DATA+AI SUMMIT 2022

Build Metadata and Lineage Driven Pipelines in Kubernetes

Yi-Hong Wang

Software Developer, IBM

Tommy Li

Software Developer, IBM

ORGANIZED BY databricks

Agenda

- Introduce Kubeflow Pipelines
- Metadata in Kubeflow Pipelines
- Benefits of Metadata and lineage driven pipeline
- Metadata enhancement in Kubeflow Pipelines v2
- Abstraction Layer to support agnostic backends
- Summary

Enterprises are still struggling to scale AI beyond experimentation

88% of corporate AI initiatives are struggling to move beyond test stages

Source: Artificial Intelligence, The Next Digital Frontier. McKinsey Global Institute, 2017

“I have no quantification of the business impact of my AI solutions”

Client Quotables

“My data scientists have developed some models, but I do not know if they always achieve the best possible solution”

“I have an analytics team that has executed multiple PoCs, but none of that has made it into production”

“We’ve deployed multiple algorithms, but we have not seen any improvement in our business KPIs”

“We find it difficult finding and hiring the right AI talent”

“My business users **do not trust** the results of my AI applications, and they do not get used”

How to deliver AI at Scale…in Production

Best practices for building accurate models are well understood...

… but less so for building productive Data Science solution at scale.

Holistic Architecture
- Application Logic
- Technical Integration
- Model Management
- Tracing, Logging, Metrics

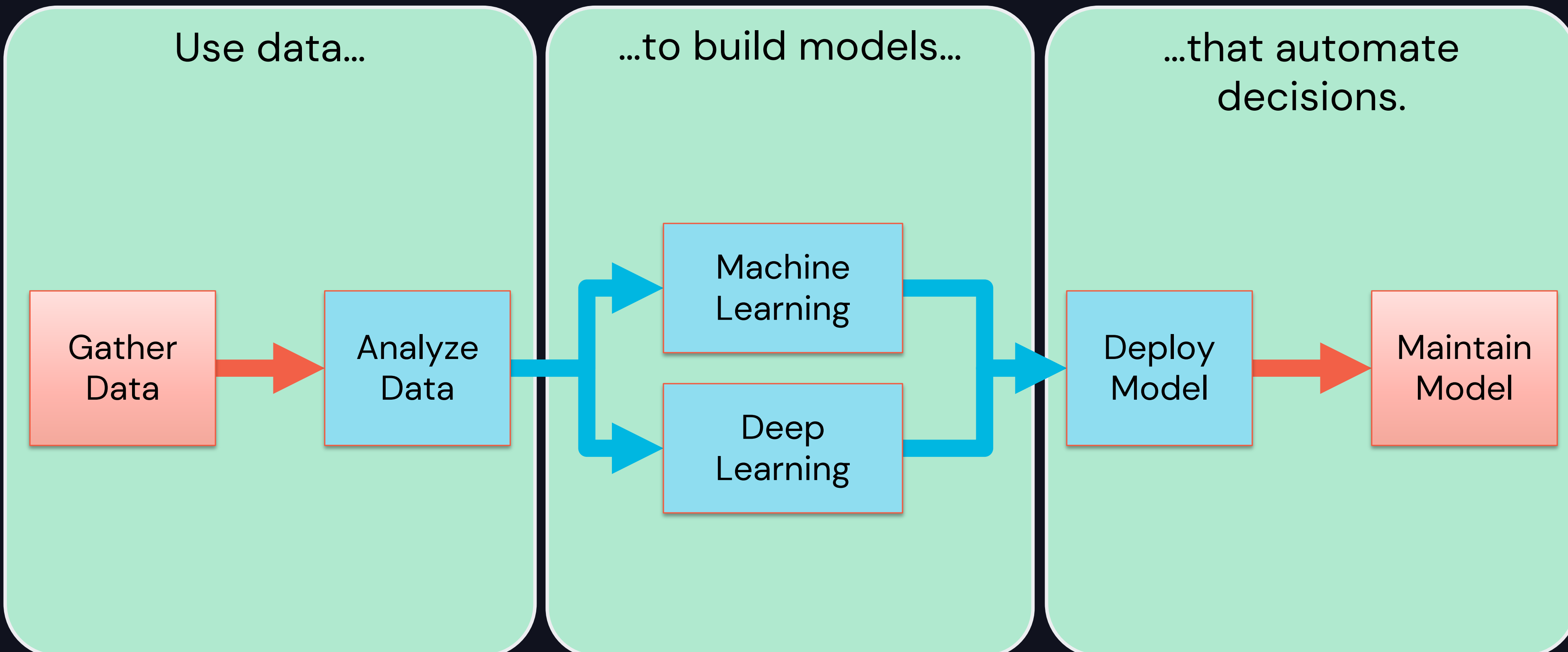
Effective Engineering
- Standards
- Pipelines
- Automation

Smooth Operations
- Technical Monitoring
- Model Monitoring
- Maintenance Strategy

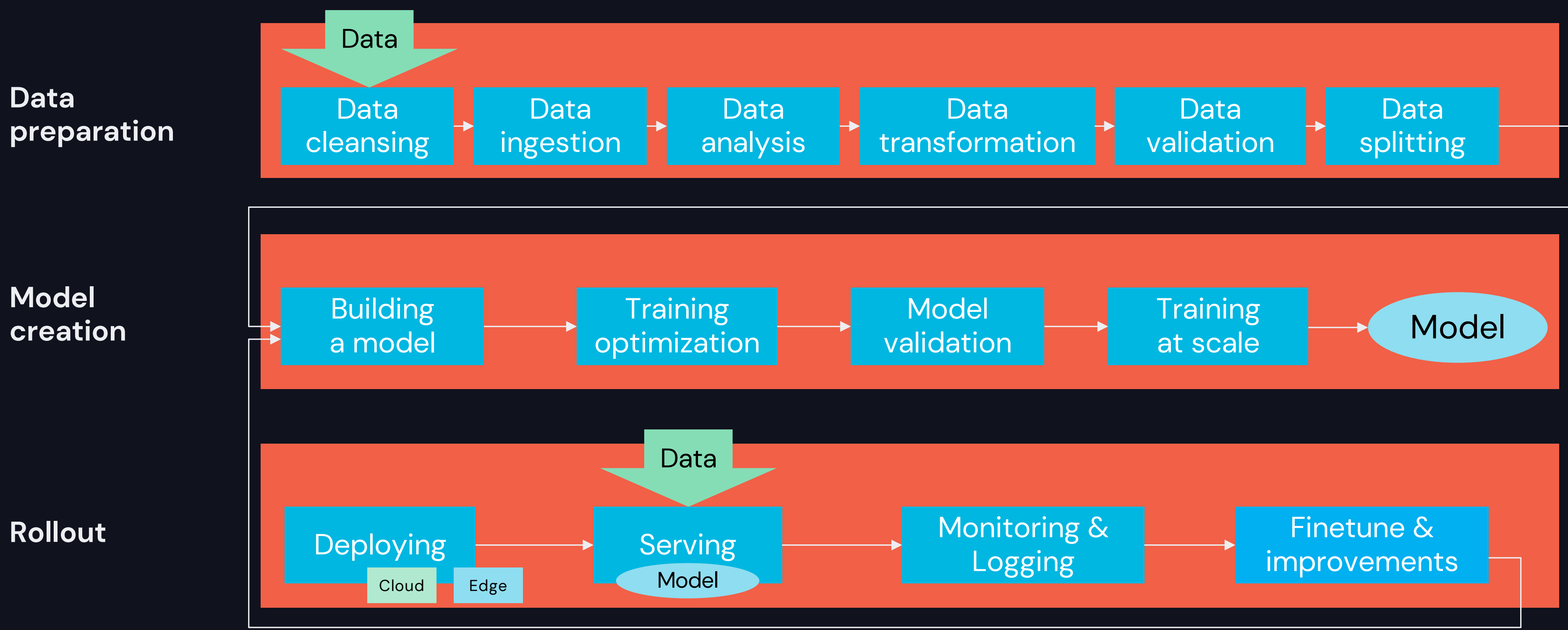
High-Performing Team

Targeted Project Approach

Pillars of AI Lifecycle – Datasets, Models…



…and Pipelines



Kubeflow Pipelines

Containerized implementations of ML Tasks

- Pre-built components: Just provide params or code snippets (e.g. training code)
- Create your own components from code or libraries
- Use any runtime, framework, data types
- Attach k8s objects - volumes, secrets

