Fully-Reproducible ML Deployments with Spark, Pachyderm and MLeap

Hollin Wilkins and Daniel Whitenack

Introductions

Hollin Wilkins
Co-Founder, Combust
@combustml

Dan Whitenack
Data Scientist, Pachyderm
@pachydermio

Our Talk in 3 Parts

1. Reproducibility in the Context of ML
2. A Specific ML Use Case
3. Demonstration of Reproducible ML Deployment

Reproducibility in the Context of ML

What is ML Reproducibility?

What is ML Reproducibility?

1. Consistent Results

What is ML Reproducibility?

1. Consistent Results

2. Data Provenance

What is ML Reproducibility?

1. Consistent Results
2. Data Provenance
3. Versioned History

Why should we care?

Why should we care?

1. Collaboration/Creativity

Why should we care?

1. Collaboration/Creativity

2. Compliance

Why should we care?

1. Collaboration/Creativity
2. Compliance
3. Unique Insights

We Propose that...

We Propose that...

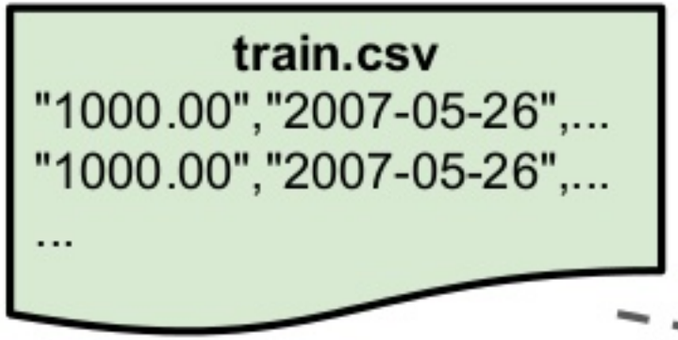
Reproducibility is essential for ML pipelines, such that they can be replayed, modified, tuned and tracked over time.

We Propose that...

Reproducibility is essential for ML pipelines, such that they can be replayed, modified, tuned and tracked over time.

Currently it is difficult to do this with standard tooling.

