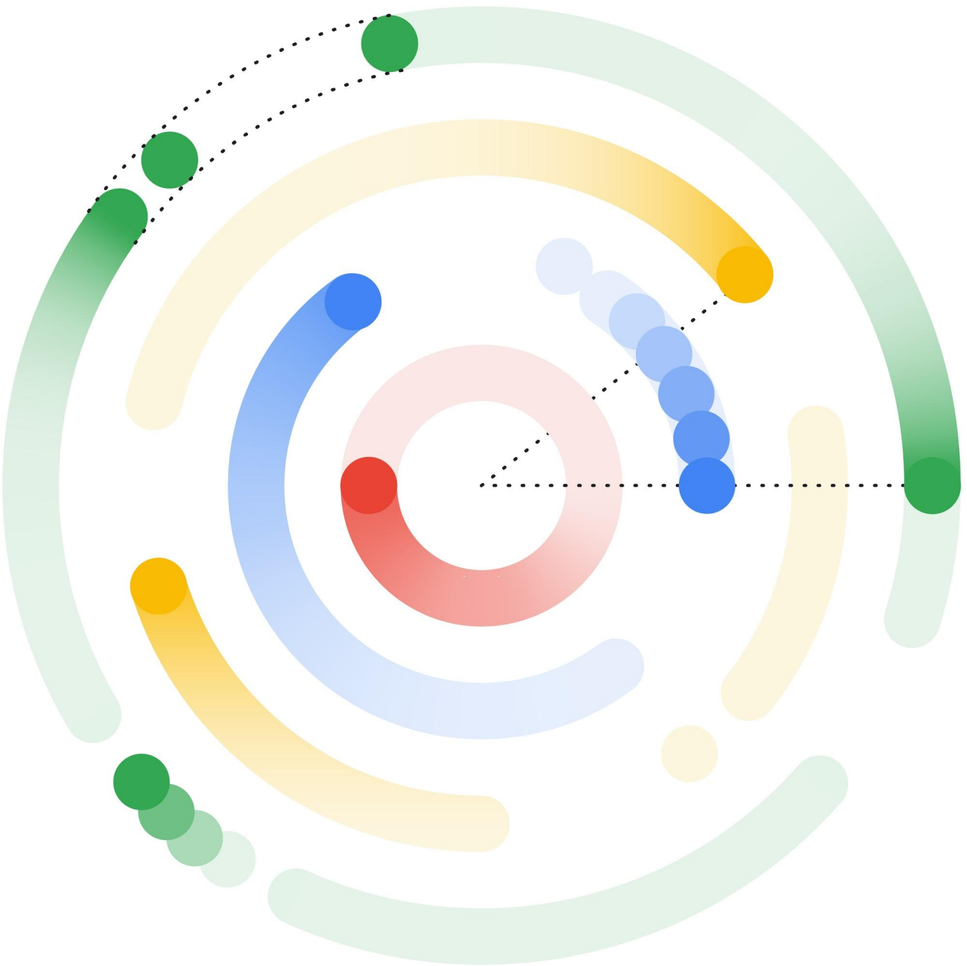
Google Cloud

Rapid Iteration with Limited DevOps Resource

Google Cloud Applied ML Summit

How Vertex Pipelines accelerate our ML projects

06/10/21

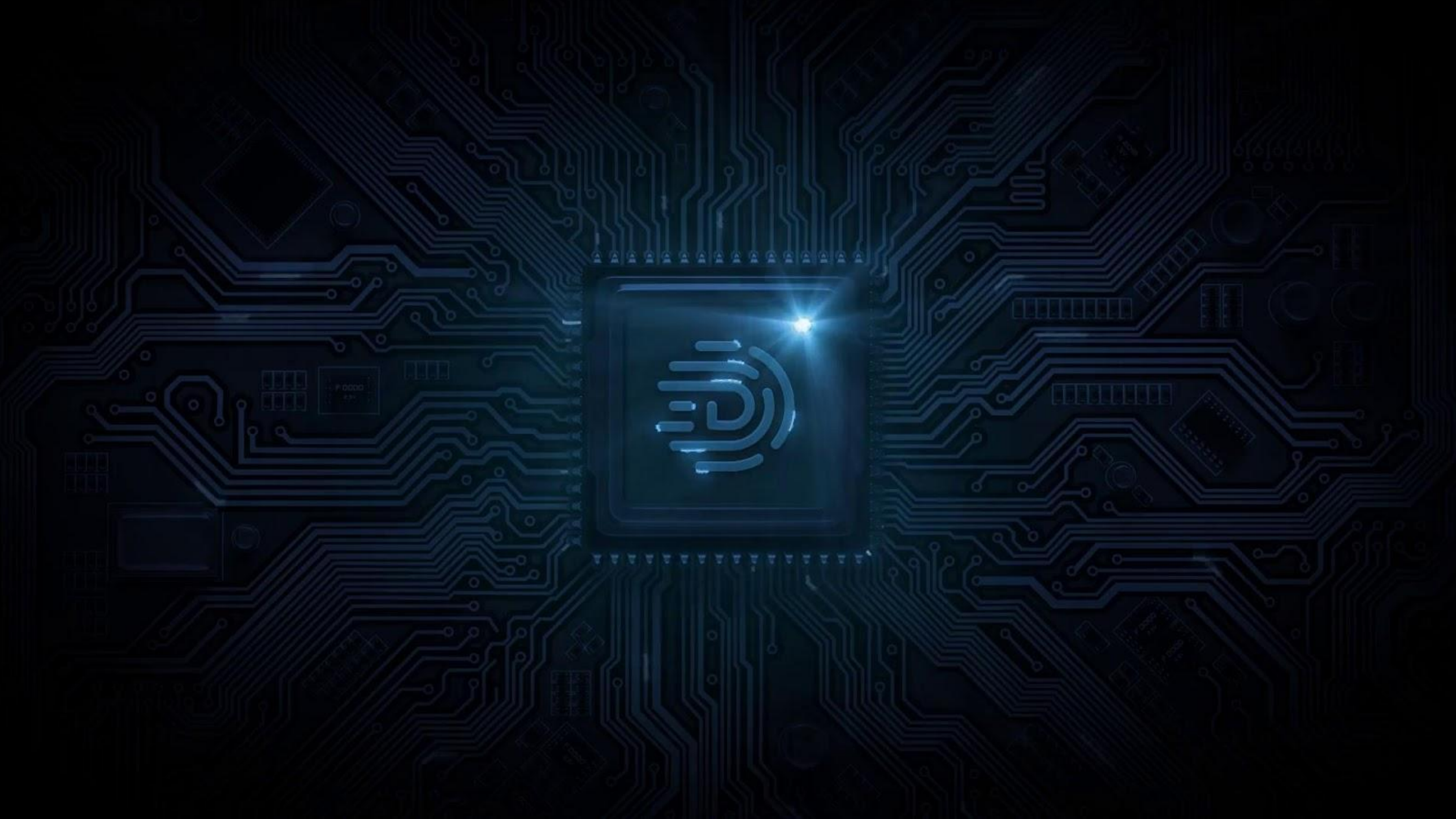


Hannes Hapke

ML Engineer,
Digits Inc.

Google Cloud

What is Digits?



Google Cloud

