

intuit



turbotax



quickbooks



mint

# Managing ML Models @ Scale

Intuit's ML Platform

Srivathsan Canchi  
Tobias Wenzel



Who we serve

Consumers  
Small businesses

Self-employed

Great refund! \$2,880

**intuit**

**turbotax**   **quickbooks**   **mint**

# ML in Action

# ML driven experiences



investment interest expenses

investment interest expenses

margin interest investment expense

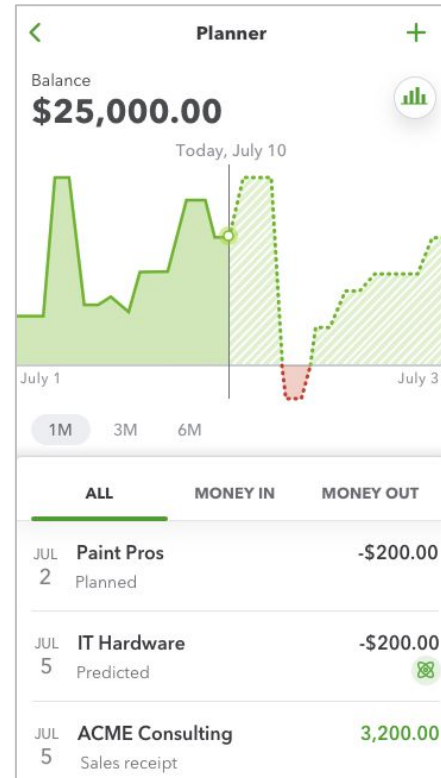
interest expense investment

investment interest expense carryover

Where do I enter a 1099-G for a state or local tax refund?  
Important: If your 1099-G shows an amount in Box 1, click or tap here for alternate instructions, as the steps below won't...

Where do I enter Form 1099-S?  
You may get a 1099-S if you sold your home, a rental property, stock in a co-op or any other real estate, including land,...

**Self-help**  
in TurboTax



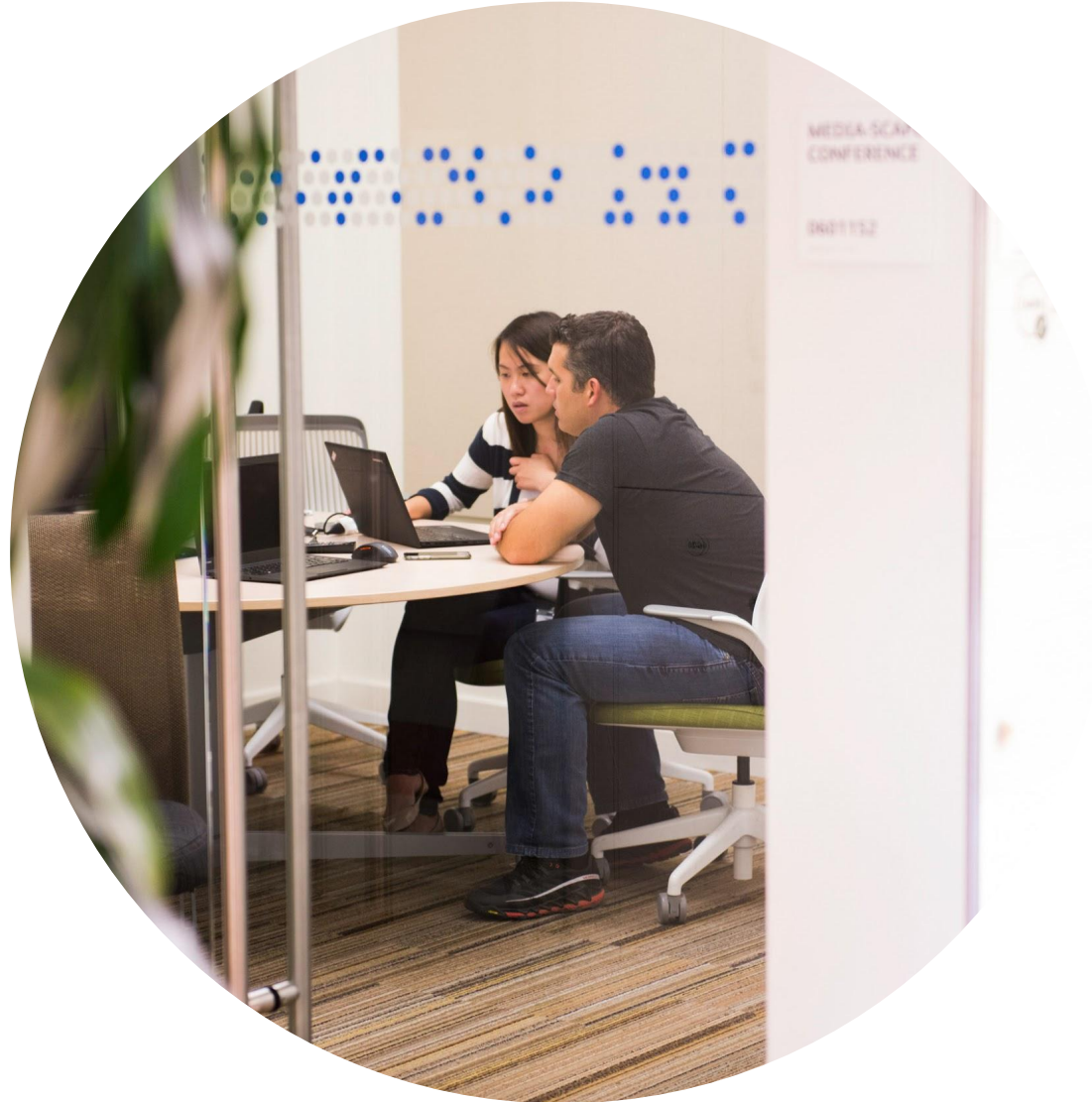
**Forecasting cash flow**  
in QuickBooks



