KubeCon | CloudNativeCon

China 2018

Introduction to Modern Data Science

Sam Kreter

Code and Slides

github.com/samkreter/KubeconAsia2018

The Process

KubeCon

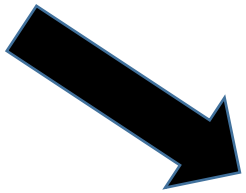
CloudNativeCon

China 2018

Business Need /
Problem Discovery

The Process

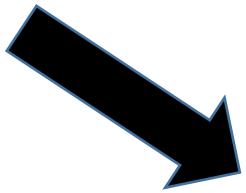
Business Need /
Problem Discovery



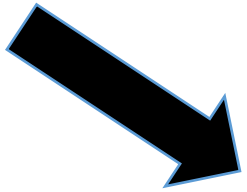
Development

The Process

Business Need /
Problem Discovery



Development



Production /
Actual User Impact

The Process

Business Need /
Problem Discovery

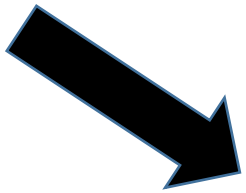
Production /
Actual User Impact

DOMINO

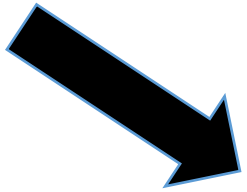
Pandas

The Process

Business Need /
Problem Discovery

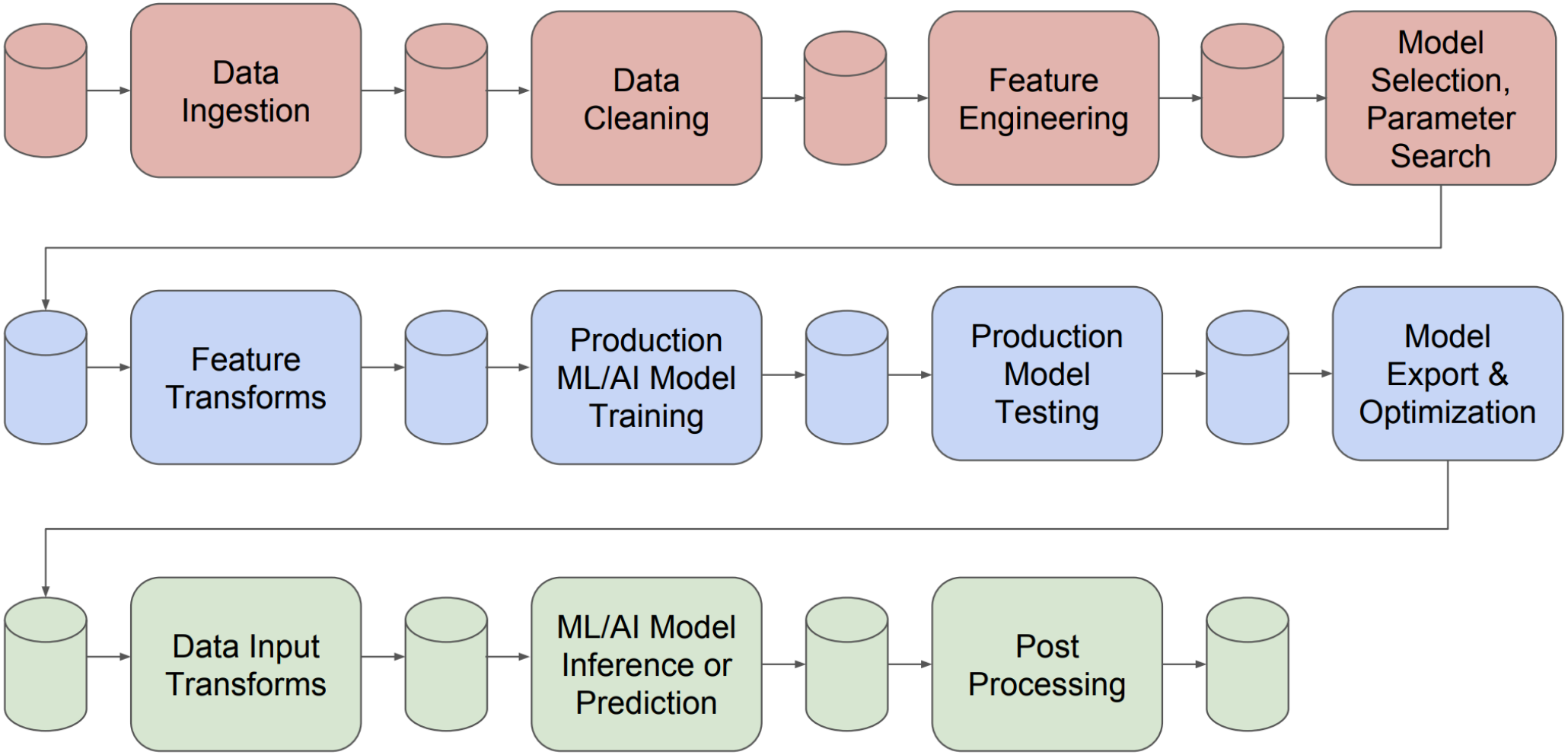


Development

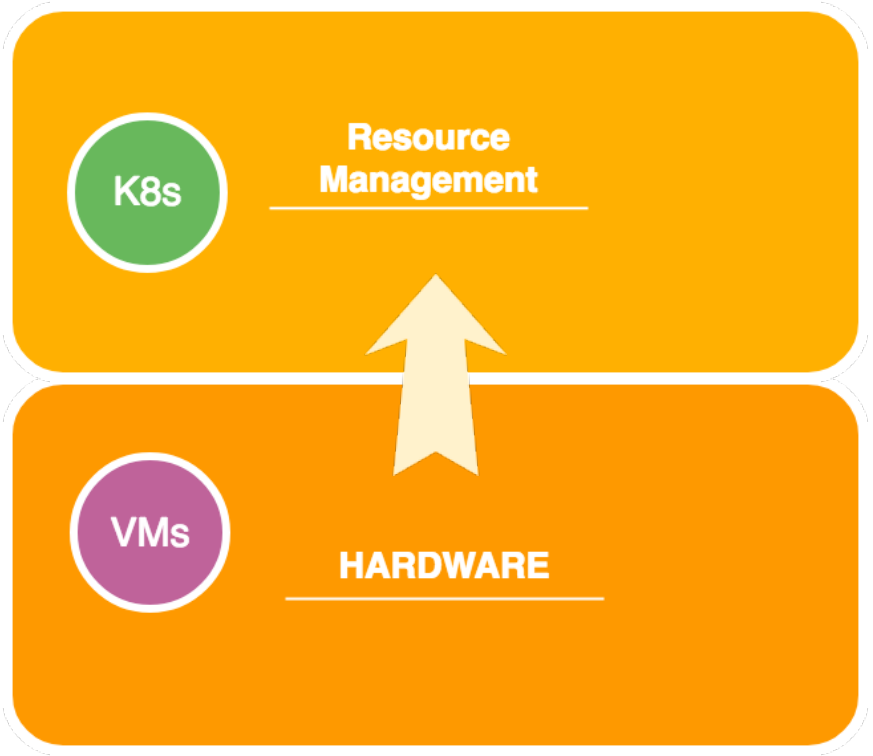


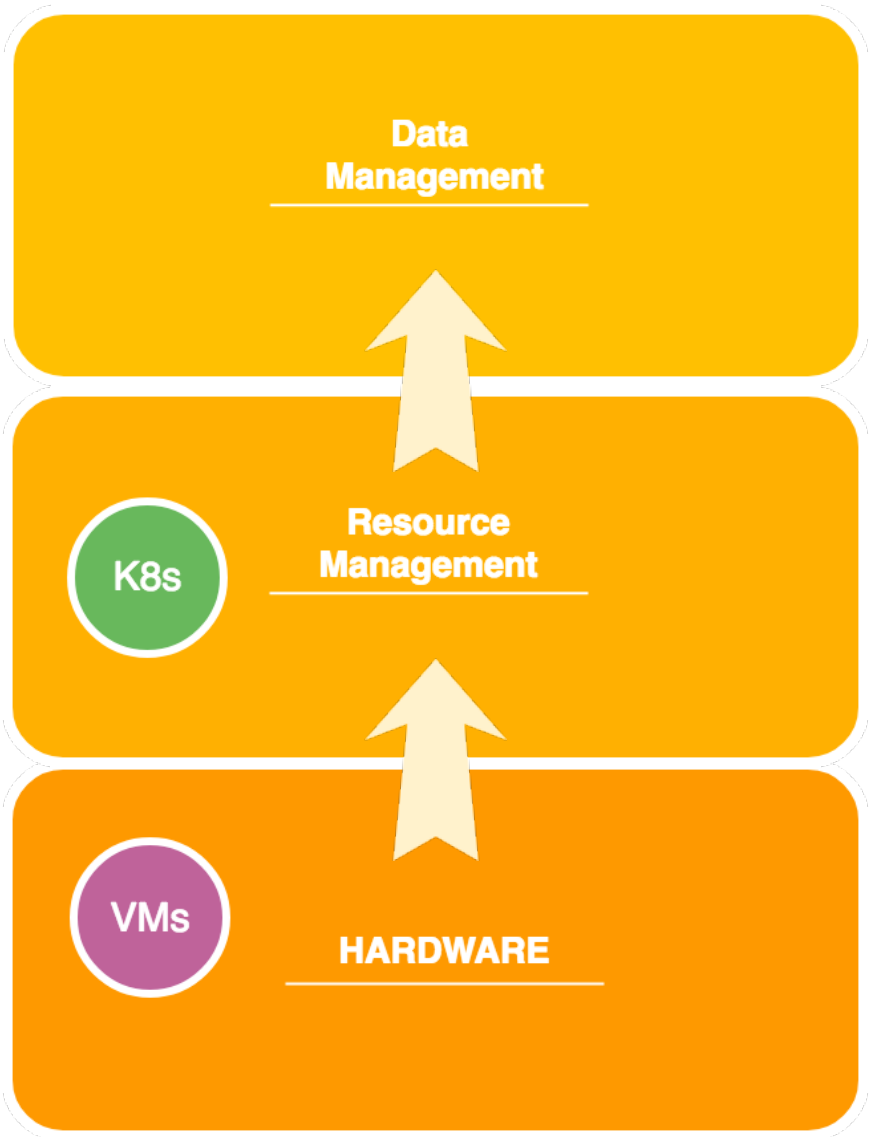
Production /
Actual User Impact

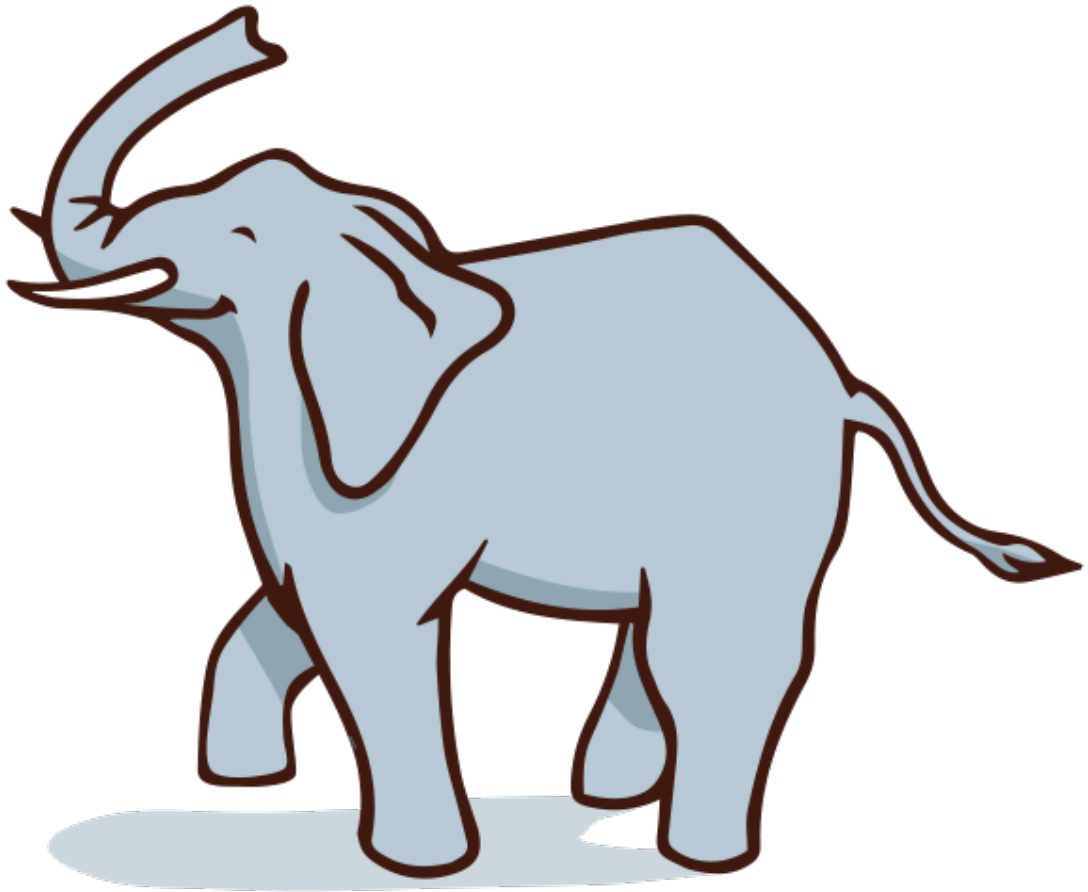
The Data Science Pipeline