Machine Learning at Digits?

ML @ Digits

- Information extraction
- Event predictions
- Clustering of information
- Deduplication

01

State of Production ML

Best practices slowly emerge

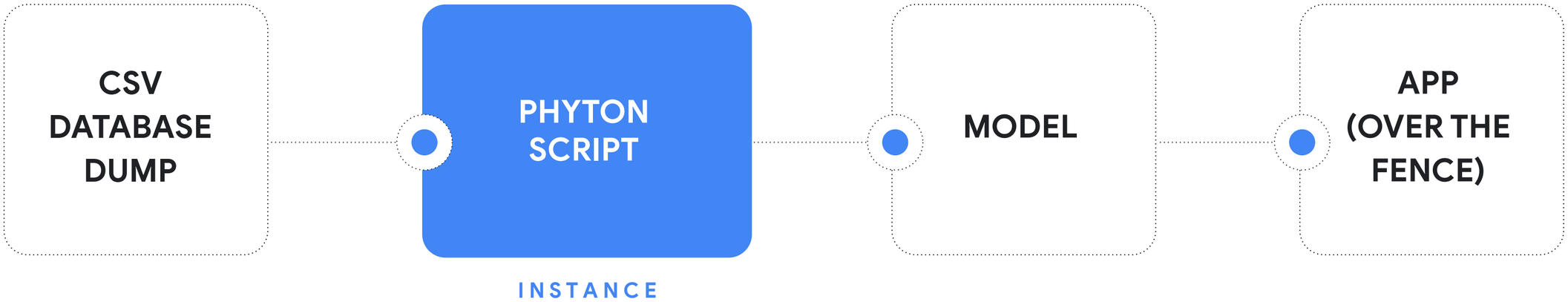
Focus in the community shifts from focus on ML architectures to production systems