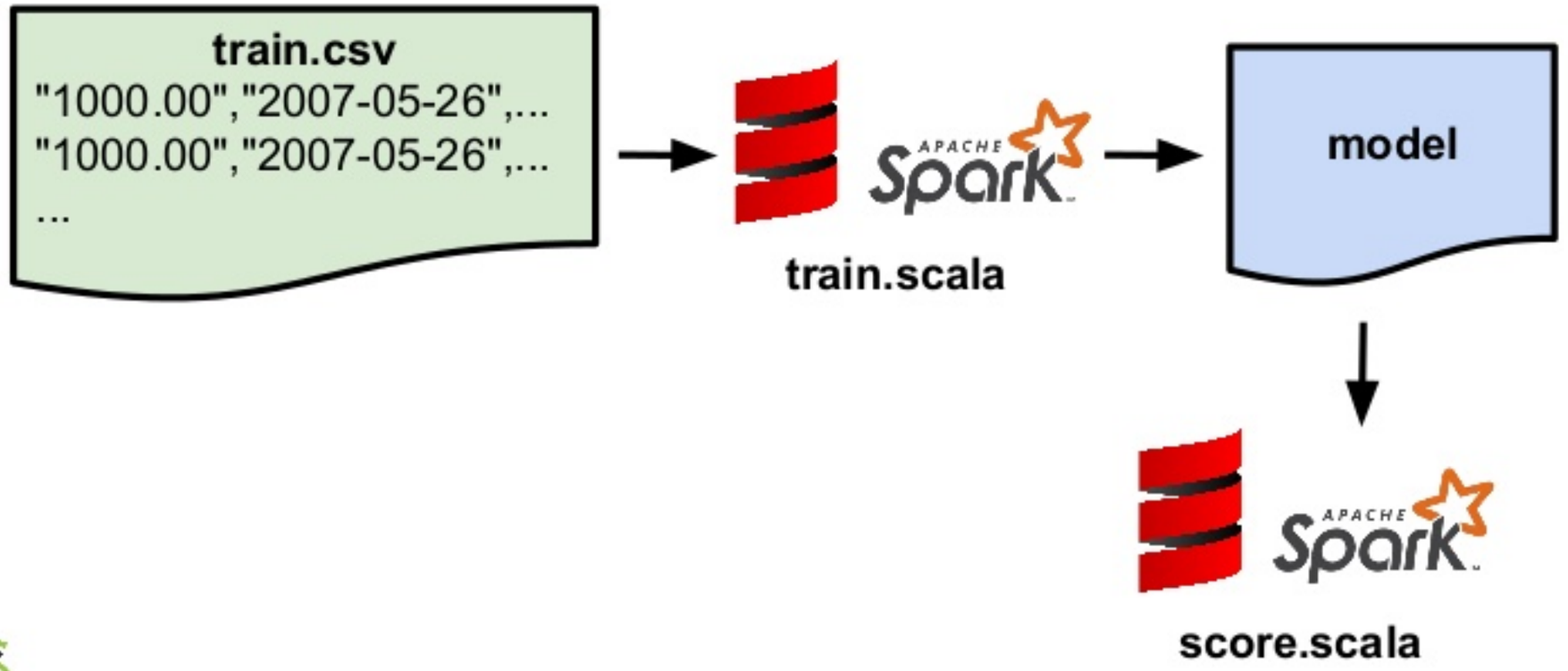
A Specific ML Use Case

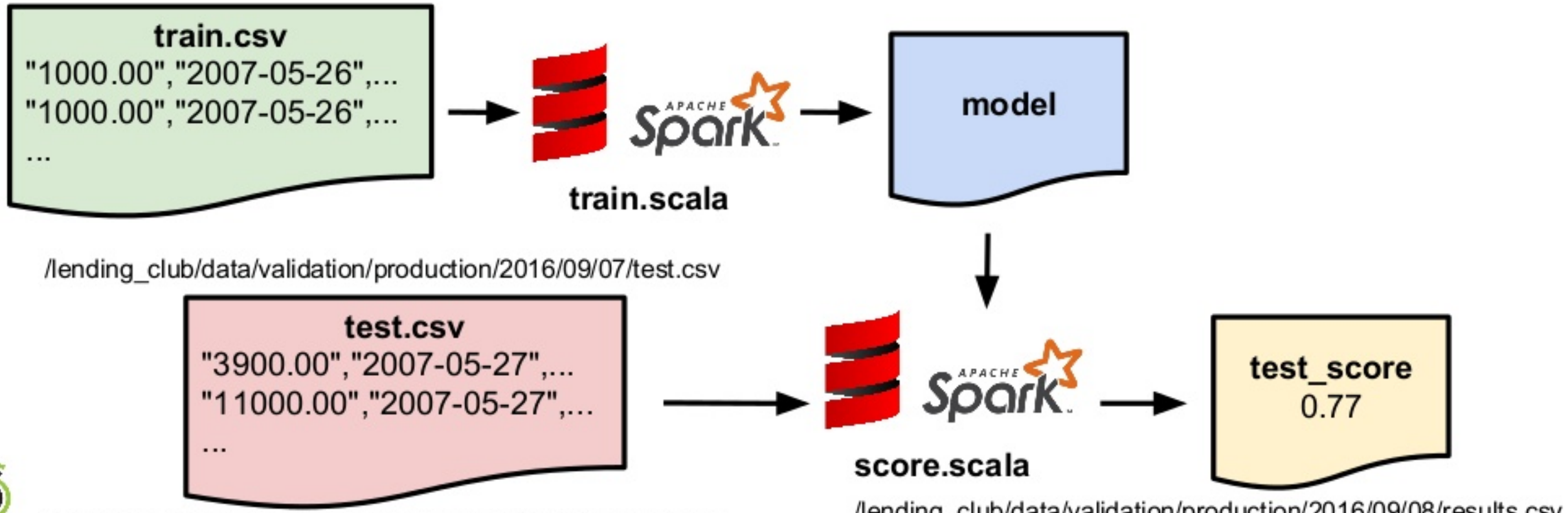


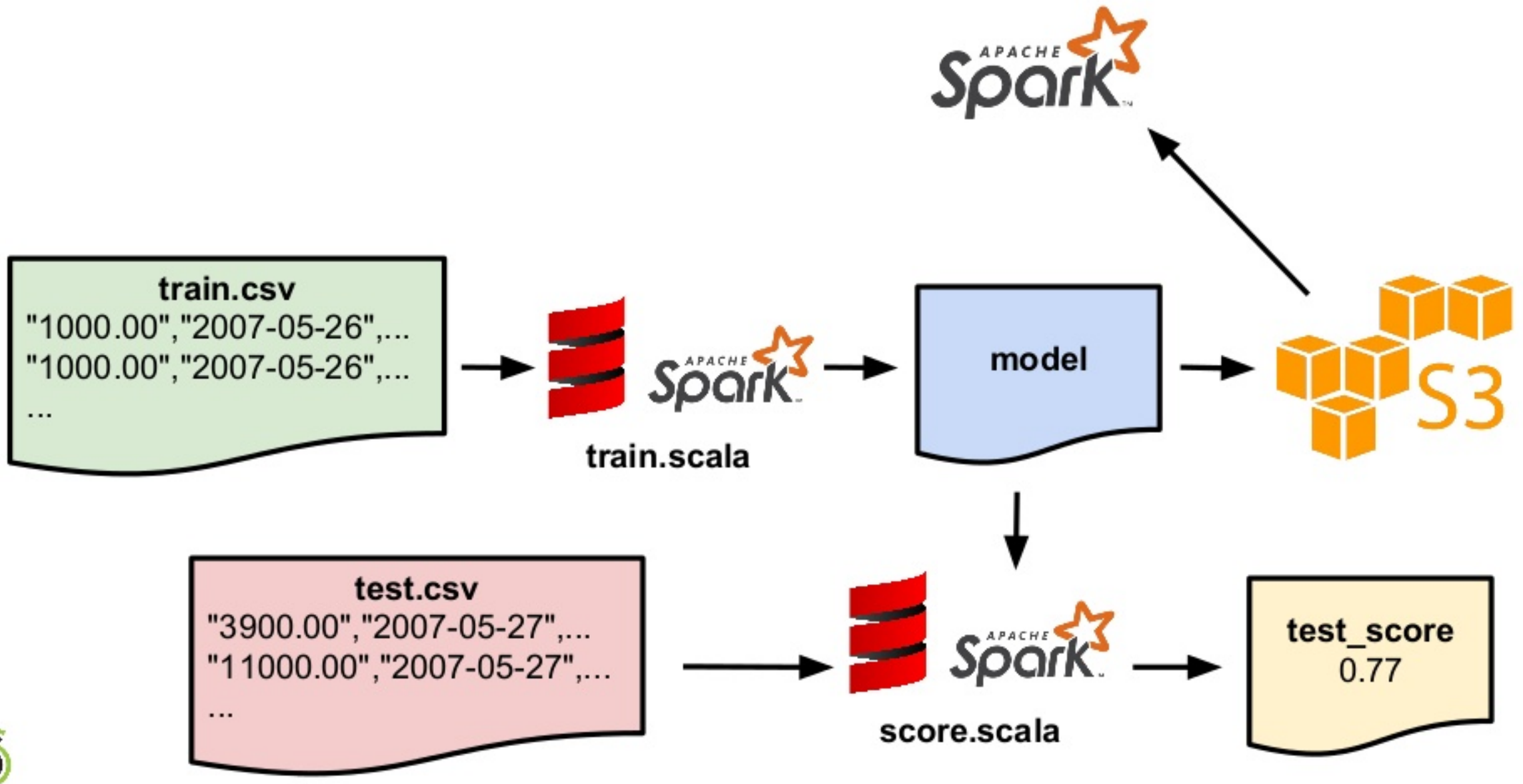
LendingClub

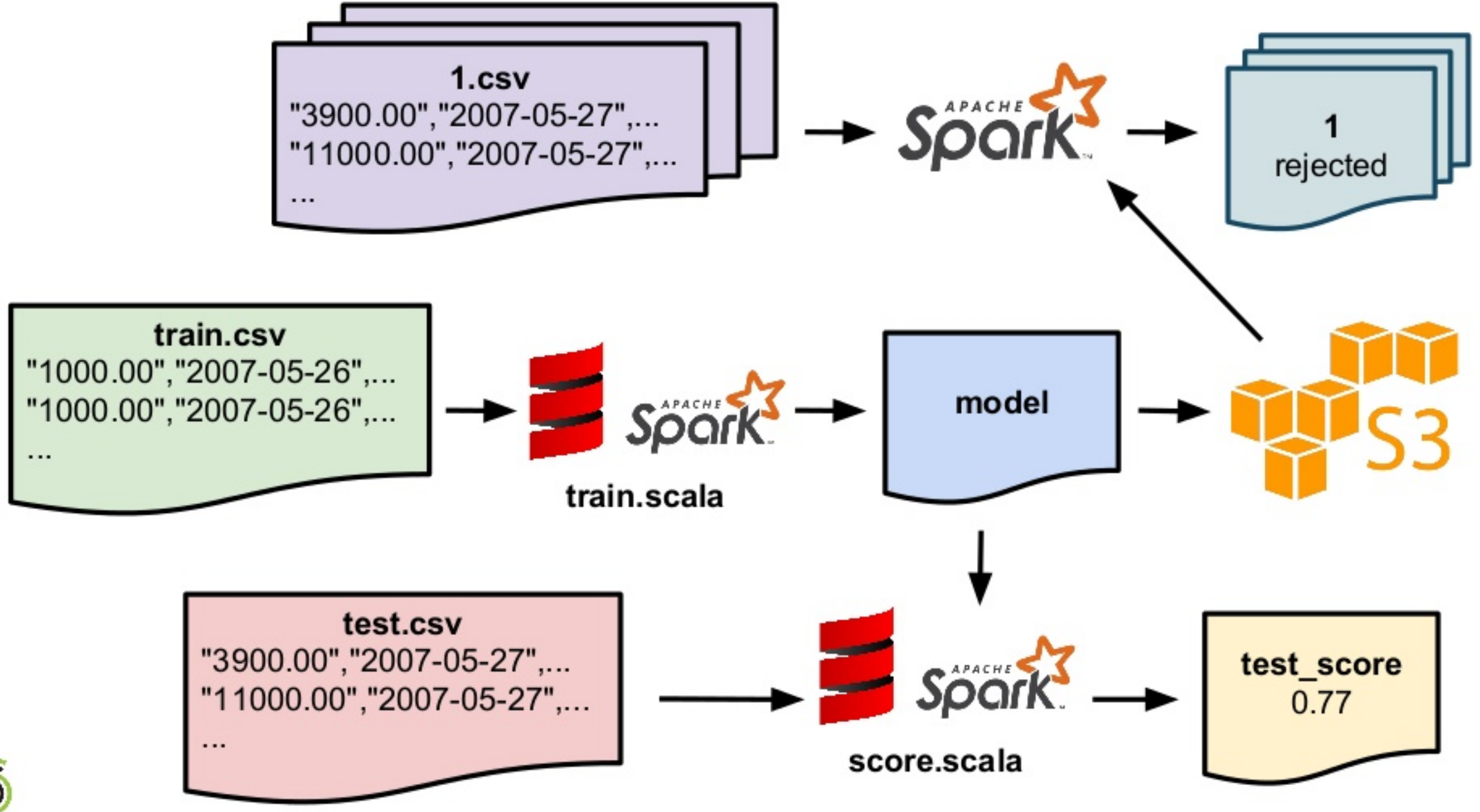
Table	Total Rows	Total Columns	Colums
loan	887383	75	index, id, member_id, loan_amnt, funded_amnt, funded_amnt_inv, term, int_rate, installment, grade, sub_grade, emp_title, emp_length, home_ownership, annual_inc, verification_status, issue_d, loan_status, pymnt_plan, url, desc, purpose, title, zip_code, addr_state, dti, delinq_2yrs, earliest_cr_line, inq_last_6mths, mths_since_last_delinq, mths_since_last_record, open_acc, pub_rec, revol_bal, revol_util, total_acc, initial_list_status, out_prncp, out_prncp_inv, total_pymnt, total_pymnt_inv, total_rec_prncp, total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_d, last_pymnt_amnt, next_pymnt_d, last_credit_pull_d, collections_12_mths_ex_med, mths_since_last_major_derog, policy_code, application_type, annual_inc_joint, dti_joint, verification_status_joint, acc_now_delinq, tot_coll_amt, tot_cur_bal, open_acc_6m, open_il_6m, open_il_12m, open_il_24m, mths_since_rcnt_il, total_bal_il, il_util, open_rv_12m, open_rv_24m, max_bal_bc, all_util, total_rev_hi_lim, inq_fi, total_cu_tl, inq_last_12m