VMs
HARDWARE


K8s
Resource Management
VMs
HARDWARE


Data Management
K8s
Resource Management
VMs
HARDWARE

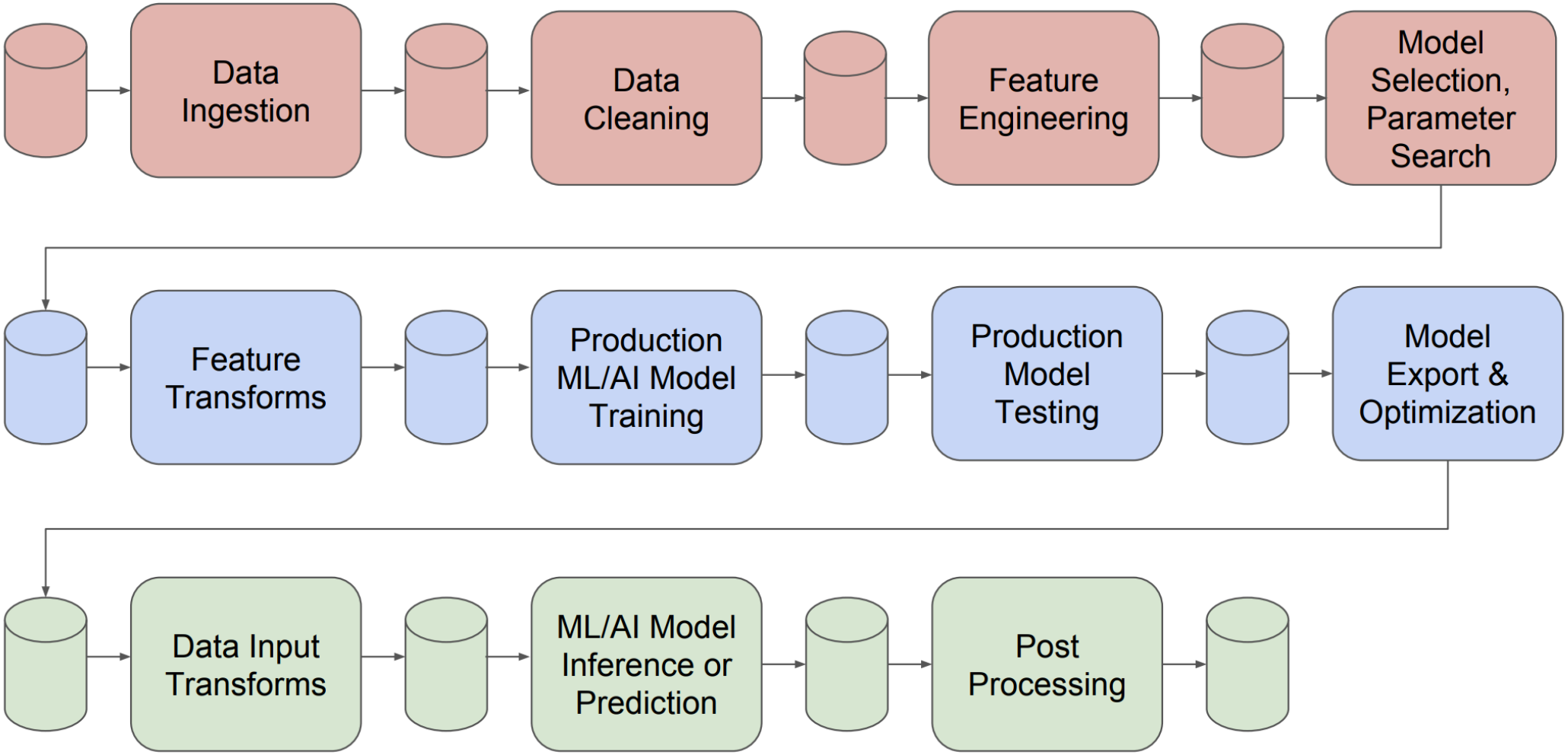


Pachyderm

Pipelines


```
{
  "pipeline": {
    "name": "wordcount"
  },
  "transform": {
    "image": "wordcount-image",
    "cmd": ["/binary", "/pfs/data", "/pfs/out"]
  },
  "parallelism_spec": {
    "coefficient": 2
  },
  "input": {
      "atom": {
        "repo": "data",
        "glob": "/*"
      }
  }
}
```