Specification of the sequence of steps

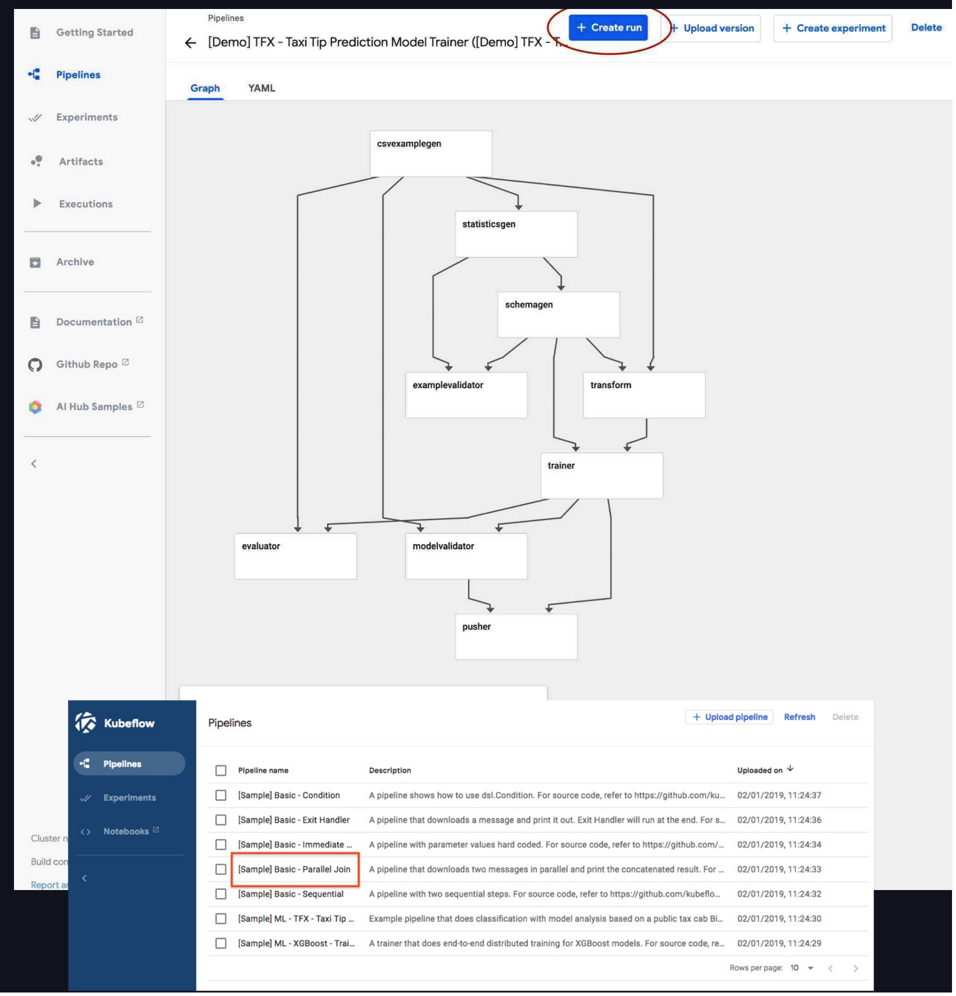
- Specified via Python DSL
- Inferred from data dependencies on input/output

Input Parameters

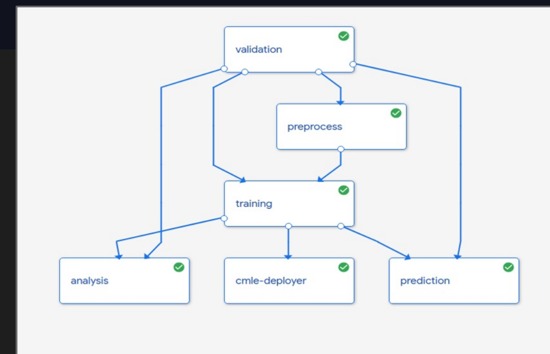
- A “Run” = Pipeline invoked w/ specific parameters
- Can be cloned with different parameters

Schedules

- Invoke a single run or create a recurring scheduled pipeline



Define Pipeline with Python SDK

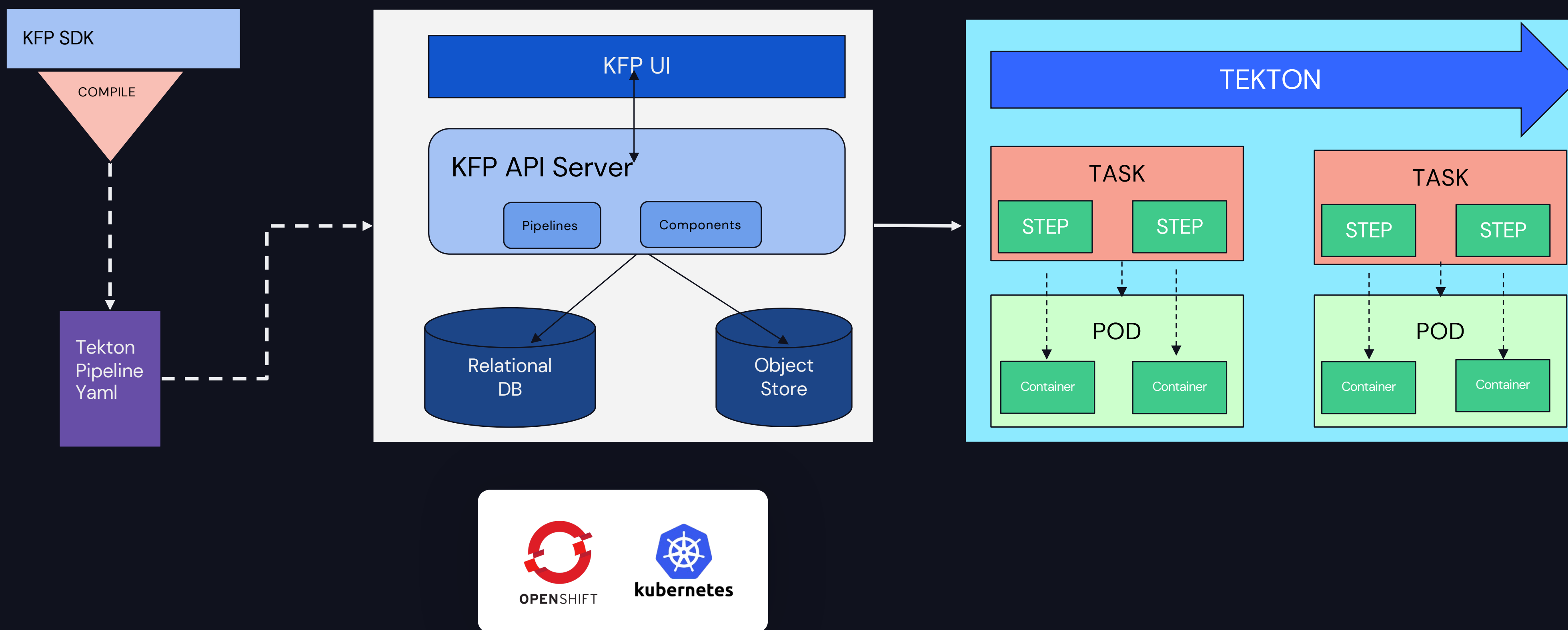



```
@dsl.pipeline(name='taxi-cab-classification-pipeline-example')
def taxi_cab_classification(
        output_dir: str, project: str,
        train_data: str = 'gs://bucket/train.csv',
        evaluation_data: str = 'gs://bucket/eval.csv',
        target: str = 'tips',
        learning_rate: int = 0.1, hidden_layer_size: str = '100,50', steps: int = 3000):

    tfdv = TfdvOp(train_data, evaluation_data, project, output_dir)
    preprocess = PreprocessOp(train_data, evaluation_data, tfdv.output['schema'], project, output_dir)
    training = DnnTrainerOp(preprocess.output, tfdv.schema, learning_rate, hidden_layer_size, steps,
                            target, output_dir)
    tfma = TfmaOp(training.output, evaluation_data,
                  tfdv.schema, project, output_dir)
    deploy = TfServingDeployerOp(training.output)
```

```
dsl.compile(taxi_cab_classification,  'tfx.tar.gz')
run = client.run_pipeline('tfx_run', 'tfx.tar.gz', params={'output': 'gs://dpa22', 'project': 'my-project-33'})
```