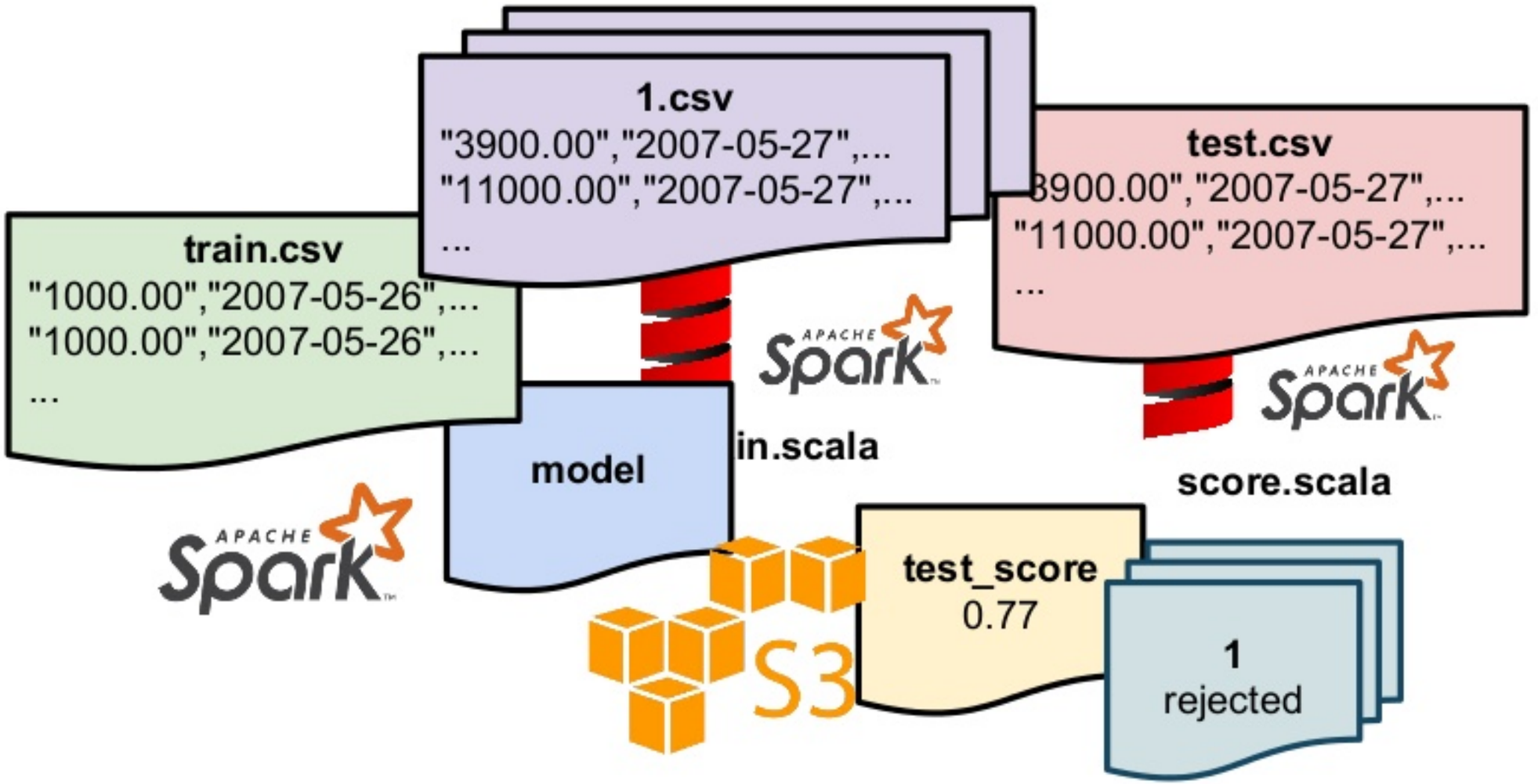

hdfs://PRODUCTION/lending_club/models/2017/05/02/22/13/model/*
train.csv
"1000.00","2007-05-26",...
"1000.00","2007-05-26",...
...
APACHE
Spark
train.scala
model

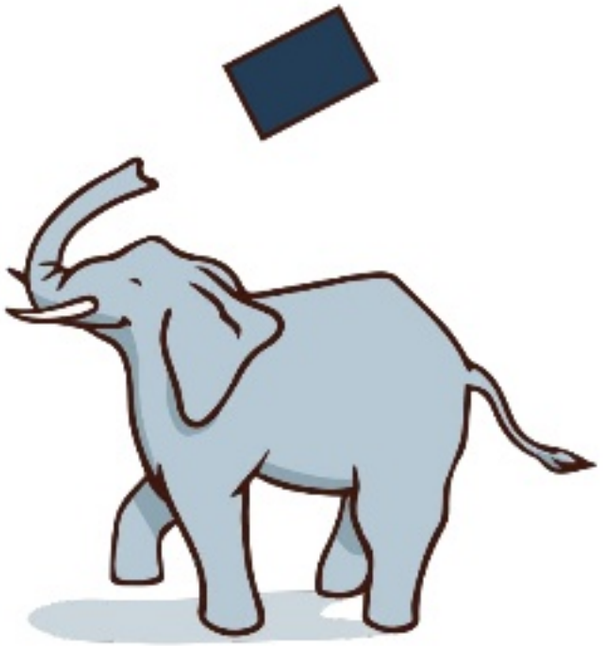

train.csv
"1000.00","2007-05-26",...
"1000.00","2007-05-26",...
...
train.scala
model
score.scala