The Data Science Pipeline



Principles

KubeCon

CloudNativeCon

China 2018

1. Autonomy


I HAVE A VERY PARTICULAR SET OF SKILLS

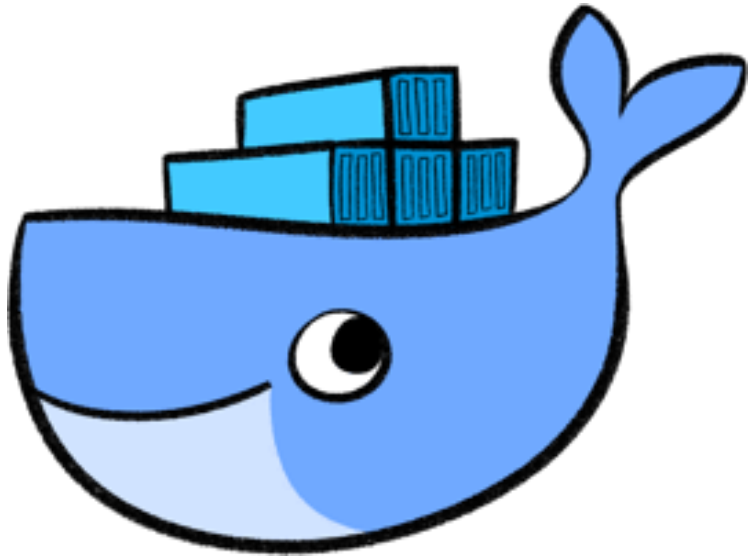
KubeCon

CloudNativeCon

China 2018

1. Autonomy

Autonomy



Rocket

cri-o

Containerization

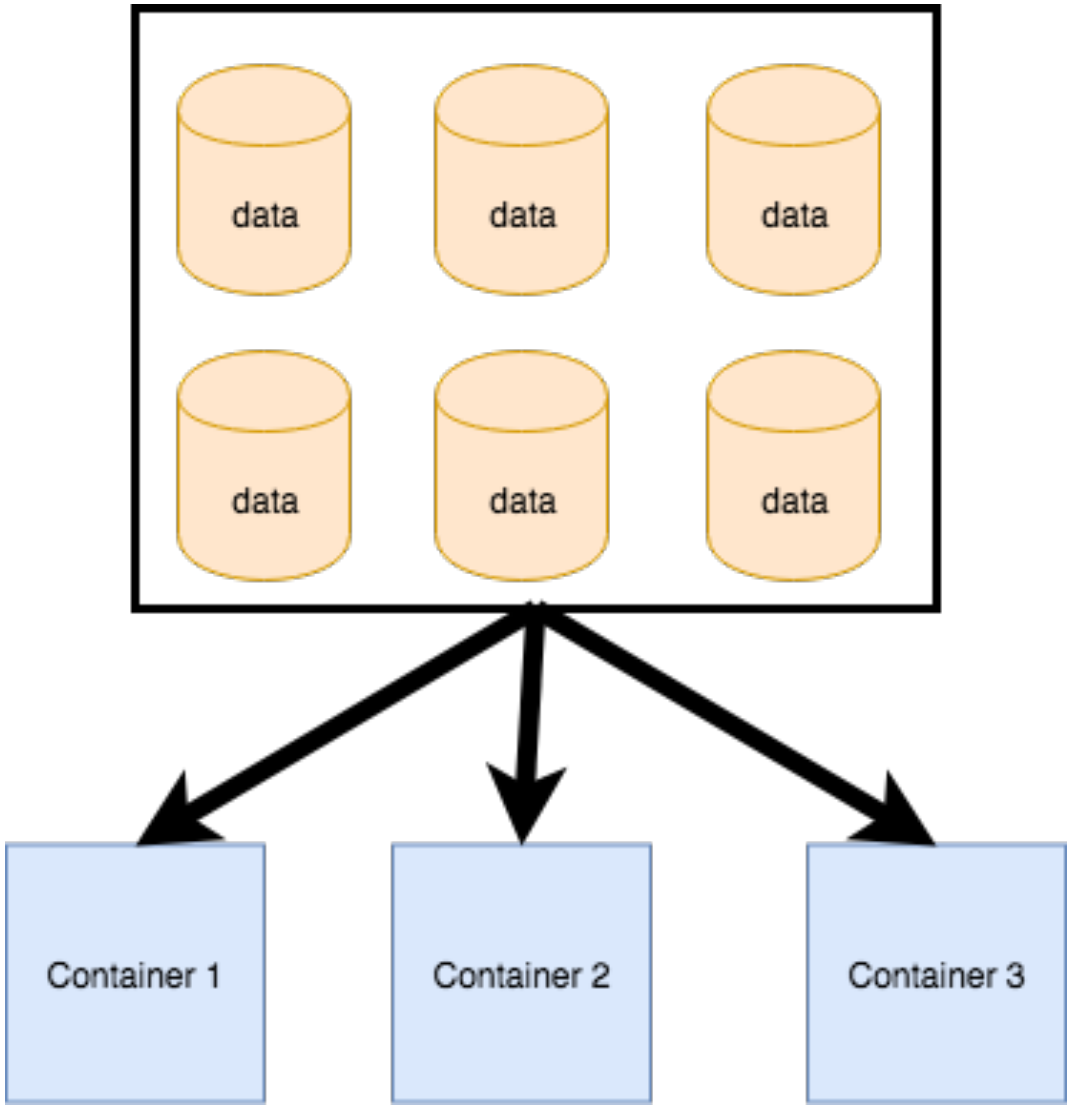
1. Single Operations per Container

Containerization

1. Single Operations per Container

2. Use Parameterize Data Flow
   - Data Inputs