Description	Category	Amount
Meet Fresh	Food & Dining	-\$9.01
Portforpits Ca	Shopping	-\$188.00
Eataly Birreria	Food & Dining	-\$30.00
Margaret Kling Maison	Personal Care	-\$74.75
Groupecdrem Ca	Shopping	-\$209.60
Menchies Belmont Vill	Food & Dining	-\$3.29
Sharetea	Food & Dining	-\$8.16

**Automatic categorization**  
in Mint

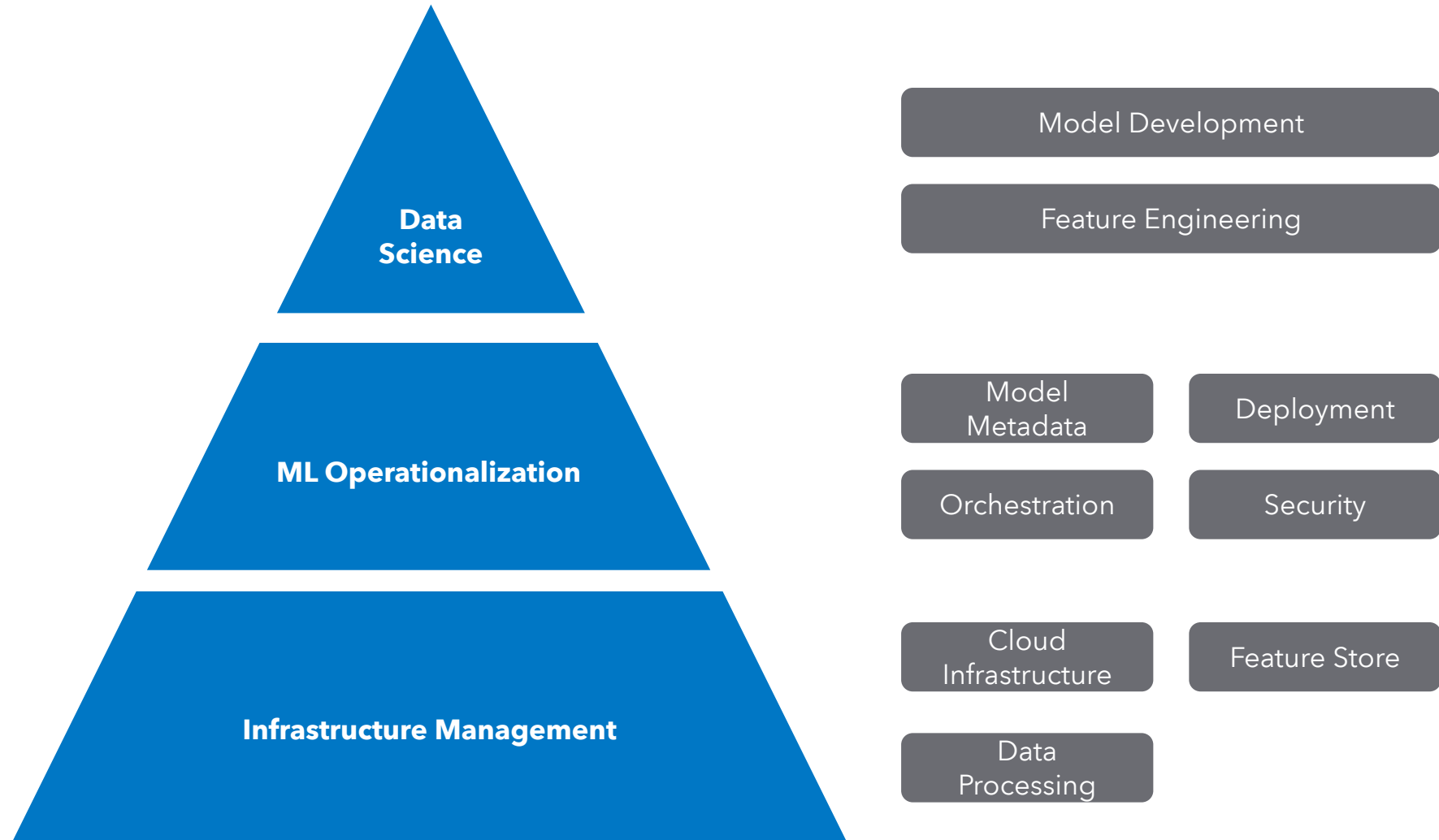
**Managing Models in  
Production is hard**