train.csv
"1000.00","2007-05-26",...
"1000.00","2007-05-26",...
...
train.scala
model
/lending_club/data/validation/production/2016/09/07/test.csv
test.csv
"3900.00","2007-05-27",...
"11000.00","2007-05-27",...
...
score.scala
test_score
0.77
/lending_club/data/validation/production/2016/09/08/test.csv
/lending_club/data/validation/production/2016/09/08/results.csv


APACHE Spark
train.csv
"1000.00","2007-05-26",...
"1000.00","2007-05-26",...
...
APACHE Spark
train.scala
model
S3
test.csv
"3900.00","2007-05-27",...
"11000.00","2007-05-27",...
...
APACHE Spark
score.scala
test_score
0.77


1.csv
"3900.00","2007-05-27",...
"11000.00","2007-05-27",...
...
APACHE Spark
1
rejected
train.csv
"1000.00","2007-05-26",...
"1000.00","2007-05-26",...
...
APACHE Spark
train.scala
model
S3
test.csv
"3900.00","2007-05-27",...
"11000.00","2007-05-27",...
...
APACHE Spark
score.scala
test_score
0.77


1.csv
"3900.00","2007-05-27",...
"11000.00","2007-05-27",...
...
train.csv
"1000.00","2007-05-26",...
"1000.00","2007-05-26",...
...
test.csv
3900.00","2007-05-27",...
"11000.00","2007-05-27",...
...
APACHE Spark
in.scala
model
APACHE Spark
score.scala
APACHE Spark
S3
test_score
0.77
1
rejected

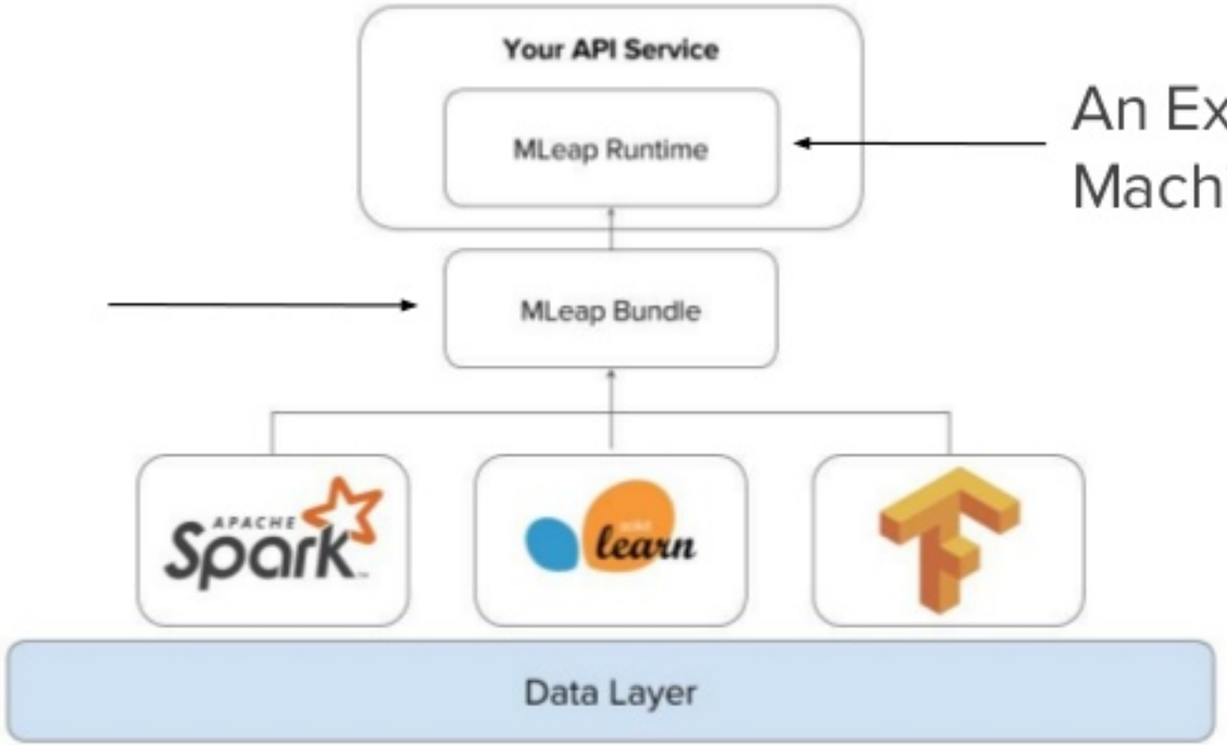


… enter MLeap + Pachyderm

Open source frameworks for reproducible ML deployments, data pipelines, and data versioning