Example

Common Machine Learning Setup

Example

Common Machine Learning Setup



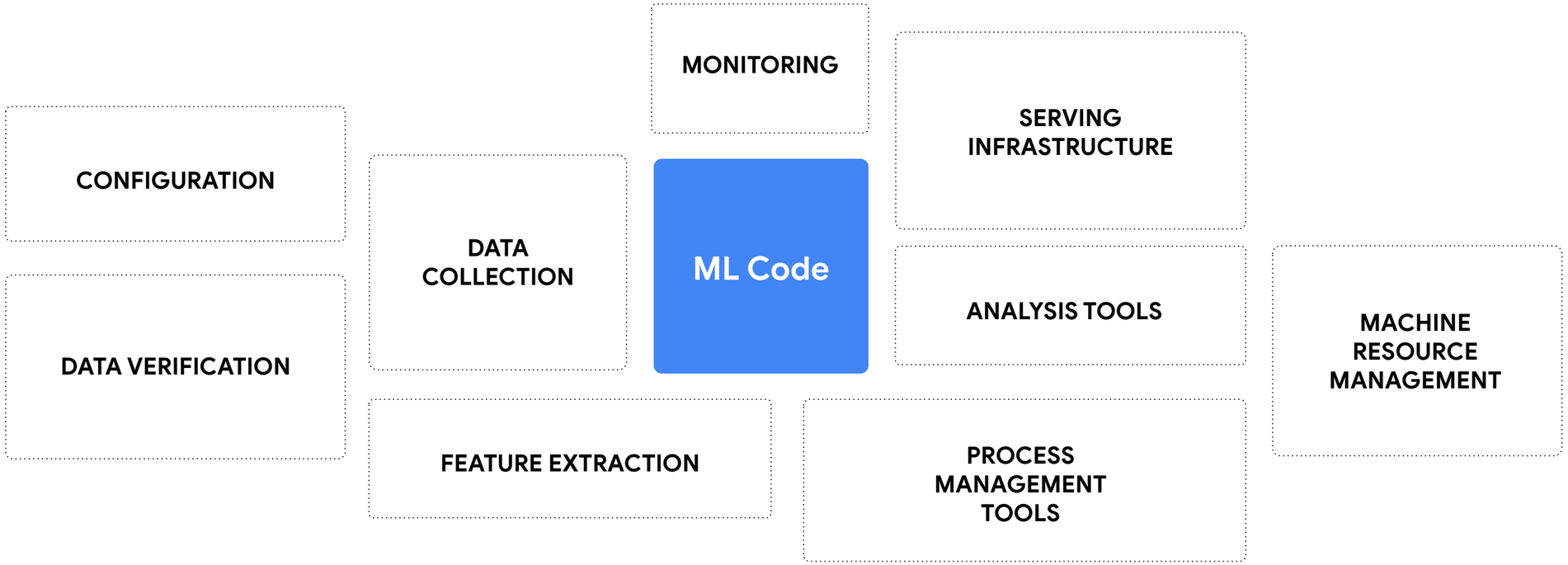
02

Machine Learning Ops

What is ML Ops?

Original image:
https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

What is ML Ops?

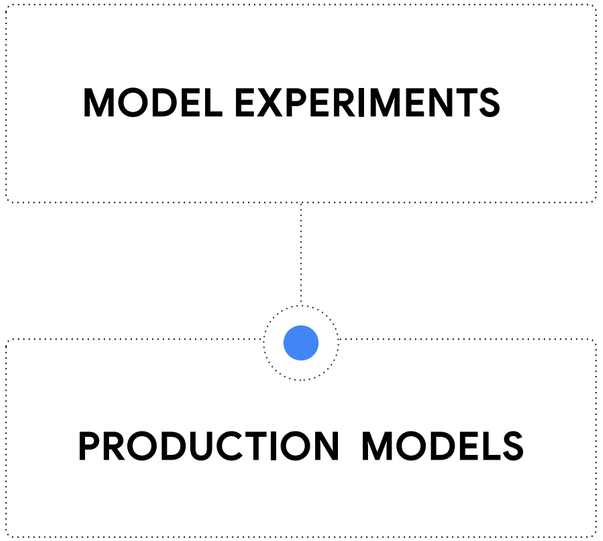


Original image:
https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

What is MLOps?

Why do we need it?

- Integrate models in Real world scenarios
- Focus on reproducibility
- Provide traceability via audit trails
- Reduce burden for data scientists

