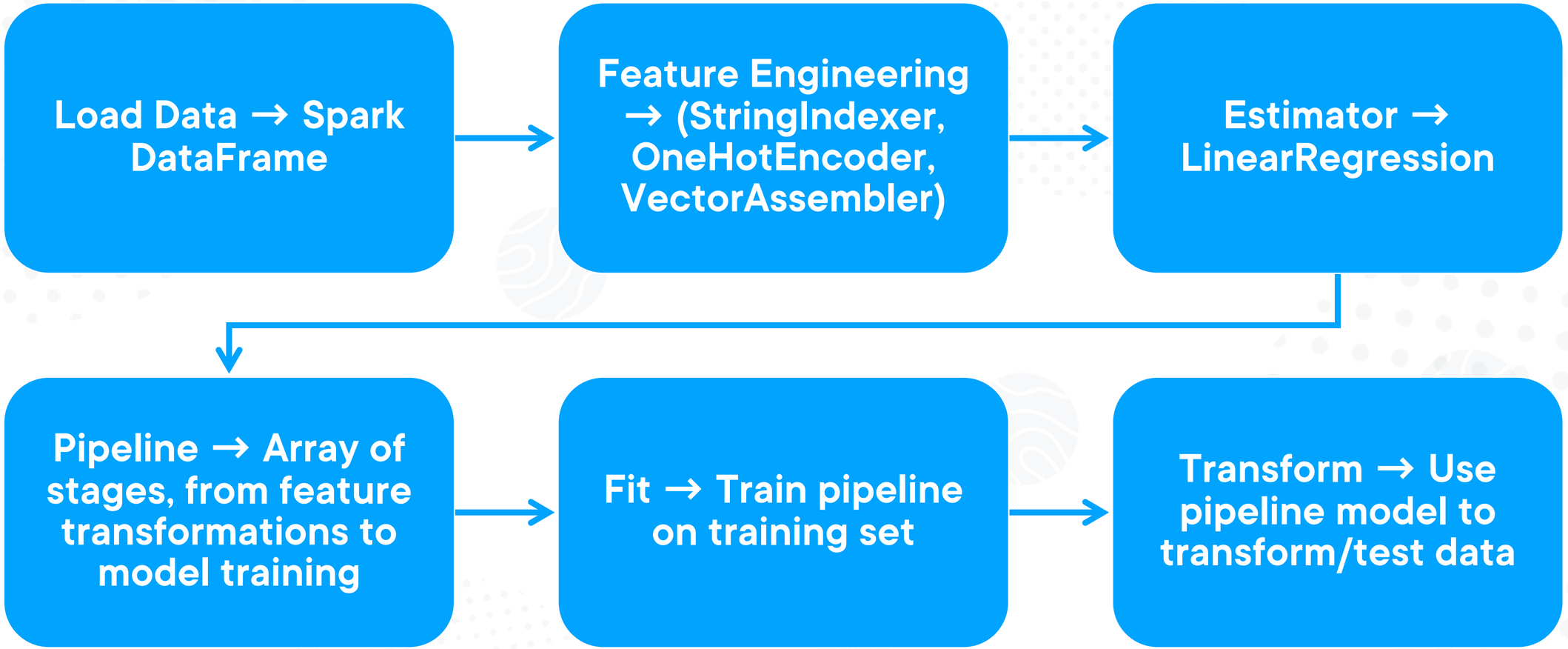
Stages: a list of transformers and estimators

Parameters: configurations for stages

Persistence: ability to save entire pipeline and models to disk

PipelineModel: A pipeline that has been fit on data

Workflow Example



Scalable Machine Learning with PySpark MLlib