Kubeflow Pipelines with Tekton hits v1.0



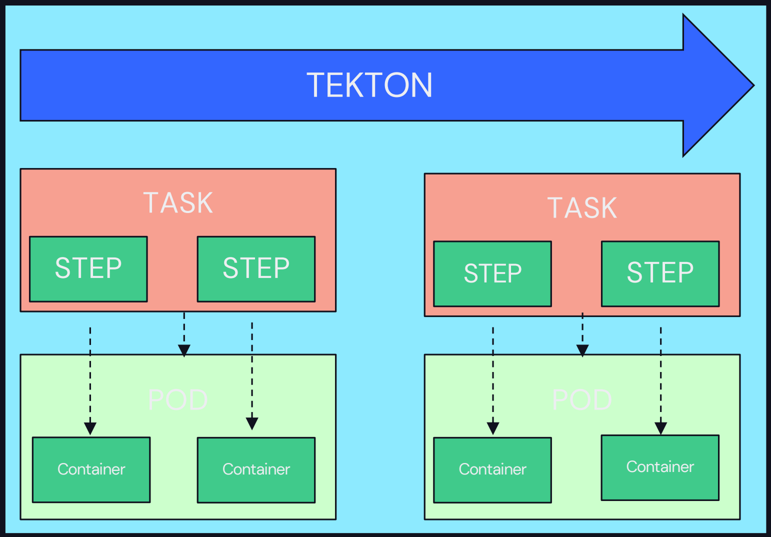
Pluggable Components

Spark | Watson Studio | WML | Open Scale | Kubeflow Training | Seldon | AIF360 | ART | KATIB | KSERVE

https://developer.ibm.com/blogs/kubeflow-pipelines-and-tekton-advances-data-workloads/

Tekton

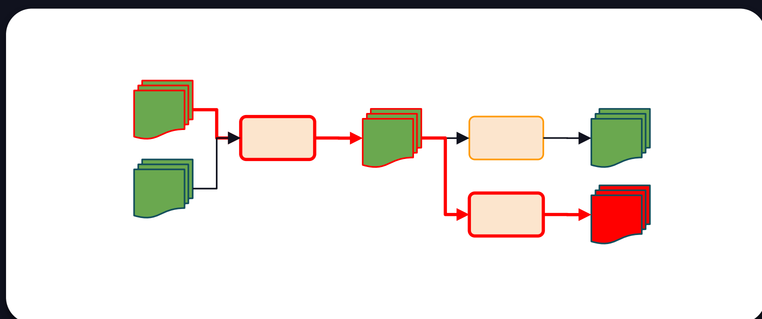
- The Tekton Pipelines project provides Kubernetes-style resources for declaring CI/CD-style pipelines.
- Tekton introduces several new CRDs including Task, Pipeline, TaskRun, and PipelineRun.
- A PipelineRun represents a single running instance of a Pipeline and is responsible for creating a Pod for each of its Tasks and as many containers within each Pod as it has Steps.



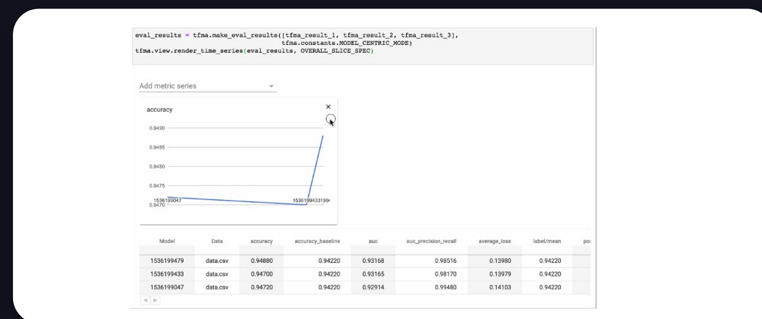
- A **PipelineRun** defines an execution of a pipeline. It references the Pipeline to run.
- A **Pipeline** defines the set of Tasks that compose a pipeline.
- A **TaskRun** defines an execution of a task. It references the task to run.
- A **Task** defines a set of build Steps such as compiling code, running tests, and building and deploying images.

Benefits of metadata and artifact tracking

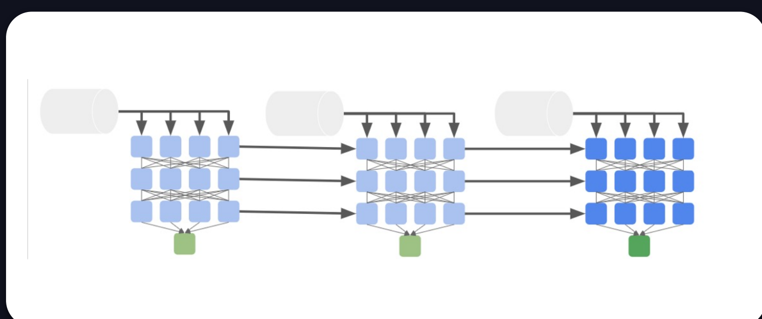
Find out which data a model was trained on



