Prof. Jean Carlos Alves

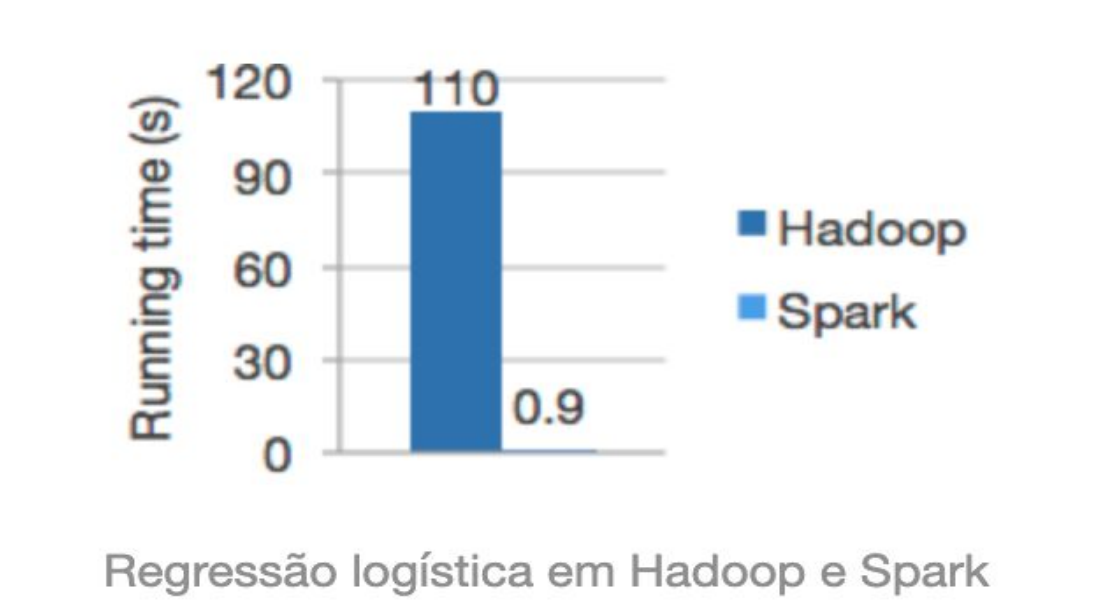
DataOps e Implantação de Sistemas de Machine Learning

PUC Minas

Spark

- Mecanismo de análise unificado para processamento de dados em grande escala
- Processamento Batch e Streaming
- Suporta Java, Scala, Python, R e SQL

Spark X MapReduce



Regressão logística em Hadoop e Spark

Bibliotecas Spark

- SQL e Dataframes
- Streaming
- Mlib
- GraphX

Spark - RDD

- A abstração principal que o Spark oferece é um conjunto de dados distribuído resiliente (RDD), que é uma coleção de elementos particionados nos nós do cluster que podem ser operados em paralelo. Os RDDs são criados começando com um arquivo no sistema de arquivos Hadoop (ou qualquer outro sistema de arquivos compatível com Hadoop) ou uma coleção Scala existente no programa do driver e transformando-o. Os usuários também podem pedir ao Spark para manter um RDD na memória, permitindo que ele seja reutilizado com eficiência em operações paralelas. Finalmente, os RDDs se recuperam automaticamente de falhas de nó.

Spark

- Transformação
- Ação

https://spark.apache.org/docs/latest/rdd-programming-guide.html#actions

Spark

- Transformação

Transformation	Meaning
map(*func*)	Return a new distributed dataset formed by passing each element of the source through a function *func*.
filter(*func*)	Return a new dataset formed by selecting those elements of the source on which *func* returns true.
flatMap(*func*)	Similar to map, but each input item can be mapped to 0 or more output items (so *func* should return a Seq rather than a single item).