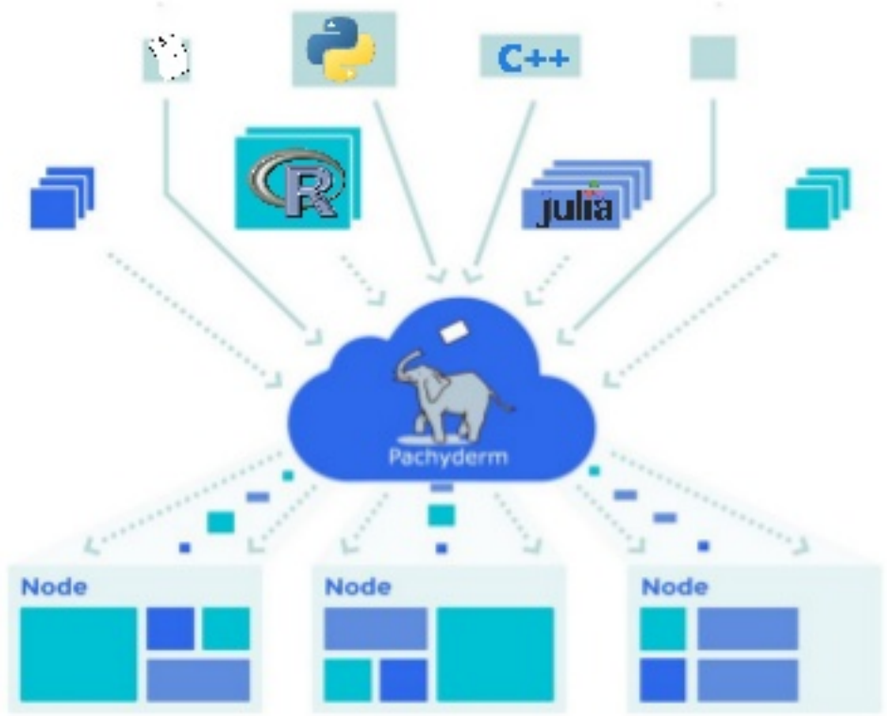
MLeap is...