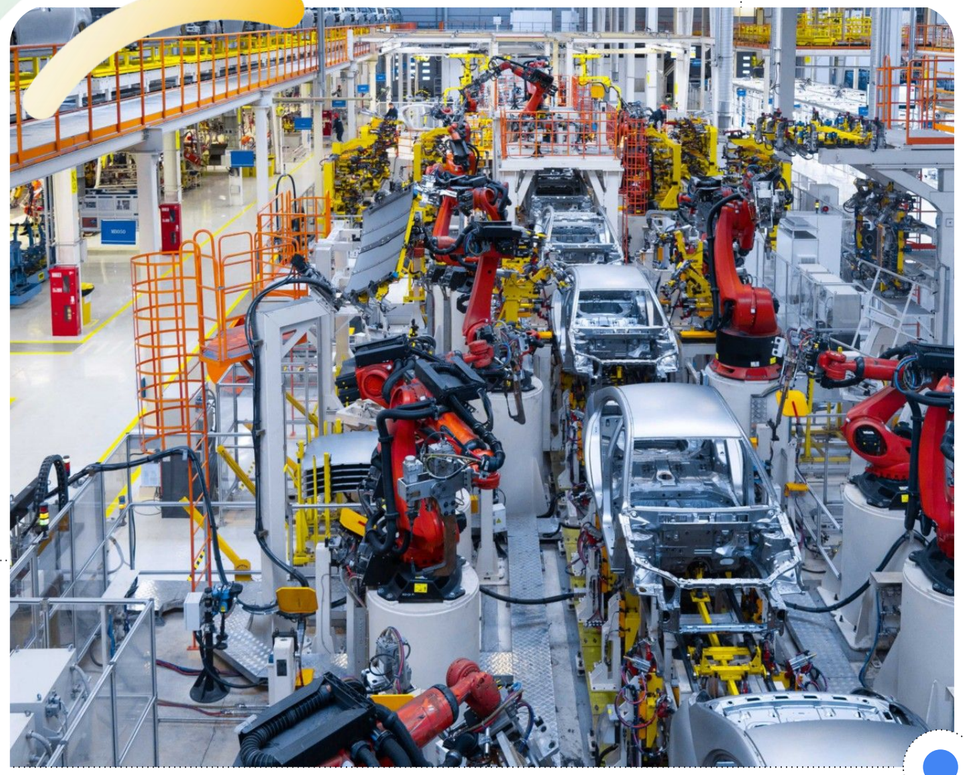


Why do we need Model Pipelines?

ML Experiments in Notebooks

Why do we need Model Pipelines?

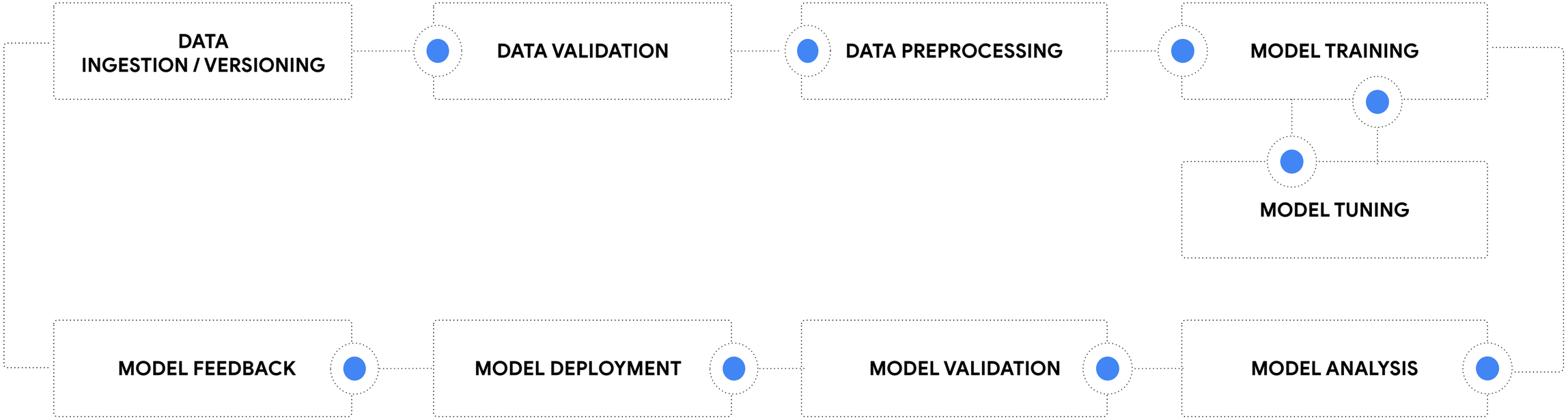
ML Experiments in Notebooks



ML Pipelines

https://www.ey.com/en_gl/advanced-manufacturing/how-digital-twins-give-automotive-companies-a-real-world-advantage