   - Data Outputs

Distributing Workloads

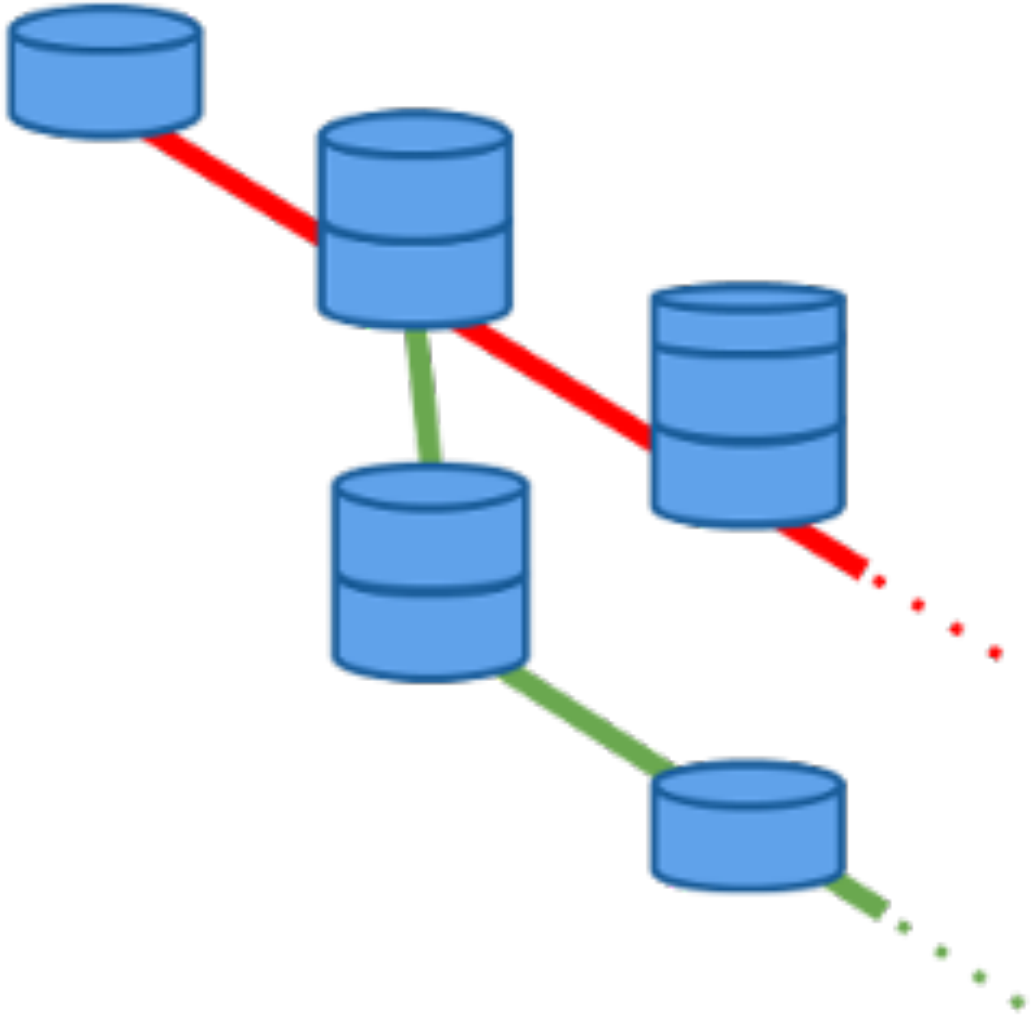


KubeCon | CloudNativeCon

China 2018

2. Reproducibility

Data Versioning



Reproducibility

For Developers

Reproducibility

For Developers

For the Team

Reproducibility

For Developers

For the Team

For Production

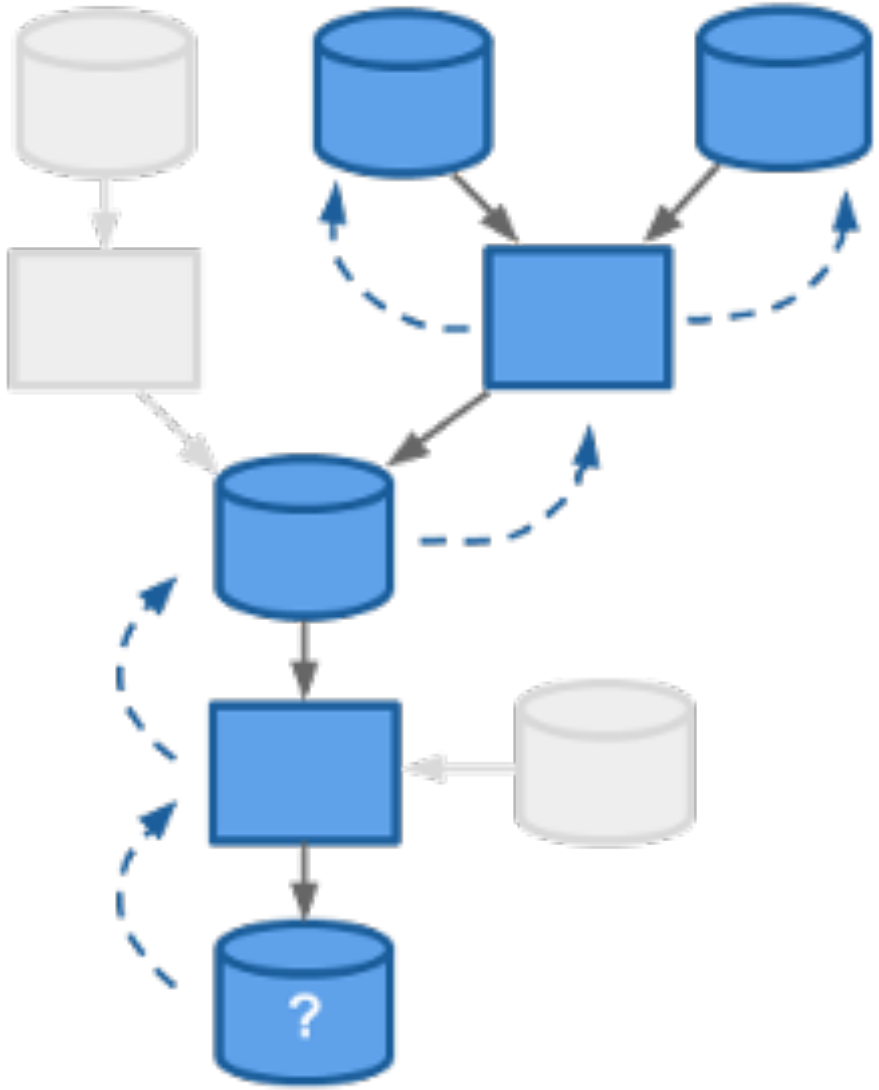
KubeCon

CloudNativeCon

China 2018

3. Data Provenance

Data Provenance



Data Provenance