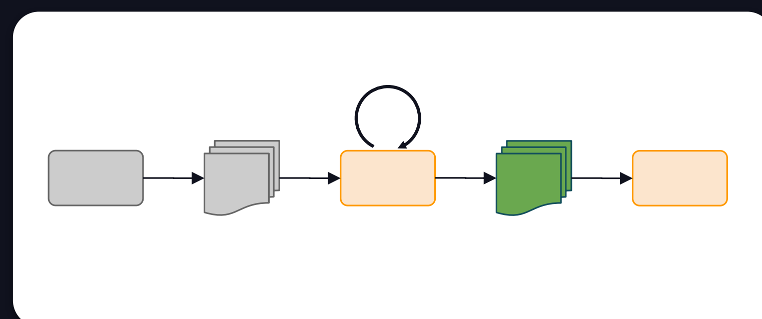
Compare previous model runs



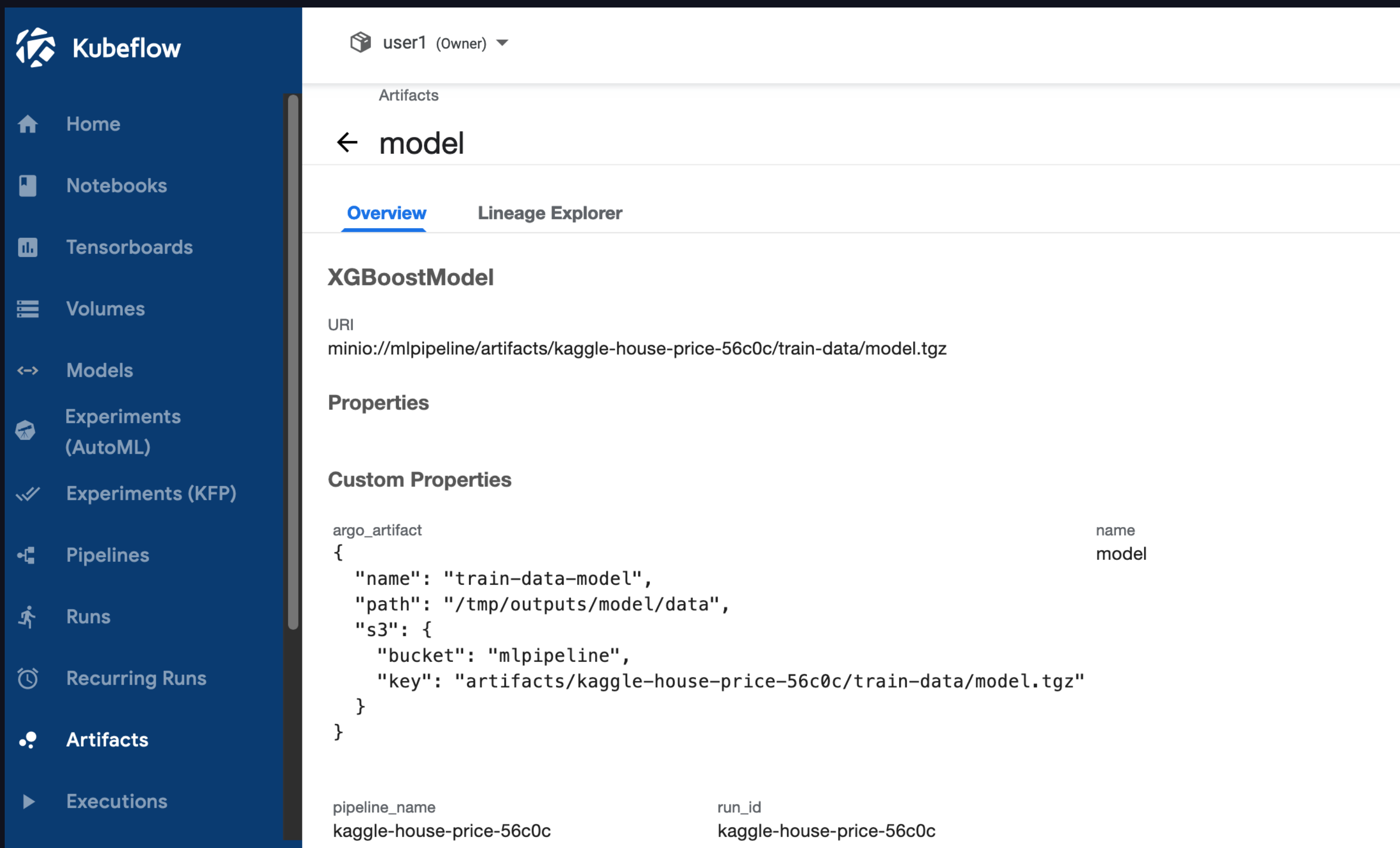
Carry-over state from previous models



Re-use previously computed outputs



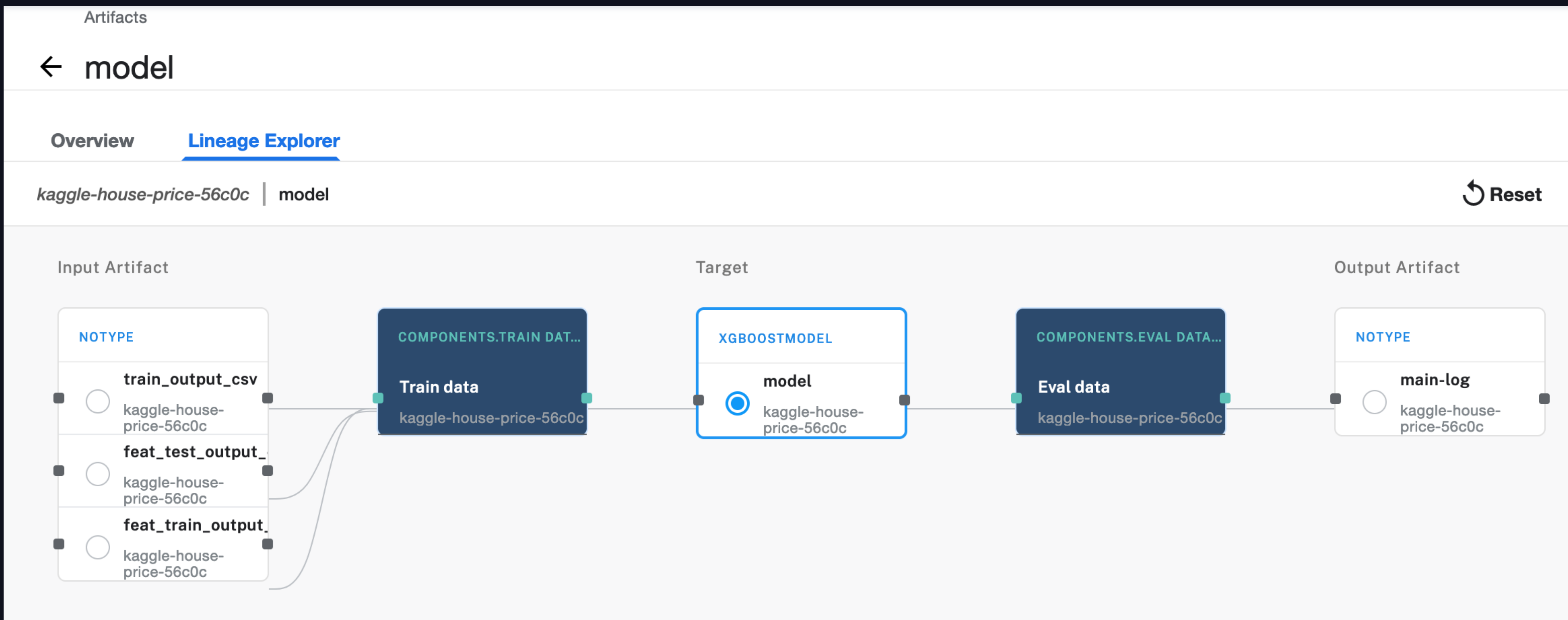
Artifact Tracking



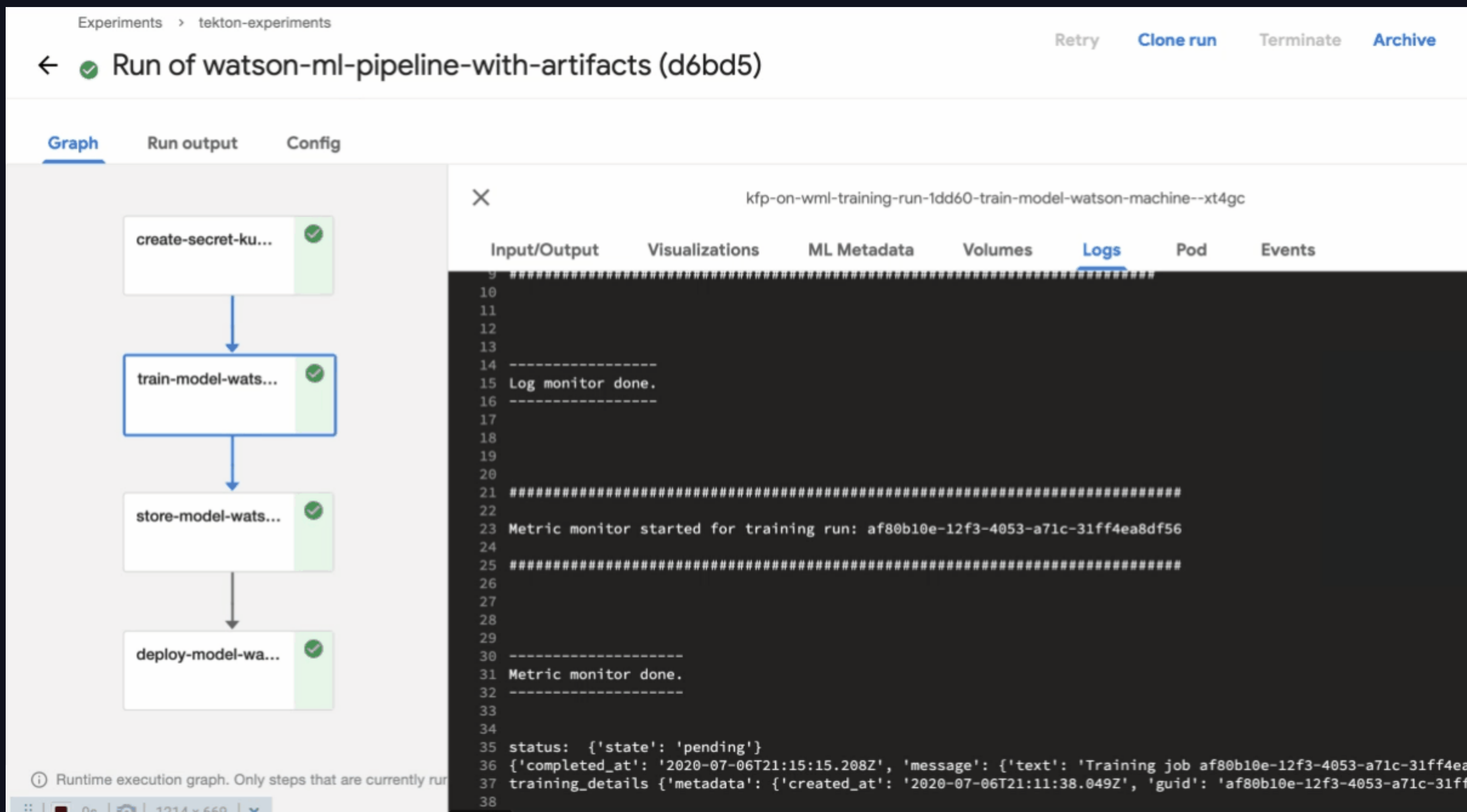
Artifacts for a run of the “Kaggle House Price” example pipeline. For each artifact, you can view details and get the artifact URL—in this case, for the model.