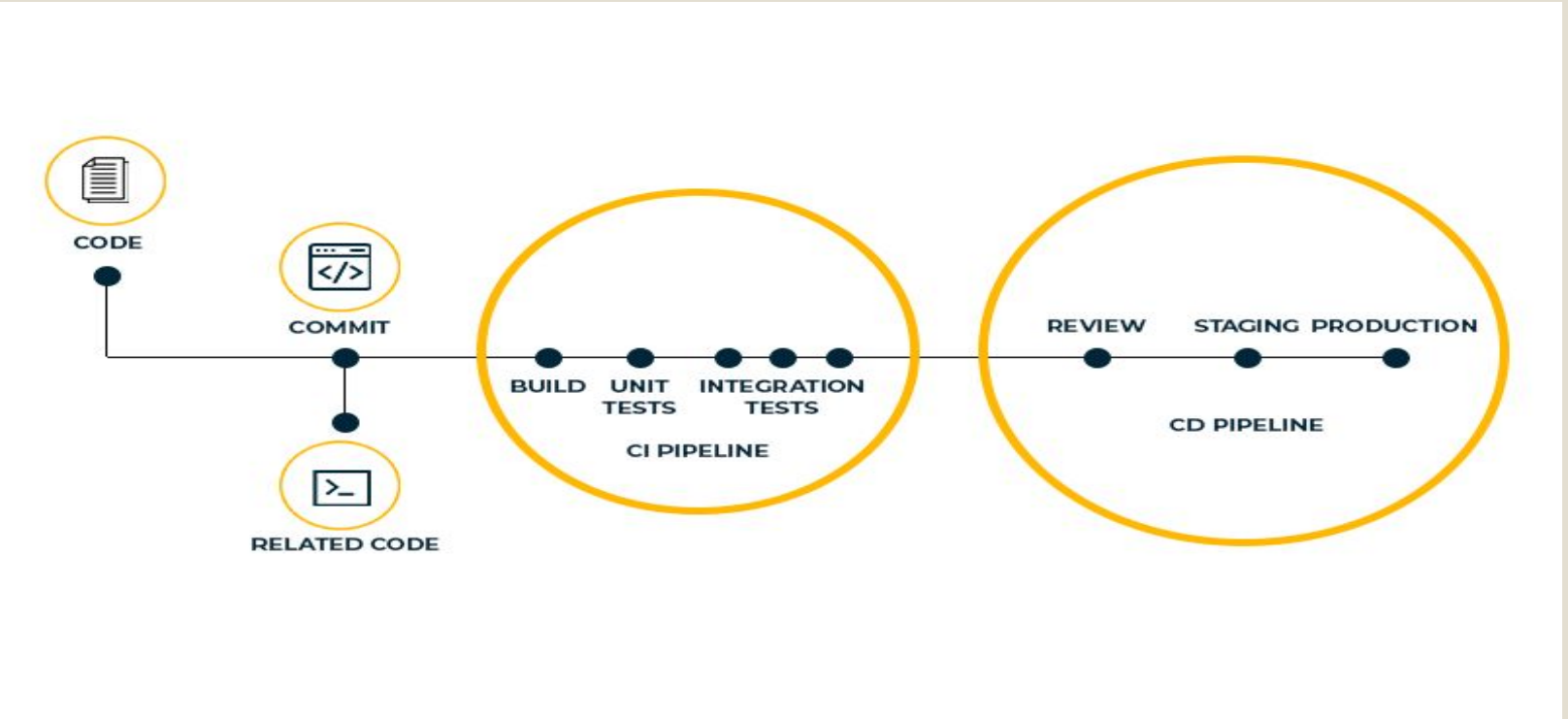
https://spark.apache.org/docs/latest/rdd-programming-guide.html#actions

Spark

Action	Meaning
reduce(*func*)	Aggregate the elements of the dataset using a function *func* (which takes two arguments and returns one). The function should be commutative and associative so that it can be computed correctly in parallel.
collect()	Return all the elements of the dataset as an array at the driver program. This is usually useful after a filter or other operation that returns a sufficiently small subset of the data.
count()	Return the number of elements in the dataset.
first()	Return the first element of the dataset (similar to take(1)).

https://spark.apache.org/docs/latest/rdd-programming-guide.html#actions

CI/CD



https://blog.convisoappsec.com/seguranca-no-seu-pipeline-ci-cd/

TDD - Test Driven Development

- Desenvolvimento de software orientado por testes.

Testes automatizados:

- xUnit -> .net
- unittest -> python
- JUnit -> java
- jest -> Node

BDD - Behavior Driven Development

- Desenvolvimento de software orientado por comportamento visando integrar regras de negócios a codificação.

Framework:

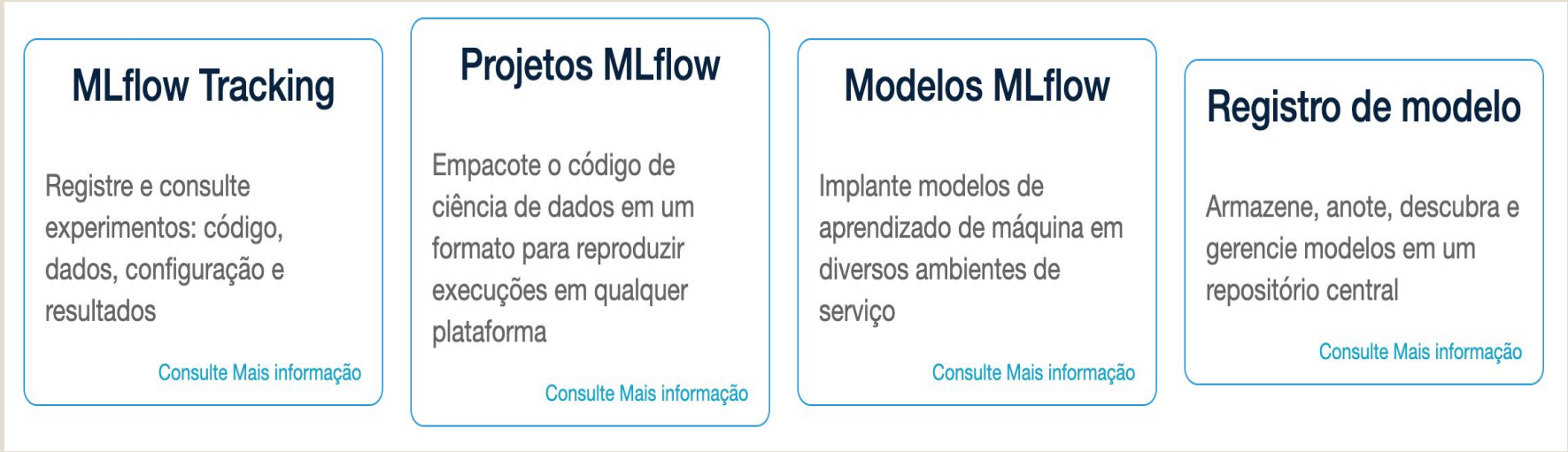
- behavior -> python
- specflow -> .net
- spock -> java