1. Retry after failure

Data Provenance

1. Retry after failure

2. Rolling Back Models

Data Provenance

1. Retry after failure
2. Rolling Back Models
3. Clarity / Organizational Trust

KubeCon

CloudNativeCon

China 2018

4. Automation (CI/CD)

Azure DevOps

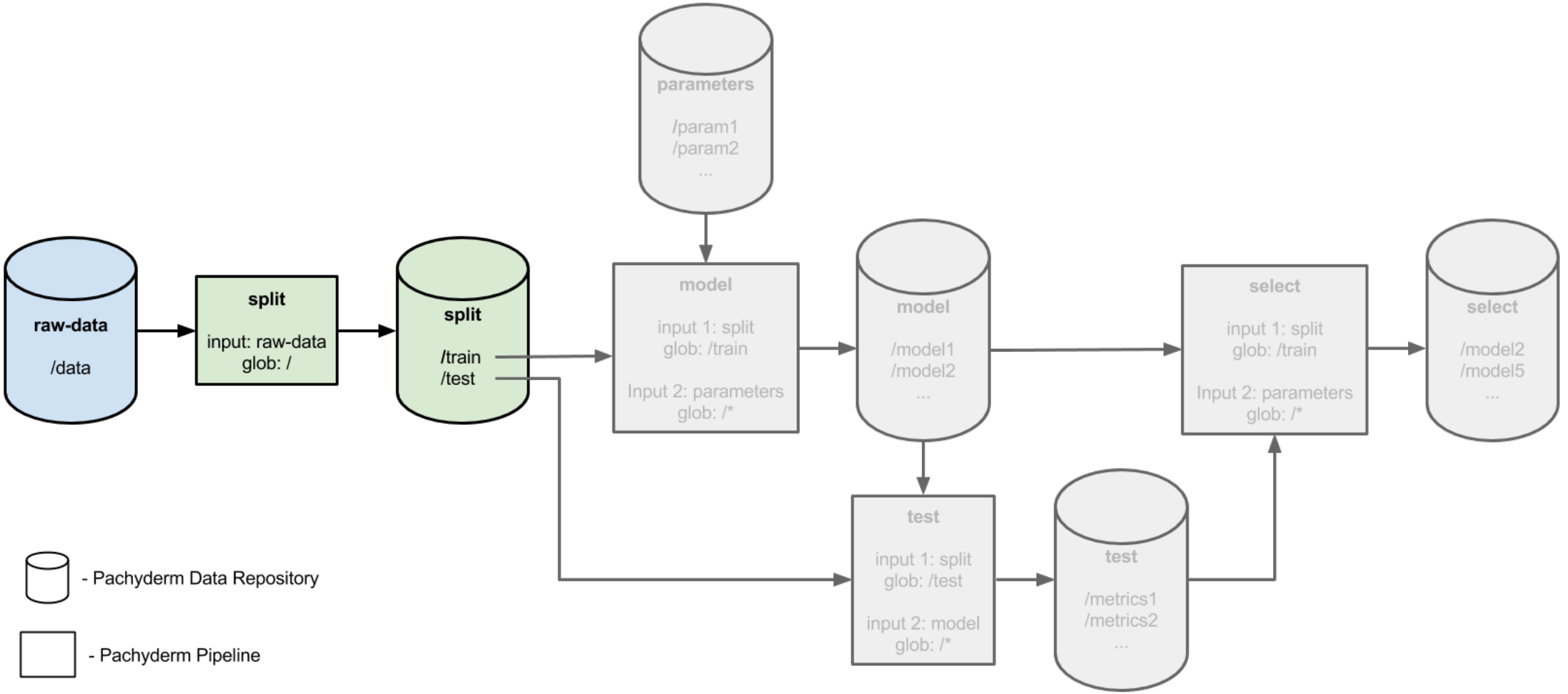




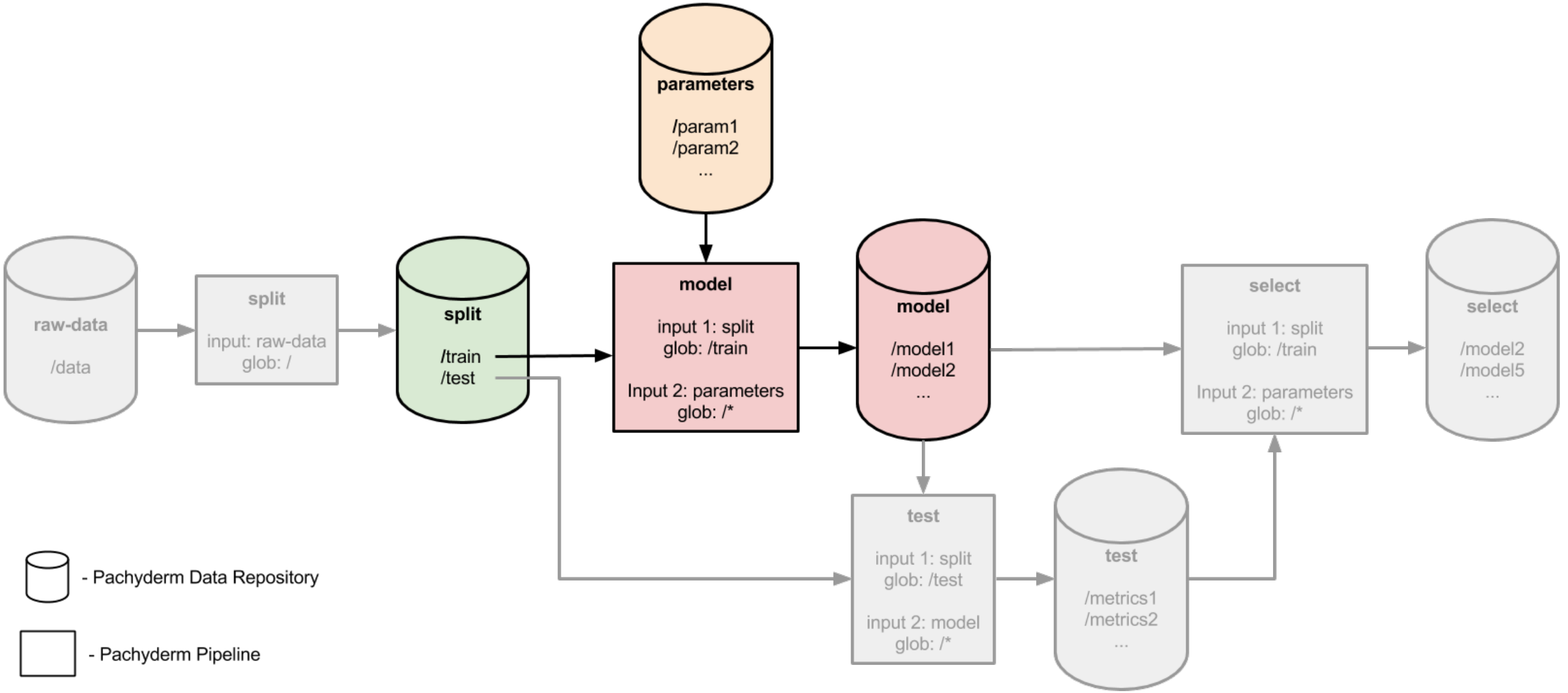
Summary

1. Autonomy
2. Reproducibility
3. Data Provenance
4. Automation