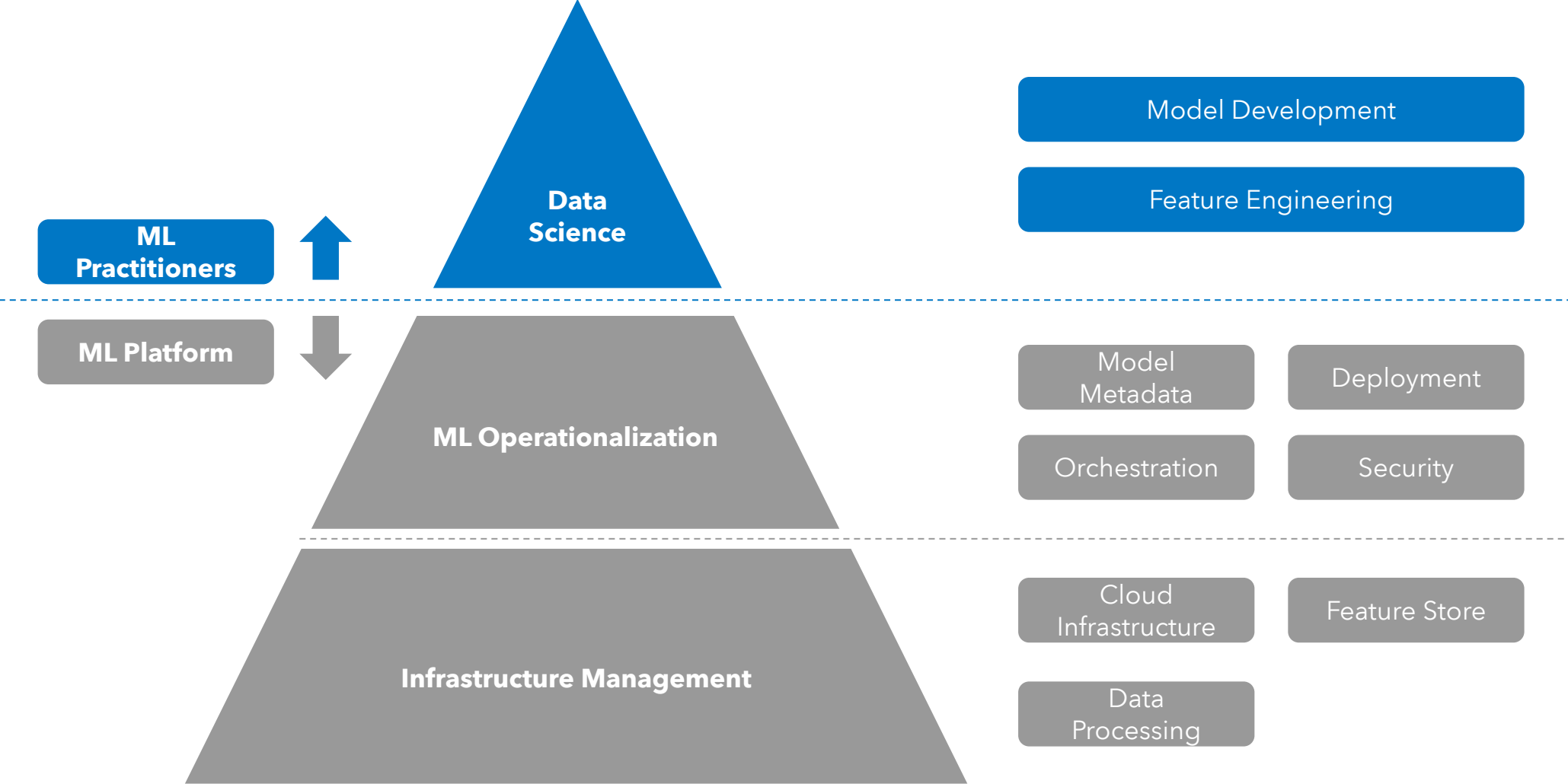
We spend more time bringing the model to production than developing and training it

