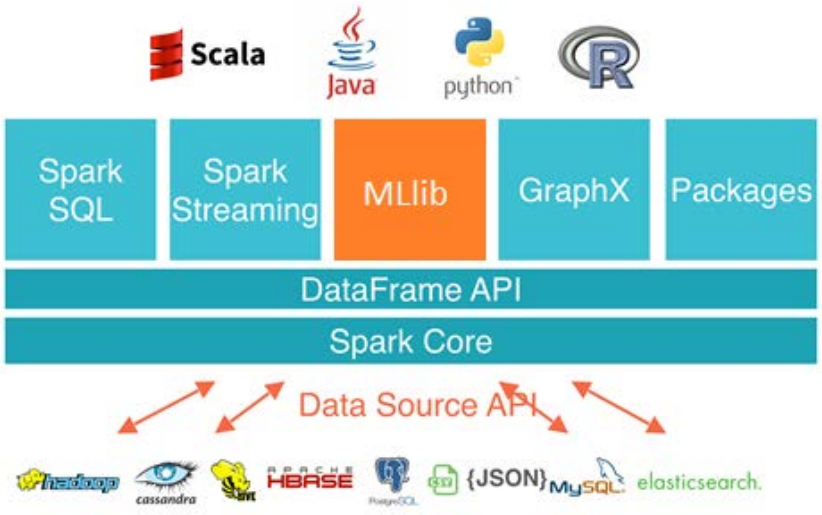
CMSC 691
High Performance Distributed Systems

Apache Spark MLlib

Dr. Alberto Cano
Assistant Professor
Department of Computer Science
acano@vcu.edu

Apache Spark Machine Learning Library (MLlib)

- Algorithms: classification, regression, frequent pattern mining, clustering, filtering, recommendation
- Data processing: feature extraction, transformation, dimensionality reduction, and feature selection
- Utilities: pipelines, persistence, linear algebra, statistics, etc



DataFrames

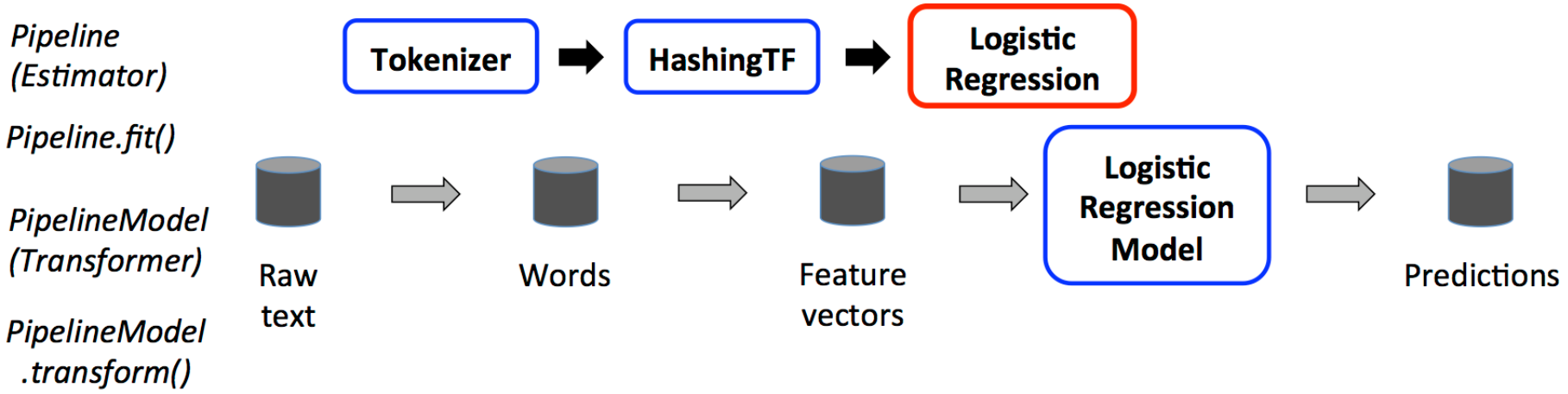
- As of Spark 2.0, the DataFrame-based API is primary API but RDD-based API is now in maintenance mode
- Data abstraction for working with structured and semi-structured data, i.e. datasets with a schema or metadata (as a table in relational databases, schema + data)
- DataFrame is a distributed collection of tabular data organized into rows and named columns storing text, feature vectors, true labels, and predictions
- Don’t worry (actually do because the documentation is a mess), but the API allows to convert from RDD to DataFrame and vice versa, right now samples work for both representations

ML Pipelines

- Defines a workflow to assemble, combine and configure multiple distributed algorithms into a single pipeline
- A practical ML pipeline often involves a sequence of data pre-processing, feature extraction, model fitting, and validation stages
- A pipeline chains multiple transformers and estimators together to specify a input/model/output sequence
- **Transformer**: A transformer is an algorithm which can transform one DataFrame into another DataFrame. E.g., an ML model which transforms a DataFrame features into predictions
- **Estimator**: An estimator is an algorithm which can be fit on a DataFrame to produce a Transformer. E.g., a learning algorithm is an Estimator which trains on a DataFrame and produces a model

ML Pipelines

- A Pipeline is specified as a sequence of stages, and each stage is either a transformer or an estimator
- For transformer stages, the *transform()* method is called on the DataFrame, generating a new transformed DataFrame
- For estimator stages, the *fit()* method is called to produce a transformer, for which *transform()* is called on the DataFrame



ML Pipelines

- Pipelines are also estimators (allows for multi-level pipelines)
- Persistent objects, can be read and loaded

Parameters

- A *Param* is a named parameter with self-contained documentation
- A *ParamMap* is a set of (parameter, value) pairs to configure a given transformer or estimator

Evaluators and CrossValidators

- A evaluator is a transformation that maps a DataFrame into a metric indicating how good a model is, e.g. Binary/MulticlassClassificationEvaluator, RegressionEvaluator

Installation

- Add Spark MLlib dependencies to your Maven project in Eclipse in addition to the Spark core ones

```
<dependency>
    <groupId>org.apache.spark</groupId>
    <artifactId>spark-core_2.11</artifactId>
    <version>2.0.1</version>
</dependency>
<dependency>
    <groupId>org.apache.spark</groupId>
    <artifactId>spark-mllib_2.10</artifactId>
    <version>2.0.1</version>
</dependency>
```

CMSC 691
High Performance Distributed Systems

Apache Spark MLlib

Dr. Alberto Cano
Assistant Professor
Department of Computer Science
acano@vcu.edu