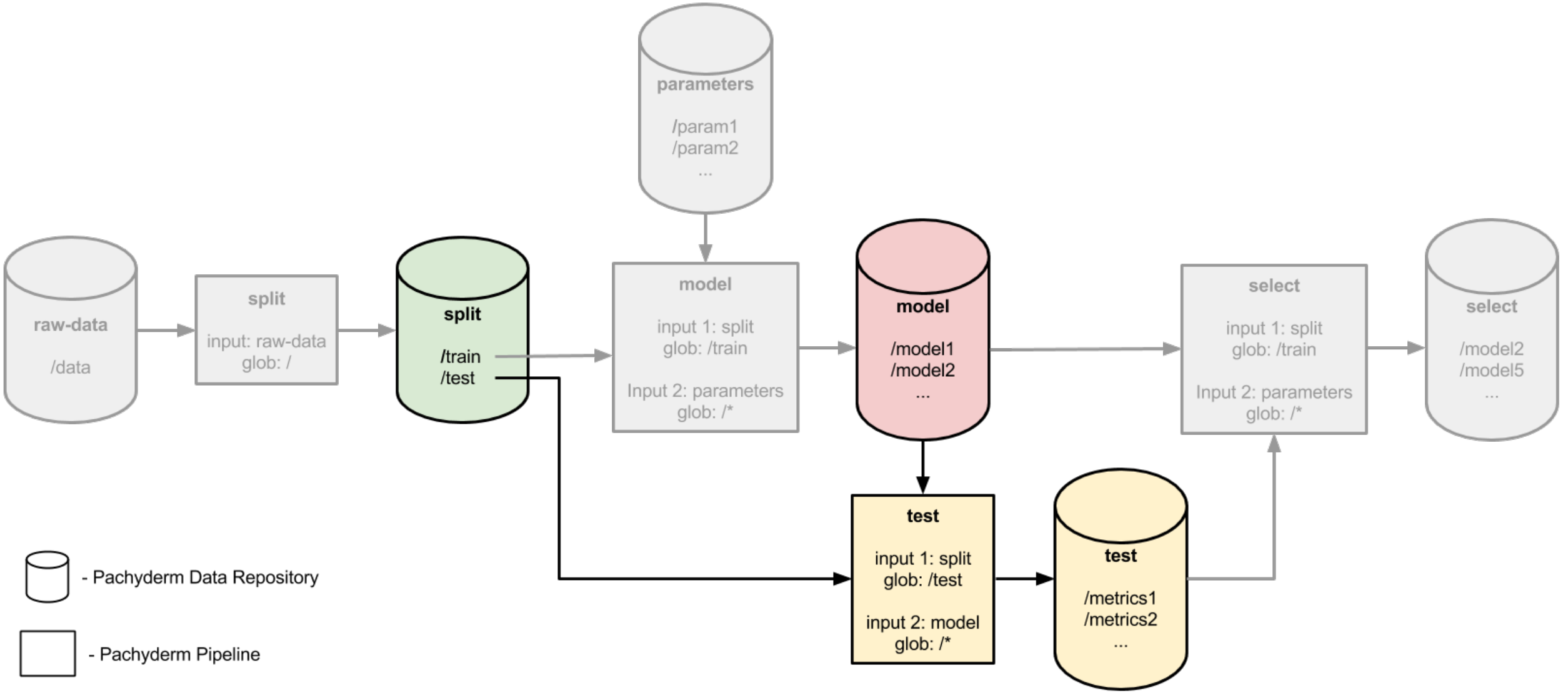
Demo


parameters
/param1
/param2
...
raw-data
/data
split
input: raw-data
glob: /
split
/train
/test
model
input 1: split
glob: /train
Input 2: parameters
glob: /*
model
/model1
/model2
...
select
input 1: split
glob: /train
Input 2: parameters
glob: /*
select
/model2
/model5
...
test
input 1: split
glob: /test
input 2: model
glob: /*
test
/metrics1
/metrics2
...
- Pachyderm Data Repository
- Pachyderm Pipeline