Model Life Cycle

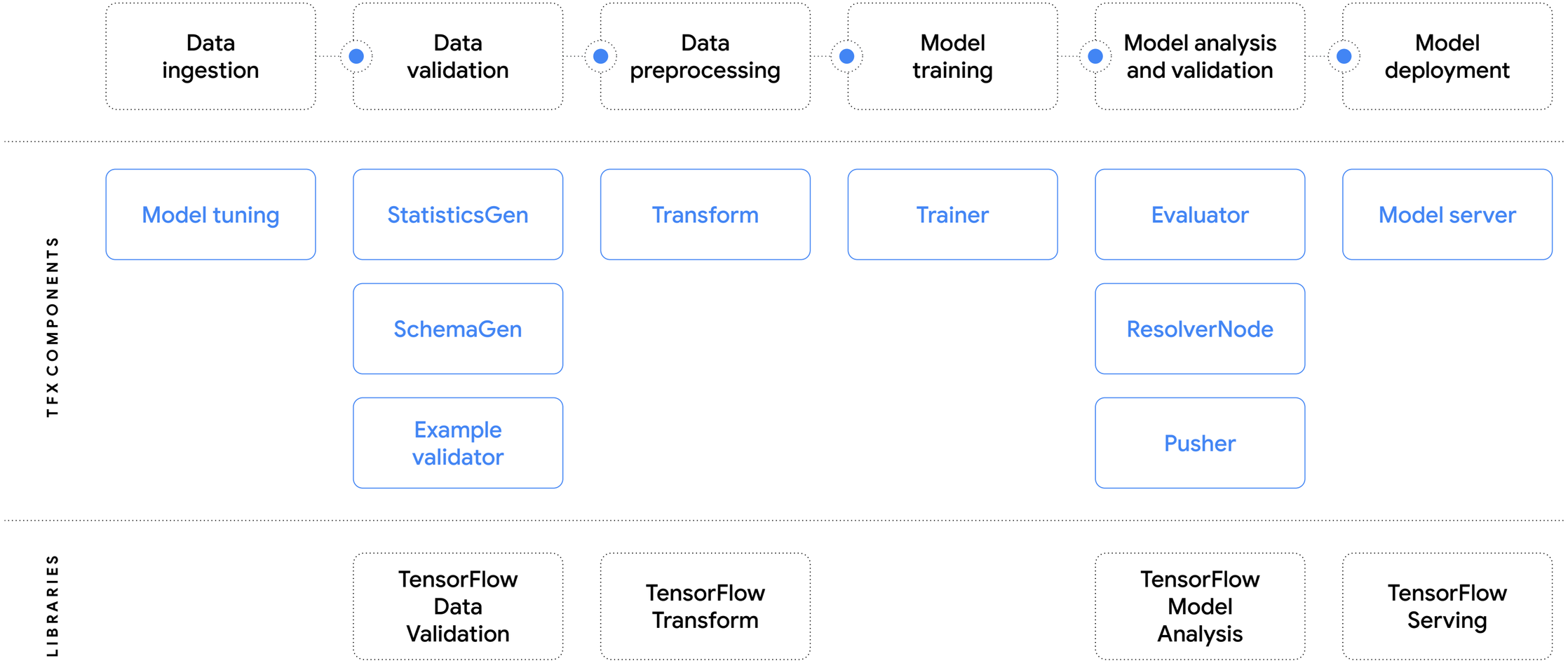


Original image: “Building Machine Learning Pipelines”
https://learning.oreilly.com/library/view/building-machine-learning/9781492053187/

03

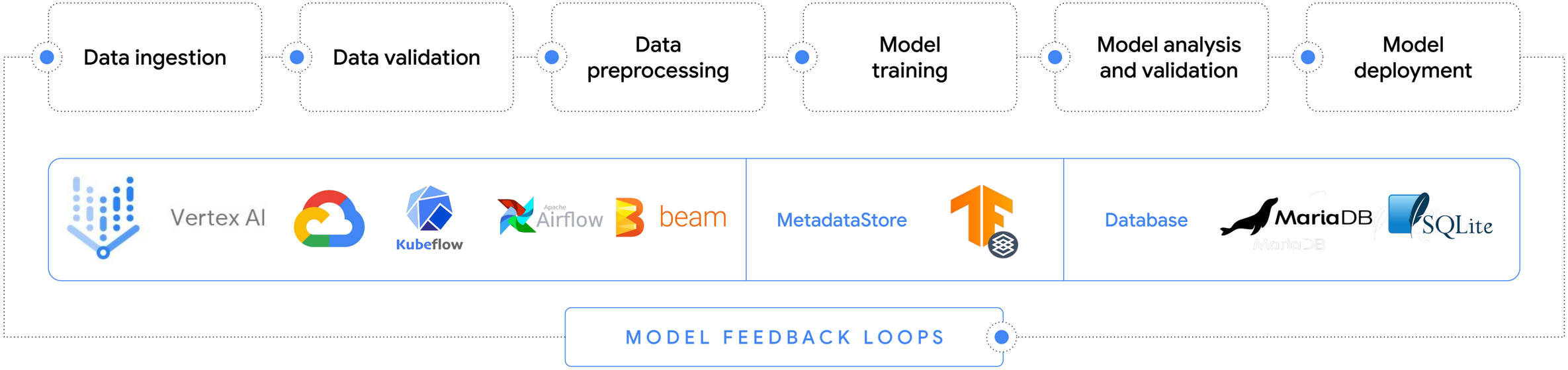
TensorFlow Extended

Why TFX?



Original image: “Building Machine Learning Pipelines”
https://learning.oreilly.com/library/view/building-machine-learning/9781492053187/

Why TFX?



Original image: “Building Machine Learning Pipelines”
https://learning.oreilly.com/library/view/building-machine-learning/9781492053187/

Why TFX?