– Data Scientists, 2018

# “Hidden Debt” of ML



# How can ML Practitioners focus on their craft?





# Principles of Intuit ML Platform



**Collaboration**



**Self-Service**



**Security**

# Powering insights and experiences with ML



## Data Scientist

*"I can quickly train and deploy models that improve the customer experience"*



## ML Engineer

*"I can quickly iterate and make models performant"*



## Data Analyst

*"I can quickly train and deploy models that inform stakeholders how the business is doing"*



## Product Manager

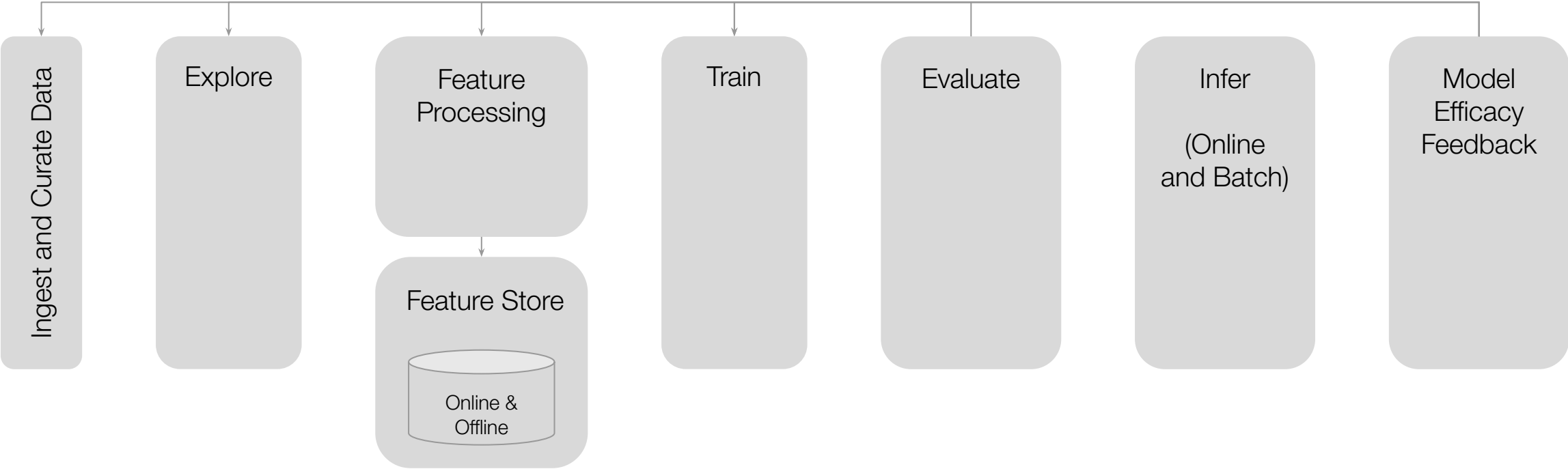
*"I can easily run experiments that rely on models to get a customer benefit"*