Lineage Tracking

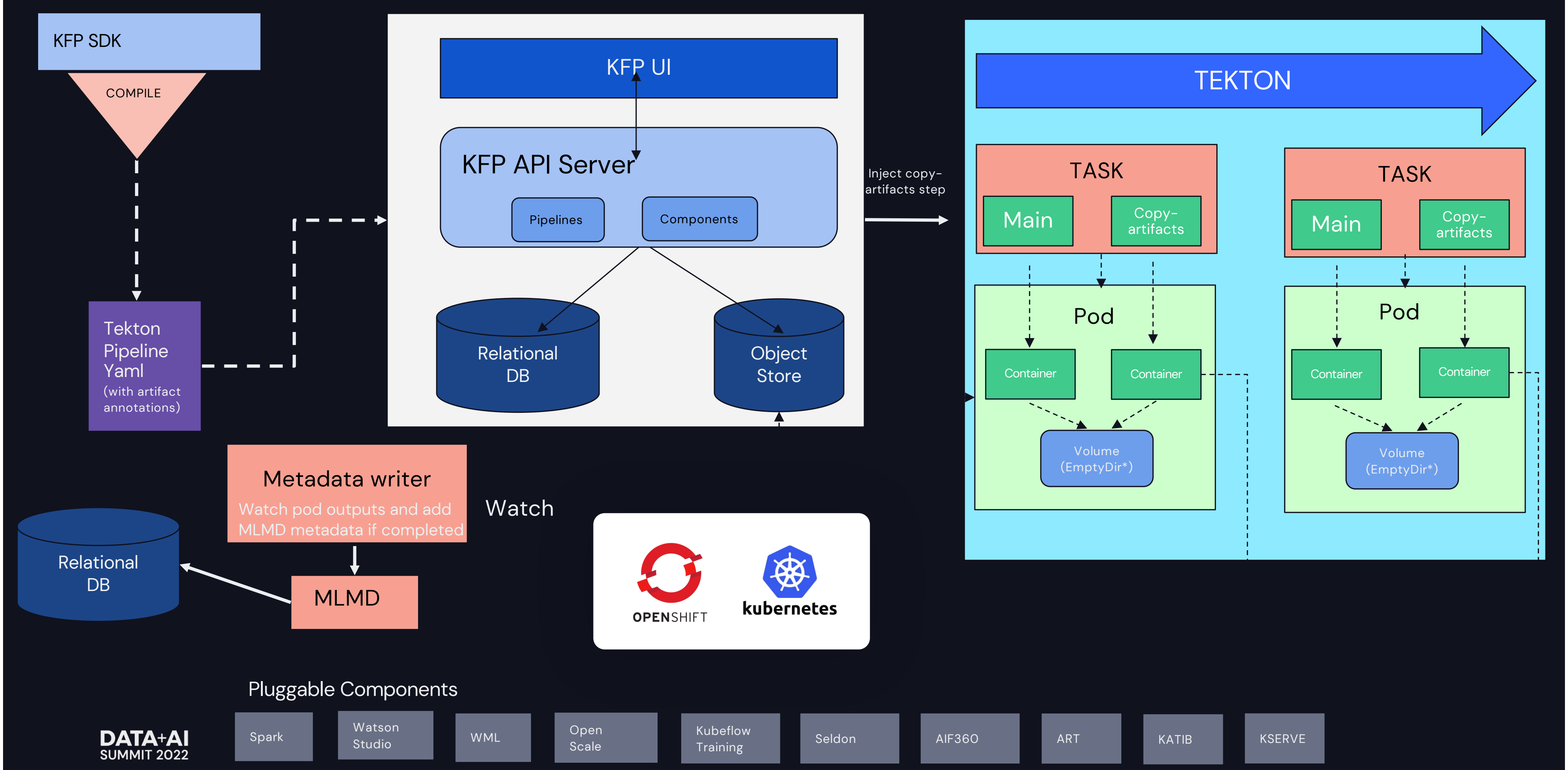
For a given run, the Pipelines Lineage Explorer lets you view the history and versions of your models, data, and more.



Kubeflow Pipelines on Tekton: Logs, Lineage Tracking and Artifact Tracking



Kubeflow Pipelines with Tekton: Metadata and Artifact tracking

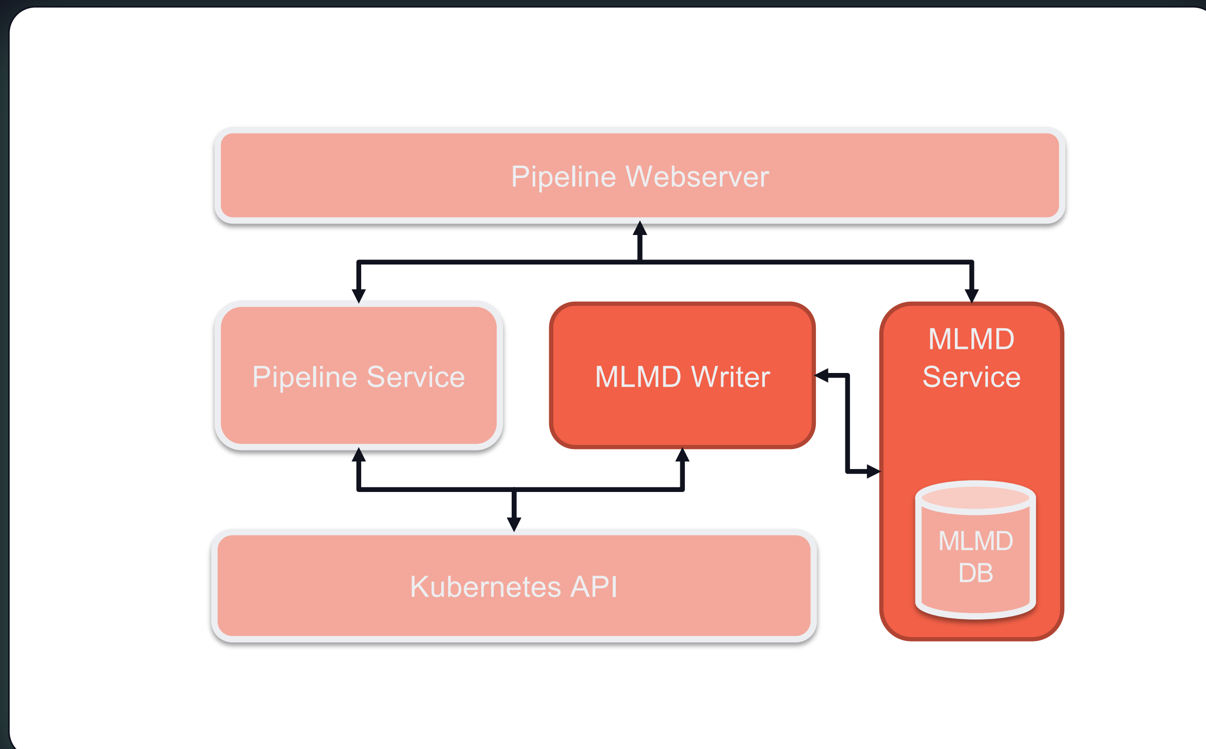


Kubeflow Pipelines v2

Machine Learning Metadata in v1

`MLMD service + MLMD writer`

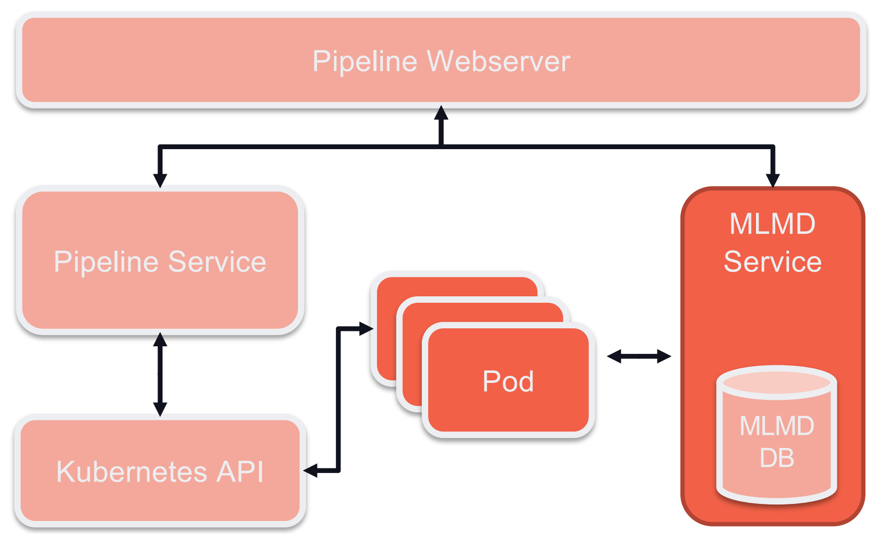
- Asynchronous process
- Preliminary data
- No way to use MLMD to do data passing for Pipeline Run



Machine Learning Metadata in v2

Integrate MLMD with pipeline execution natively

- Integrate MLMD into pipeline execution
- Extend metadata, including pipeline status, parameters, artifacts, etc.
- Use MLMD in pipeline tasks
- Caching key calculation
- Source of truth of the Pipeline Run UI



Pipeline Spec in v1

`Platform-dependent Pipeline Spec`

- SDK generates Argo/Tekton YAML
- Pipeline UI only understands specific CR

```yaml
apiVersion: argoproj.io/v1alpha1
kind: Workflow
metadata:
  annotations:
    pipelines.kubeflow.org/pipeline_spec: '{"description": "Get Most Frequent Word and Save to GCS", "inputs"
  generateName: save-most-frequent-
spec:
  arguments:
    parameters:
    - name: message
    - name: outputpath
  entrypoint: save-most-frequent
  serviceAccountName: pipeline-runner
  onExit: exiting
  templates:
  - dag:
      tasks:
      - arguments:
          parameters:
          - name: message
            value: '{{inputs.parameters.message}}'
        name: get-frequent
        template: get-frequent
```