- Functionality covers the entire model life cycle
- Orchestration agnostic
- Metadata at its core
- TFX can be easily customized with custom components

```
# Data ingestion
QUERY = “SELECT * FROM examples;”
example_gen = BigQueryExampleGen(query=QUERY)

# Computes statistics
statistics_gen = StatisticsGen(examples=example_gen.outputs['examples'])

# Generates schema
schema_gen = SchemaGen(
    statistics=statistics_gen.outputs['statistics'], infer_feature_shape=True)

# Performs feature engineering
transform = Transform(
    examples=example_gen.outputs['examples'],
    schema=schema_gen.outputs['schema'],
    module_file=module_file)
...
```

Orchestration

- Orchestrate entire ML pipelines
- Google Clouds Vertex Pipelines
- Kubeflow Pipelines
- Apache Airflow
- Apache Beam

```
components = [example_gen, statistics_gen, ...]


mlmd_conn_config = \
   metadata.sqlite_metadata_connection_config(metadata_path)


beam_arg = [f"--direct_num_workers={direct_num_workers}"]


tfx_pipeline = pipeline.Pipeline(
       components=components,
       enable_cache=True,
       metadata_connection_config=mlmd_conn_config,
       beam_pipeline_args=beam_arg,
       …
)

BeamDagRunner().run(tfx_pipeline)
```

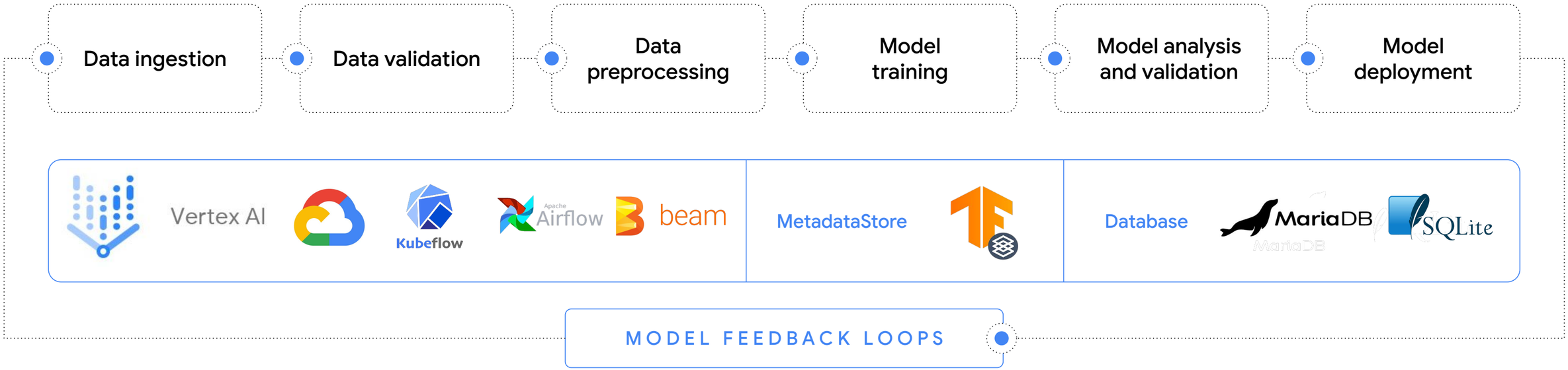
Orchestration

```
...
INFO:absl:Component FileBasedExampleGen depends on [].
INFO:absl:Component FileBasedExampleGen is scheduled.
INFO:absl:Component ResolverNode.latest_blessed_model_resolver depends on [].
INFO:absl:Component ResolverNode.latest_blessed_model_resolver is scheduled.
INFO:absl:Component StatisticsGen depends on ['Run[FileBasedExampleGen]'].
INFO:absl:Component StatisticsGen is scheduled.
INFO:absl:Component SchemaGen depends on ['Run[StatisticsGen]'].
INFO:absl:Component SchemaGen is scheduled.
INFO:absl:Component ExampleValidator depends on ['Run[SchemaGen]', 'Run[StatisticsGen]'].
INFO:absl:Component ExampleValidator is scheduled.
INFO:absl:Component Transform depends on ['Run[SchemaGen]', 'Run[FileBasedExampleGen]'].
INFO:absl:Component Transform is scheduled.
INFO:absl:Component Trainer depends on ['Run[SchemaGen]', 'Run[Transform]'].
INFO:absl:Component Trainer is scheduled.
...
```