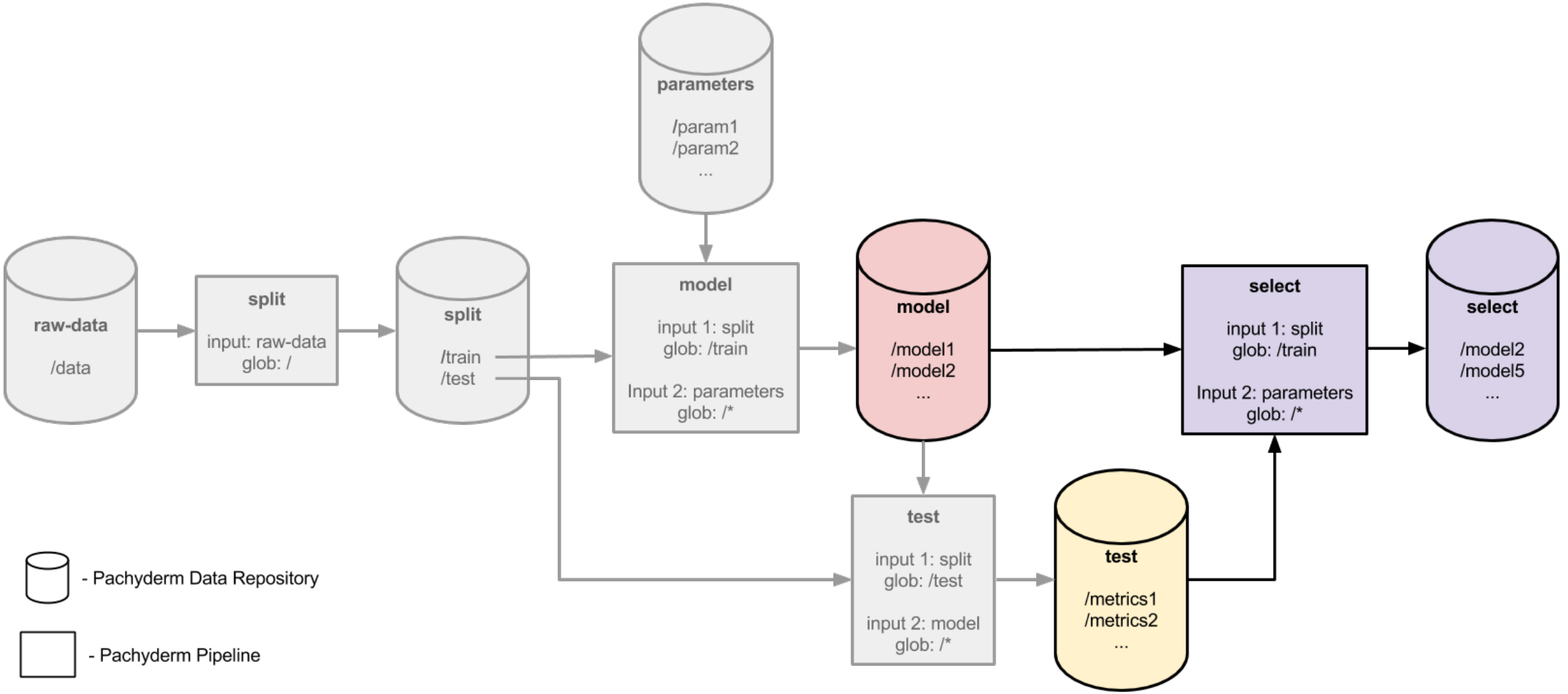
Demo


parameters
/param1
/param2
...
raw-data
/data
split
input: raw-data
glob: /
split
/train
/test
model
input 1: split
glob: /train
Input 2: parameters
glob: /*
model
/model1
/model2
...
select
input 1: split
glob: /train
Input 2: parameters
glob: /*
select
/model2
/model5
...
test
input 1: split
glob: /test
input 2: model
glob: /*
test
/metrics1
/metrics2
...
- Pachyderm Data Repository
- Pachyderm Pipeline

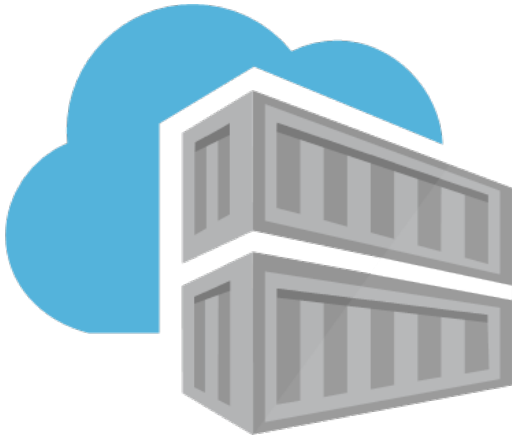
Demo

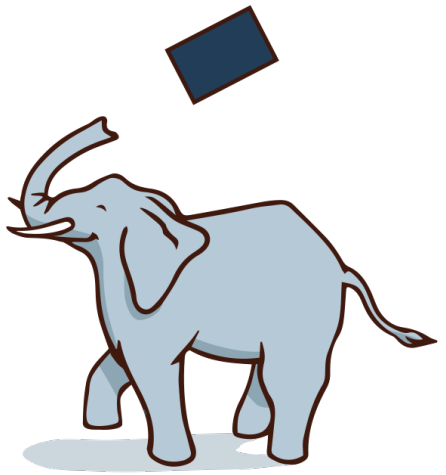

parameters
/param1
/param2
...
raw-data
/data
split
input: raw-data
glob: /
split
/train
/test
model
input 1: split
glob: /train
Input 2: parameters
glob: /*
model
/model1
/model2
...
select
input 1: split
glob: /train
Input 2: parameters
glob: /*
select
/model2
/model5
...
test
input 1: split
glob: /test
input 2: model
glob: /*
test
/metrics1
/metrics2
...
- Pachyderm Data Repository
- Pachyderm Pipeline

Demo


parameters
/param1
/param2
...
raw-data
/data
split
input: raw-data
glob: /
split
/train
/test
model
input 1: split
glob: /train
Input 2: parameters
glob: /*
model
/model1
/model2
...
select
input 1: split
glob: /train
Input 2: parameters
glob: /*
select
/model2
/model5
...
test
input 1: split
glob: /test
input 2: model
glob: /*
test
/metrics1
/metrics2
...
- Pachyderm Data Repository
- Pachyderm Pipeline

Demo





Pachyderm

kubernetes

Contact Me

Twitter: @samkreter

Github: samkreter

Medium: samkreter

Sources / Resources

Pachyderm: http://www.pachyderm.io/

Data Science Bill of Rights: http://www.pachyderm.io/dsbor.html

Azure Container Registry Build: https://docs.microsoft.com/en-us/azure/container-registry/container-registry-tasks-overview

Azure Kubernetes Service: https://azure.microsoft.com/en-us/services/kubernetes-service/

Pipeline Images: https://github.com/pachyderm/pachyderm/tree/master/doc/examples/ml/hyperparameter

KubeCon

CloudNativeCon

China 2018

谢谢！

KubeCon | CloudNativeCon

China 2018