Workflow

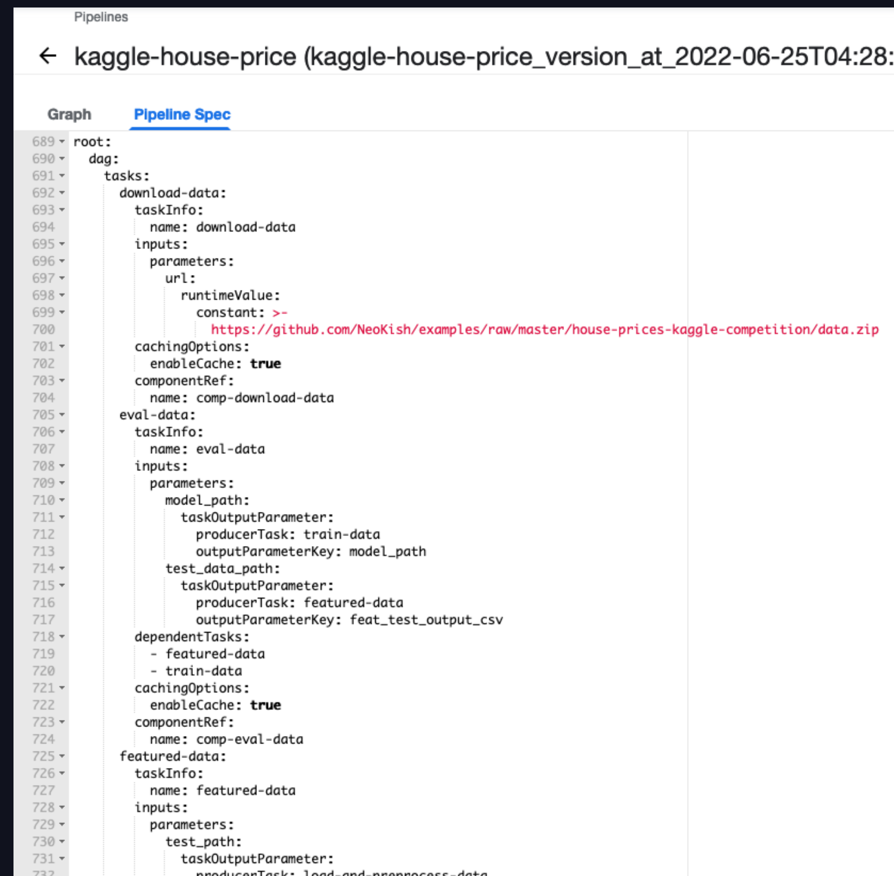
```yaml
apiVersion: tekton.dev/v1beta1
kind: PipelineRun
metadata:
  name: save-most-frequent
  annotations:
    tekton.dev/output_artifacts: '{"get-frequent": [{"key": "artifacts/$PIPELINERUN/get-frequent/word.tgz",
      "name": "get-frequent-word", "path": "/tmp/message.txt"}]}'
    tekton.dev/input_artifacts: '{"save": [{"name": "get-frequent-word", "parent_task":
      "get-frequent"}]}'
    tekton.dev/artifact_bucket: mlpipeline
    tekton.dev/artifact_endpoint: minio-service.kubeflow:9000
    tekton.dev/artifact_endpoint_scheme: http://
    tekton.dev/artifact_items: '{"exiting": [], "get-frequent": [["word", "$(results.word.path)"]],
      "save": []}'
    sidecar.istio.io/inject: "false"
    pipelines.kubeflow.org/big_data_passing_format: $(workspaces.$TASK_NAME.path)/artifacts/$ORIG_PR_NAME/$TA
    pipelines.kubeflow.org/pipeline_spec: '{"description": "Get Most Frequent Word
      and Save to GCS", "inputs": [{"name": "message", "type": "String"}, {"name":
      "outputpath", "type": "String"}], "name": "Save Most Frequent"}'
spec:
  params:
  - name: message
    value: ''
  - name: outputpath
    value: ''
  pipelineSpec:
    params:
```

PipelineRun

Intermediate Representation in v2

Agnostic Pipeline Spec

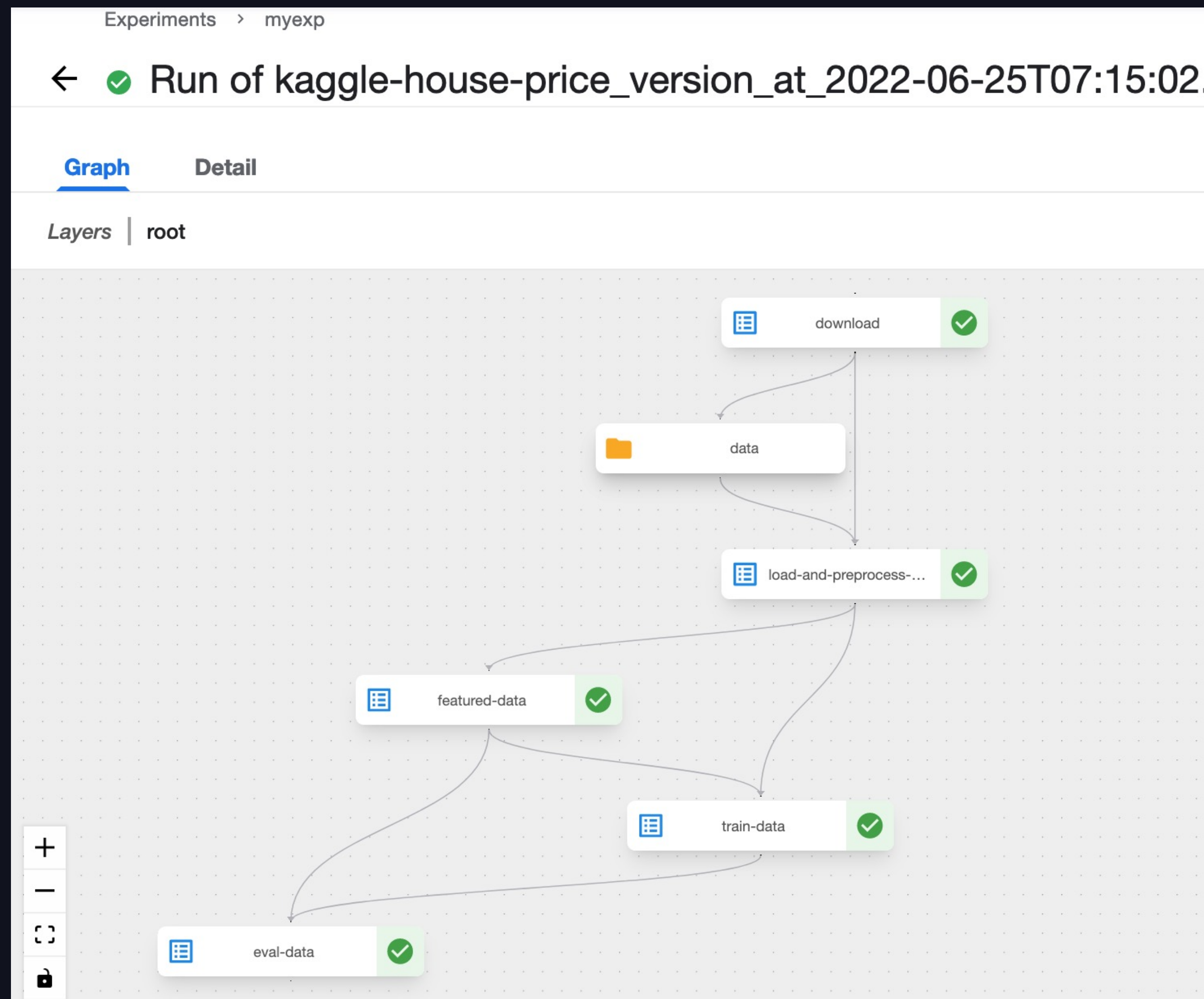
- SDK generates IR in YAML format
- Easy to interpret
- Speed up Low Code/No Code integration