A Serialization Framework For Machine Learning Pipelines

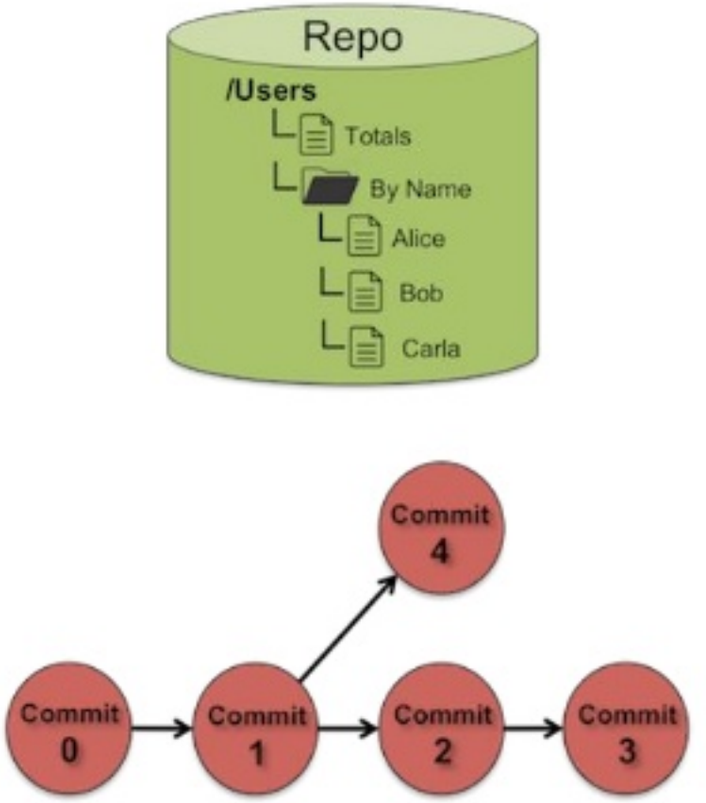


An Execution Engine for Machine Learning Pipelines

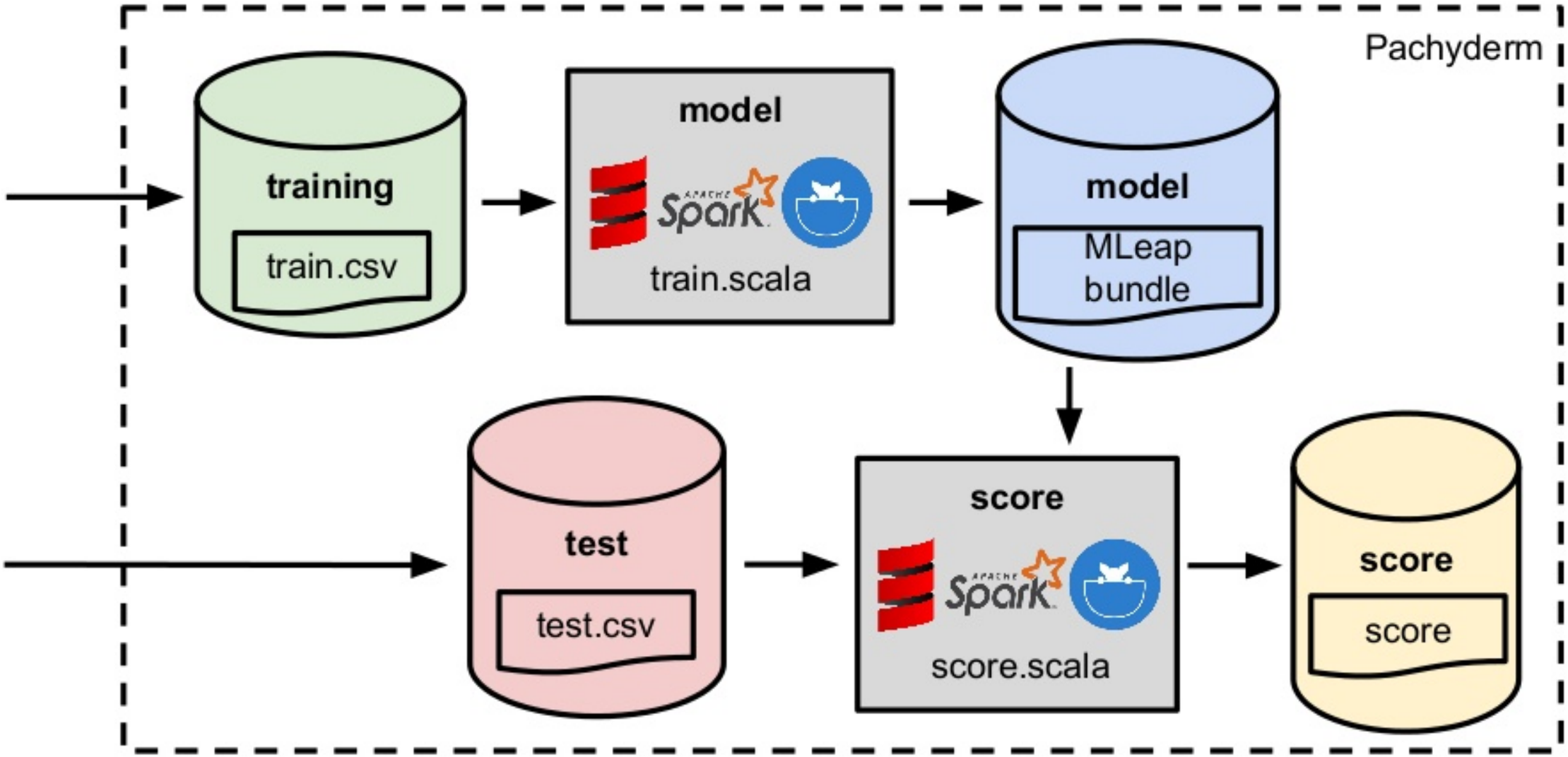
Pachyderm is...