MLFlow

- MLflow é uma plataforma de código aberto para gerenciar o ciclo de vida de ML, incluindo experimentação, reprodutibilidade, implantação e um registro de modelo central.

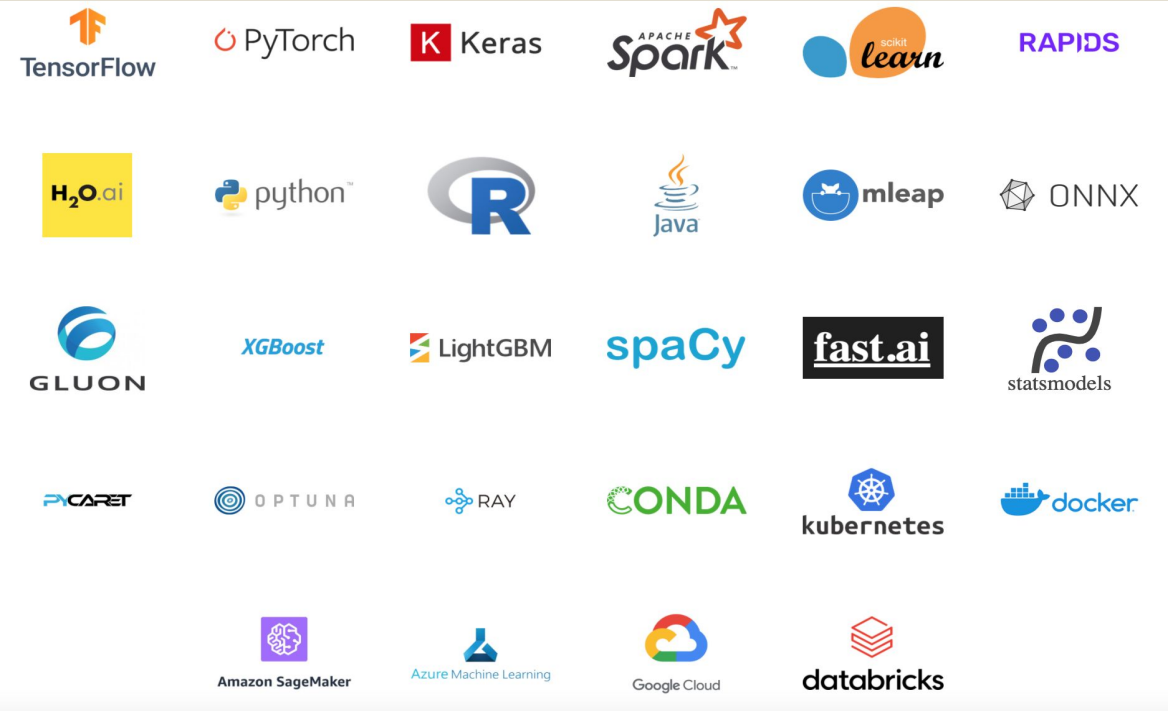
https://mlflow.org/

MLFlow - Componentes



https://mlflow.org/

MLFlow - Integrações



https://mlflow.org/

Bibliografia

- https://cio.com.br/tendencias/o-que-e-dataops-analytics-colaborativo-e-multifuncional/
- https://aws.amazon.com/pt/devops/what-is-devops/
- https://medium.com/data-ops/dataops-is-not-just-devops-for-data-6e03083157b7
- https://aws.amazon.com/pt/devops/what-is-devops/
- https://www.dataopsmanifesto.org/
- https://medium.com/data-ops/dataops-is-not-just-devops-for-data-6e03083157b7
- https://www.aitrends.com/machine-learning/mlops-not-just-ml-business-new-competitive-frontier/
- https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning?hl=pt-br
- https://www.aitrends.com/machine-learning/mlops-not-just-ml-business-new-competitive-frontier/
- https://martinfowler.com/articles/cd4ml.html
- https://hopsworks.readthedocs.io/en/1.1/featurestore/featurestore.html
- https://mlflow.org/docs/latest/tutorials-and-examples/tutorial.html
- https://mlflow.org/

OBRIGADO

Jean Carlos Alves