DATA+AI SUMMIT 2022

New UX for v2

- New v2 IR
- Retrieve Pipeline Run information from MLMD



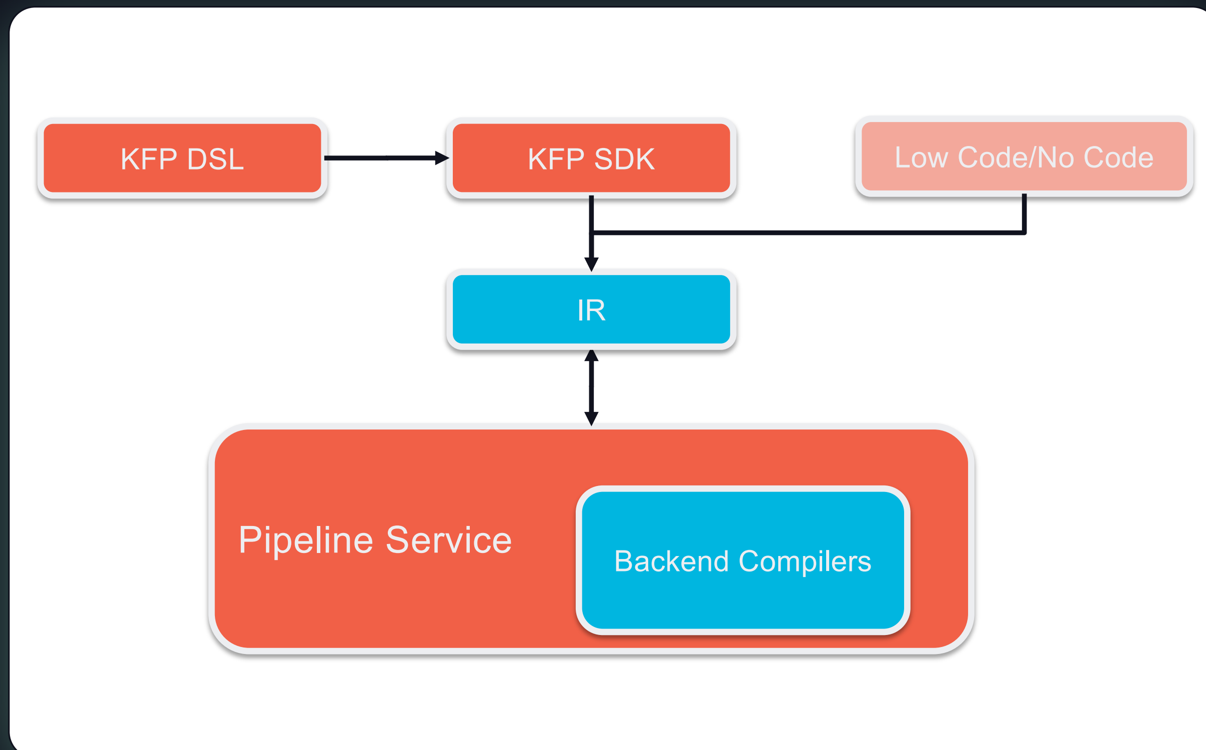
DATA+AI SUMMIT 2022

New Orchestration Controllers

Smart Compiler → Smart Runtime

Backend Compiler

- Encapsulate backend engines
- Hide the platform specific CR from users

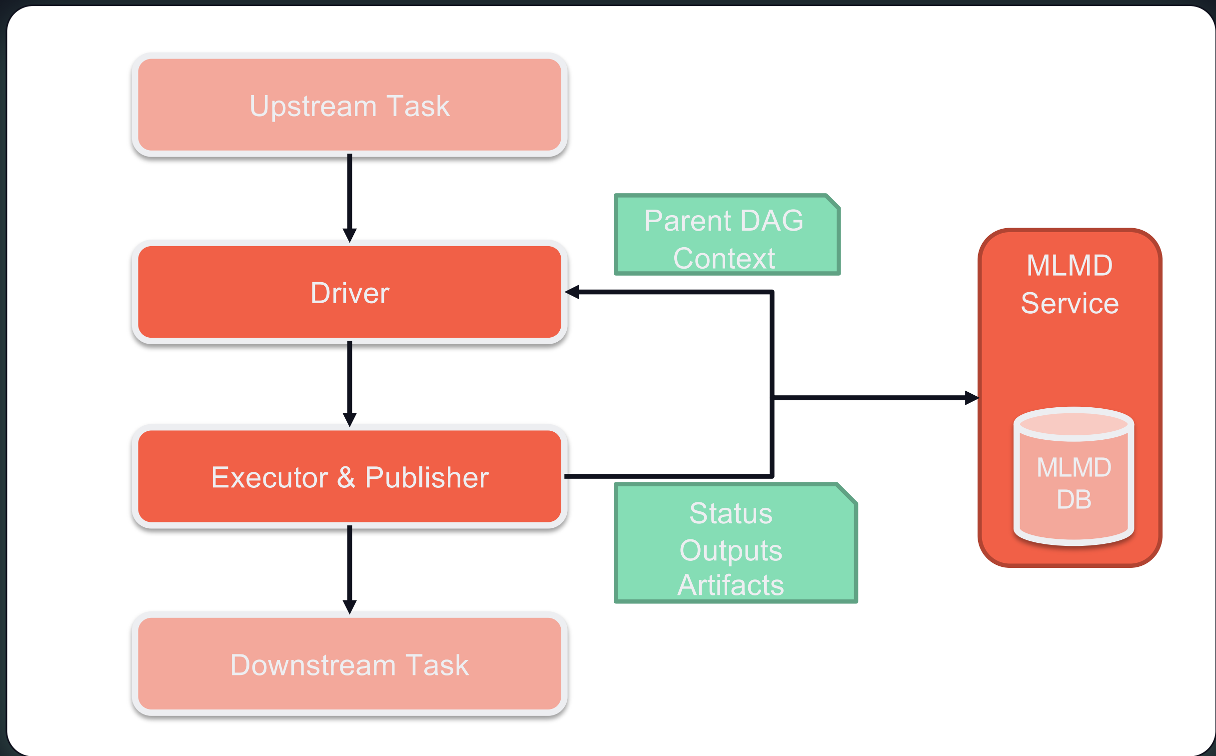


DATA+AI SUMMIT 2022

Smart Runtime

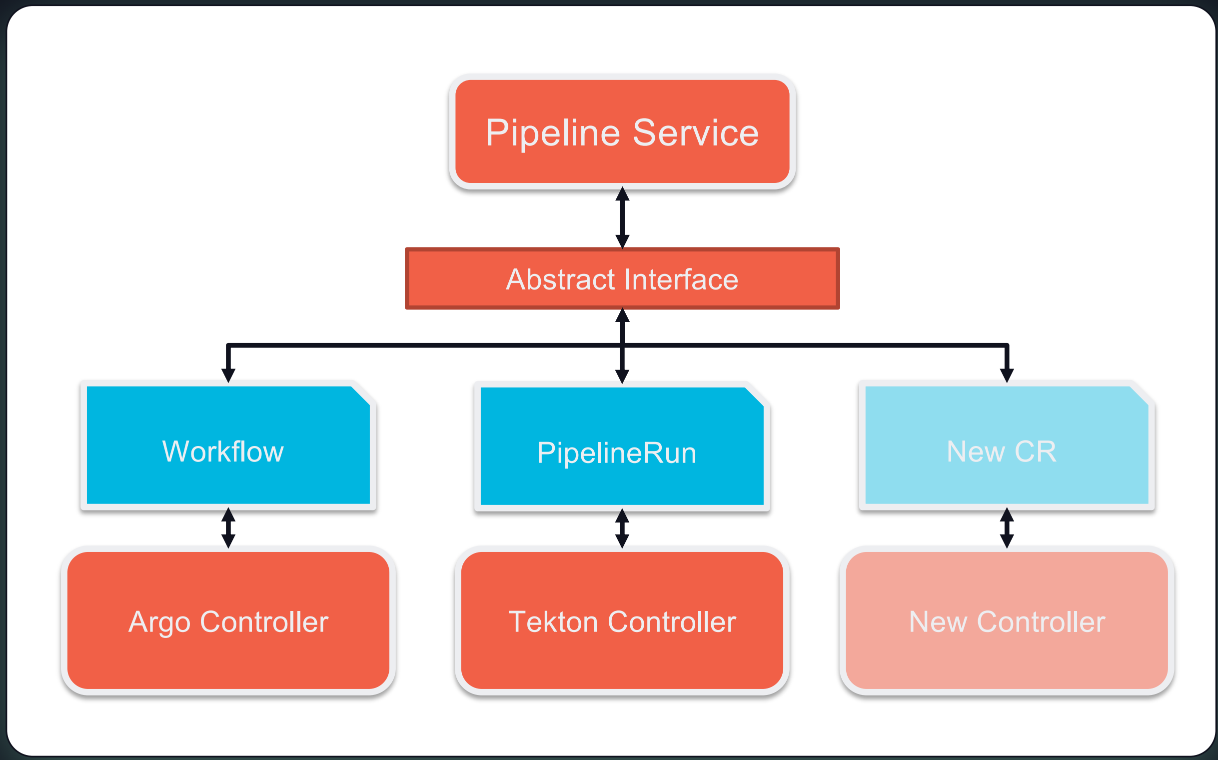
Driver/Executor/Publisher

- Gain more controls of the pipeline execution
- Easier to add features
- Natively integrate with MLMD