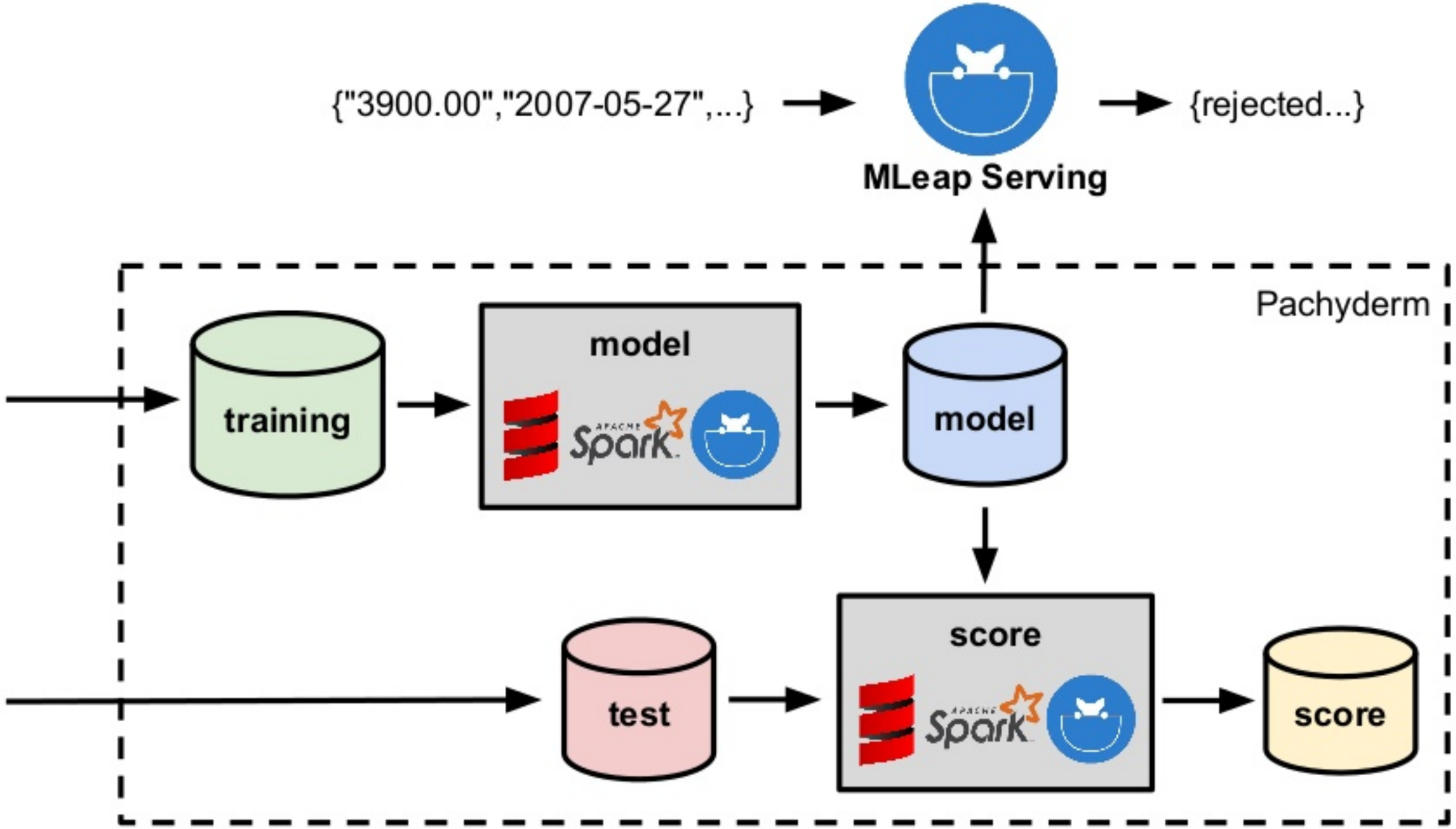


Containerized Data Pipelines



Data Versioning

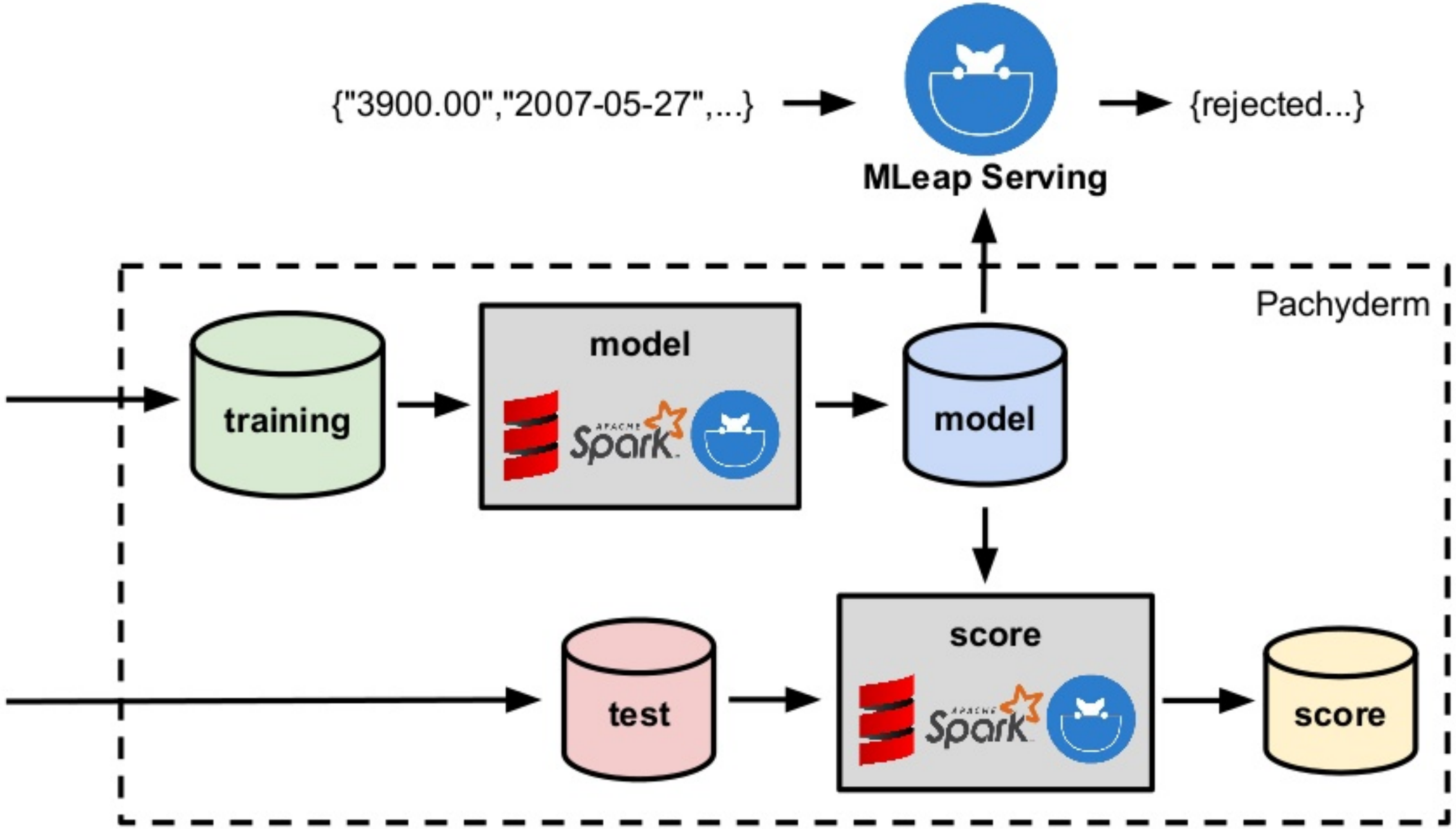

Pachyderm
training
train.csv
model
Spark
train.scala
model
MLeap
bundle
test
test.csv
score
Spark
score.scala
score
score


{"3900.00","2007-05-27",...}
MLeap Serving
{rejected...}
Pachyderm
training
model
Spark
model
test
score
Spark
score

Existing Solutions, Comparison

	Plain Spark	Prediction.io	Data Robot	Model DB	Pachyderm + MLeap
Data Versioning	✗	✗	✗	✗	✓
Model Versioning	✗	✓	✓	✓	✓
Open Sourced	✓	✓	✗	✓	✓
Works with ML Pipelines	✓	⊖	✗	✓	✓
Commercial Support	✓	✗	✓	✗	✓

Demonstration of Reproducible ML Deployment


{"3900.00","2007-05-27",...}
MLeap Serving
{rejected...}
Pachyderm
training
model
Spark
model
test
score
Spark
score

Git Repositories

Pachyderm: https://github.com/pachyderm/pachyderm

MLeap: https://github.com/combust/mleap

Demo: https://github.com/combust/pachyderm-mleap-demo

Thank You.

Hollin Wilkins
Combust, @combustml, combust.ml

Daniel Whitenack
Pachyderm, @pachydermIO, pachyderm.io