04

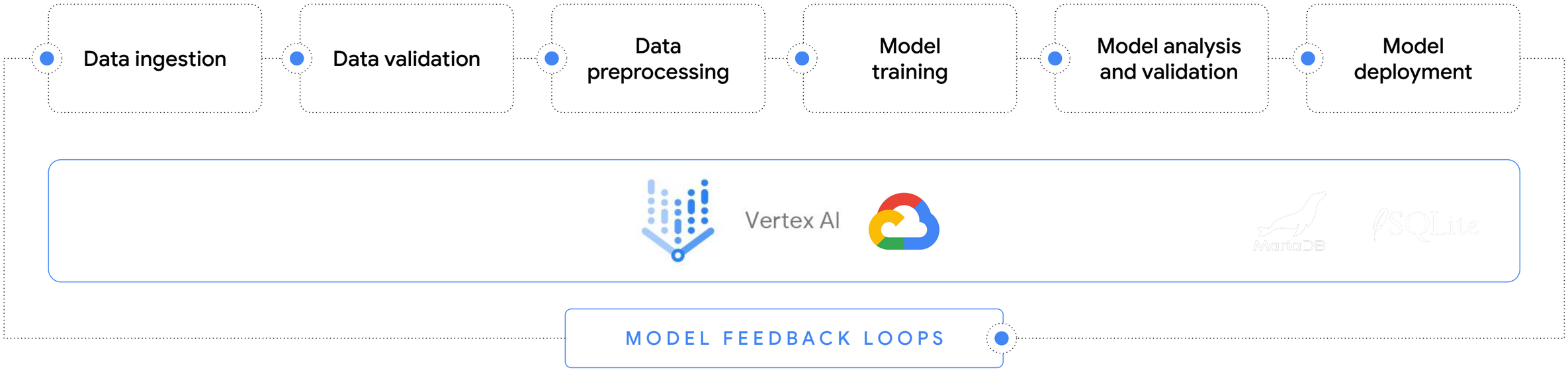
Google Cloud's Vertex Pipelines

Managed Pipelines

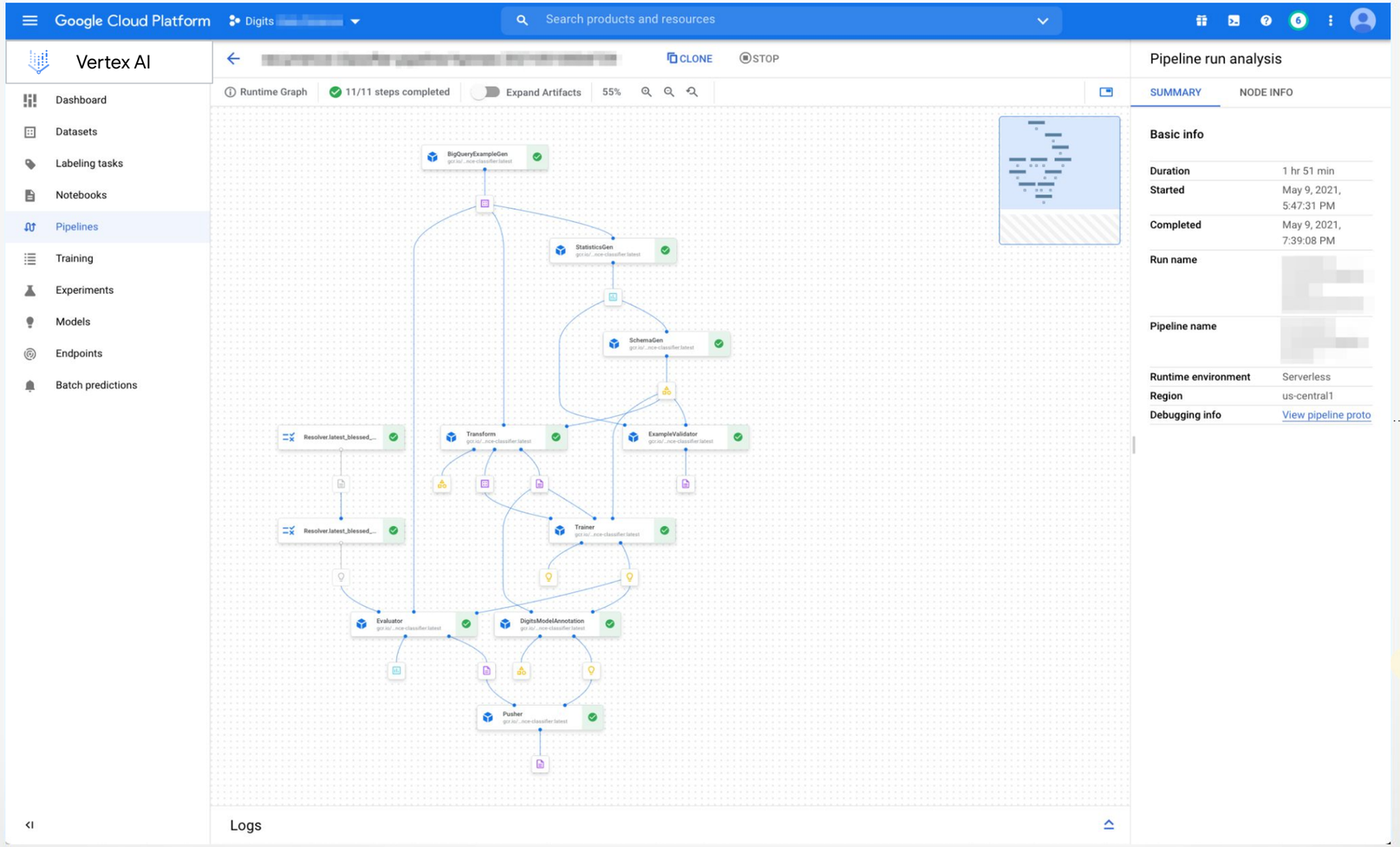


Original image: “Building Machine Learning Pipelines”
https://learning.oreilly.com/library/view/building-machine-learning/9781492053187/

Managed Pipelines



Original image: “Building Machine Learning Pipelines”
https://learning.oreilly.com/library/view/building-machine-learning/9781492053187/



Google Cloud

05

Why Vertex Pipelines

“We can use Software Engineering CI/CD systems for ML.”