Abstraction Layer for Orchestration Engines

- Communicate with orchestration engines with single interface
- Expand the support to other orchestration engines



DATA+AI SUMMIT 2022

Components

Roadmap

Components

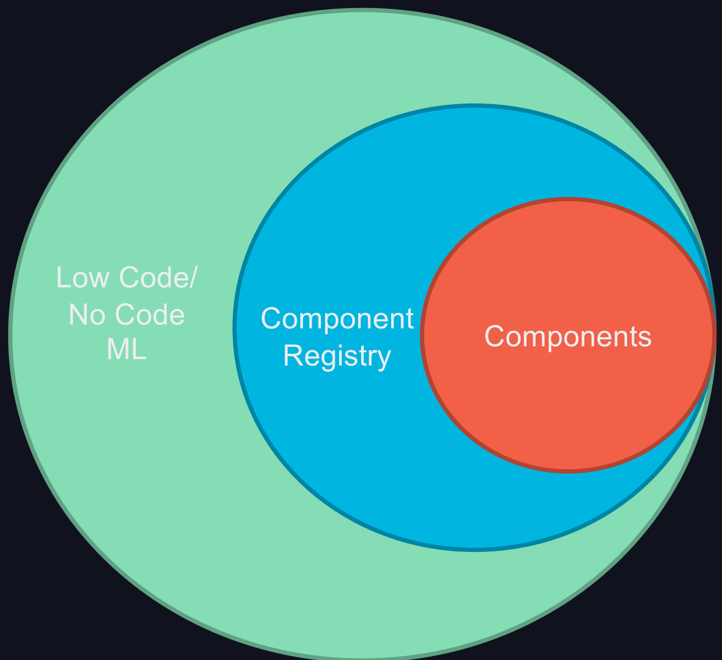
A rich set of components from community and vendors.

Component/Pipeline Registry

KFP SDK can directly load the components from the registry as long as it follows a standardized protocol.

Low code/No code ML

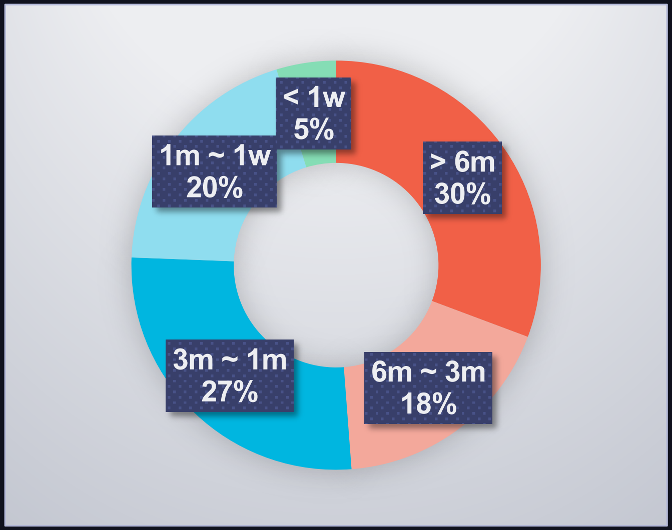
Create E2E ML workflow via drag-and-drop



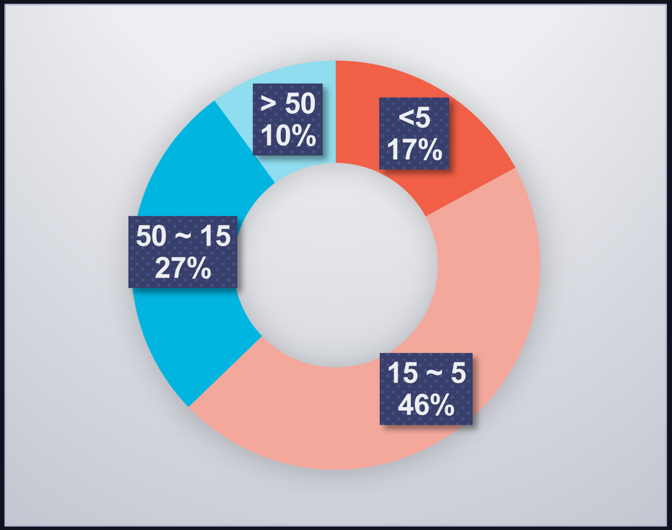
Challenges for ML

Kubeflow Community Survey

Average life of a model in production



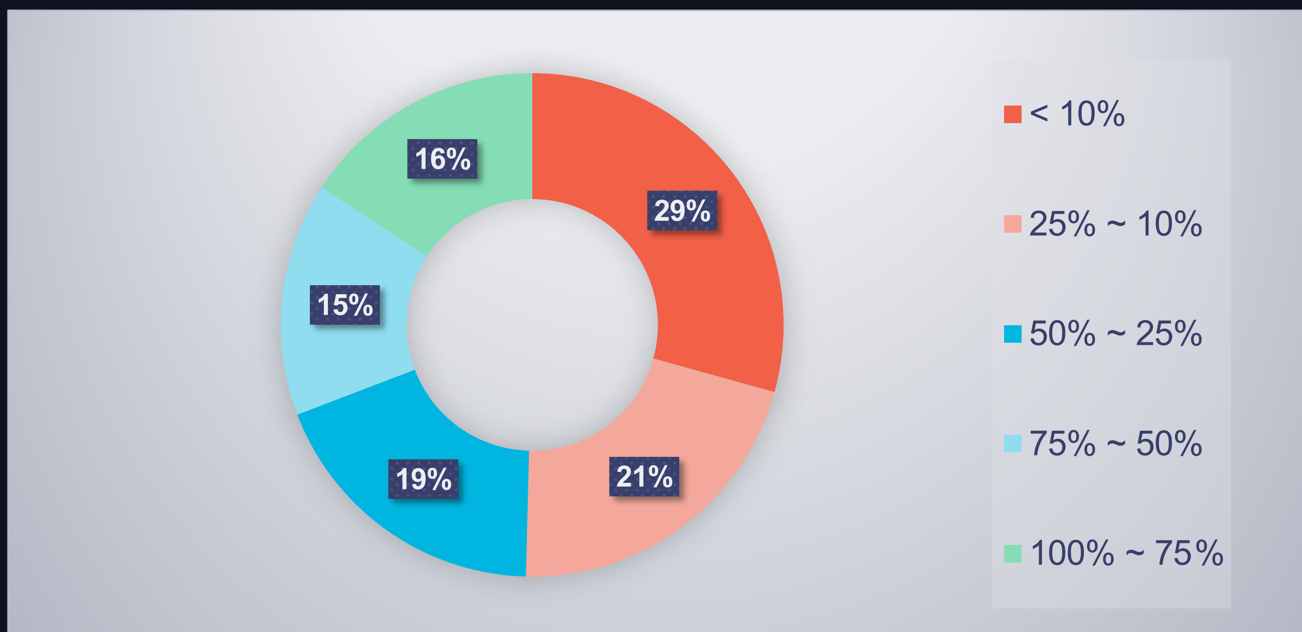
How many iterations does it take to produce a production model



Source: [Kubeflow Community](https://www.youtube.com/watch?v=zR-Ji1VP3pI)
https://www.youtube.com/watch?v=zR-Ji1VP3pI

Cont.

What percentage of your 2021 models were successfully deployed into production and are delivering business values?



Source: Kubeflow Community
https://www.youtube.com/watch?v=zR-Ji1VP3pI

Summary

- Embrace MLMD as first step toward MLOps
  - https://github.com/kubeflow/pipelines/tree/master/third_party/ml-metadata
- Use IR or component-based strategy to compose ML pipelines
  - https://www.kubeflow.org/docs/components/pipelines/sdk-v2/component-development/
- Leverage abstraction layer to bring your orchestration engine to Kubeflow Pipelines
  - https://github.com/kubeflow/pipelines/blob/master/backend/src/common/util/execution_spec.go#L51

References

- Kubeflow Pipelines
  https://github.com/kubeflow/pipelines/
- Kubeflow Pipelines on Tekton
  https://github.com/kubeflow/kfp-tekton
- Kubeflow Pipelines v2 Design
  - http://bit.ly/kfp-v2

DATA+AI SUMMIT 2022

Thank you

Yi-Hong Wang
Software Developer

Tommy Li
Software Developer