>This is not recommended!

MLOps isn’t CI/CD

Learn from software engineering, but don’t treat it like it

- Compare model versions with previous versions
- Intermediate pipeline artifacts matter
- Components / steps are entangled
- Pipelines need to scale, e.g. for data processing
- Tracking pipeline metadata

“Metadata is an insurance policy for Machine Learning”

ML in the real world

Dirty data is just the beginning ...

- Every changing data sets / schemas
- Privacy concerns
- Legal compliance, e.g. GDPR
- Extension data preprocessing
- Multiple models required for the same problem
- Model inventory requirements
- Model audit trials

“Machine Learning Pipelines is for large machine learning teams with dedicated DevOps colleagues.”

> Not true, no dedicated team needed

```python
from aiplatform.pipelines import schedule

pipeline_config_file_name = f'{constants.MODEL_NAME}_pipeline_config.json'
runner = kubeflow_v2_dag_runner.KubeflowV2DagRunner(
   config=kubeflow_v2_dag_runner.KubeflowV2DagRunnerConfig(
       project_id=PROJECT_ID,
   ),
   output_filename=pipeline_config_file_name)
_ = runner.compile(pipeline, write_out=True)

schedule.create_from_pipeline_file(
   pipeline_path=pipeline_config_file_name,
   schedule='0 5 * * 1',  # Monday's at 5 am
   project_id=PROJECT_ID,
   region=REGION,
   time_zone='America/Los_Angeles',
   parameter_values={}
)
```

```python
from aiplatform.pipelines import schedule

pipeline_config_file_name = f'{constants.MODEL_NAME}_pipeline_config.json'
runner = kubeflow_v2_dag_runner.KubeflowV2DagRunner(
   config=kubeflow_v2_dag_runner.KubeflowV2DagRunnerConfig(
       project_id=PROJECT_ID,
   ),
   output_filename=pipeline_config_file_name)
_ = runner.compile(pipeline, write_out=True)

schedule.create_from_pipeline_file(
   pipeline_path=pipeline_config_file_name,
   schedule='0 5 * * 1',  # Monday's at 5 am
   project_id=PROJECT_ID,
   region=REGION,
   time_zone='America/Los_Angeles',
   parameter_values={}
)
```

```
from aiplatform.pipelines import schedule

pipeline_config_file_name = f'{constants.MODEL_NAME}_pipeline_config.json'
runner = kubeflow_v2_dag_runner.KubeflowV2DagRunner(
   config=kubeflow_v2_dag_runner.KubeflowV2DagRunnerConfig(
       project_id=PROJECT_ID,
   ),
   output_filename=pipeline_config_file_name)
_ = runner.compile(pipeline, write_out=True)

schedule.create_from_pipeline_file(
   pipeline_path=pipeline_config_file_name,
   schedule='0 5 * * 1',  # Monday's at 5 am
   project_id=PROJECT_ID,
   region=REGION,
   time_zone='America/Los_Angeles',
   parameter_values={}
)
```

```
from aiplatform.pipelines import schedule

pipeline_config_file_name = f'{constants.MODEL_NAME}_pipeline_config.json'
runner = kubeflow_v2_dag_runner.KubeflowV2DagRunner(
   config=kubeflow_v2_dag_runner.KubeflowV2DagRunnerConfig(
       project_id=PROJECT_ID,
   ),
   output_filename=pipeline_config_file_name)
_ = runner.compile(pipeline, write_out=True)

schedule.create_from_pipeline_file(
   pipeline_path=pipeline_config_file_name,
   schedule='0 5 * * 1',  # Monday's at 5 am
   project_id=PROJECT_ID,
   region=REGION,
   time_zone='America/Los_Angeles',
   parameter_values={}
)
```

Machine Learning Pipelines at Startups

- Managed pipelines reduce DevOps needs
- Reduced expenses due to managed pipelines instead of 24/7 running clusters
- Automated model updates free up time of ML Engineers
- Consistent model updates across ML projects
- One-stop place for ML related data
- Automated audit tracking

Conclusion

Managed Pipelines

- Save time
- Save Money
- Reduce the burden on data scientists and ML engineers
- Force consistency across ML projects in your company

Conclusion

Investment in ML Engineering pays off

- Time to first model reduced from weeks to days
- Managed Pipelines reduce costs of ML projects
- Tooling provided consistent ML workflows
- Models are comparable
- Models are reproducible
- Processes are repeatable

Google Cloud

Thank you.

DIGITS

digits.com

WEB

@digits

CONTACT

Google Cloud

Google Cloud Summit

Thank you for joining

Timeline

Provision GKE Cluster for

Set up GKE cluster for self-hosted ML pipelines

Integrate GPUs in GKE

Install Kubeflow Pipelines

Item two

Quam lacus suspendisse faucibus interdum posuere lorem ipsum dolor sit.

Item three

In nulla posuere sollicitudin aliquam ultrices sagittis orci.