## Marketer

*"I can easily target particular users that have a certain characteristics"*

# Generic Model Lifecycle



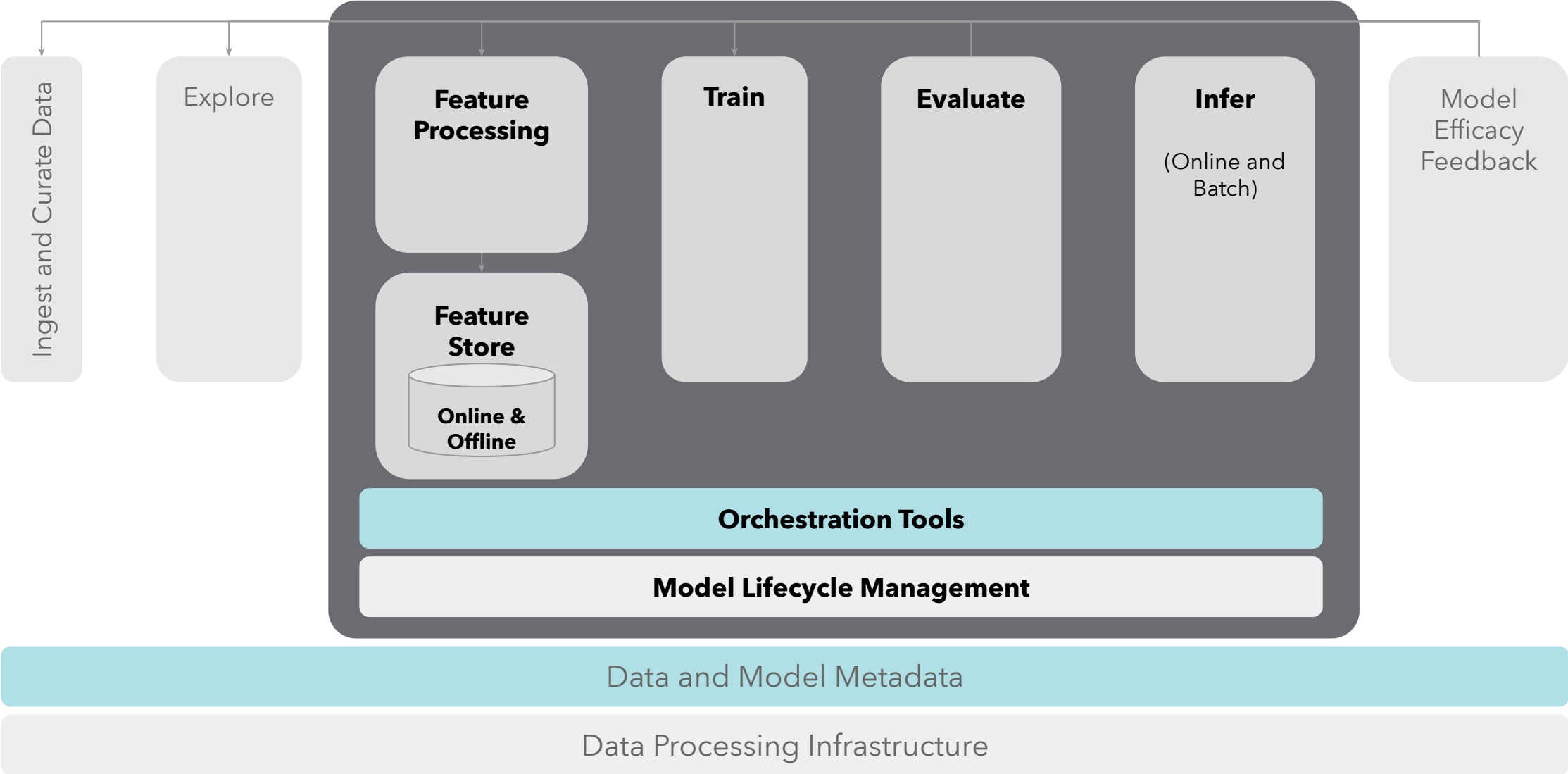
Orchestration Tools

Model Lifecycle Management

Data and Model Metadata

Data Processing Infrastructure

# Intuit's ML Platform Today



# Technologies in use



AWS SageMaker



Kubernetes

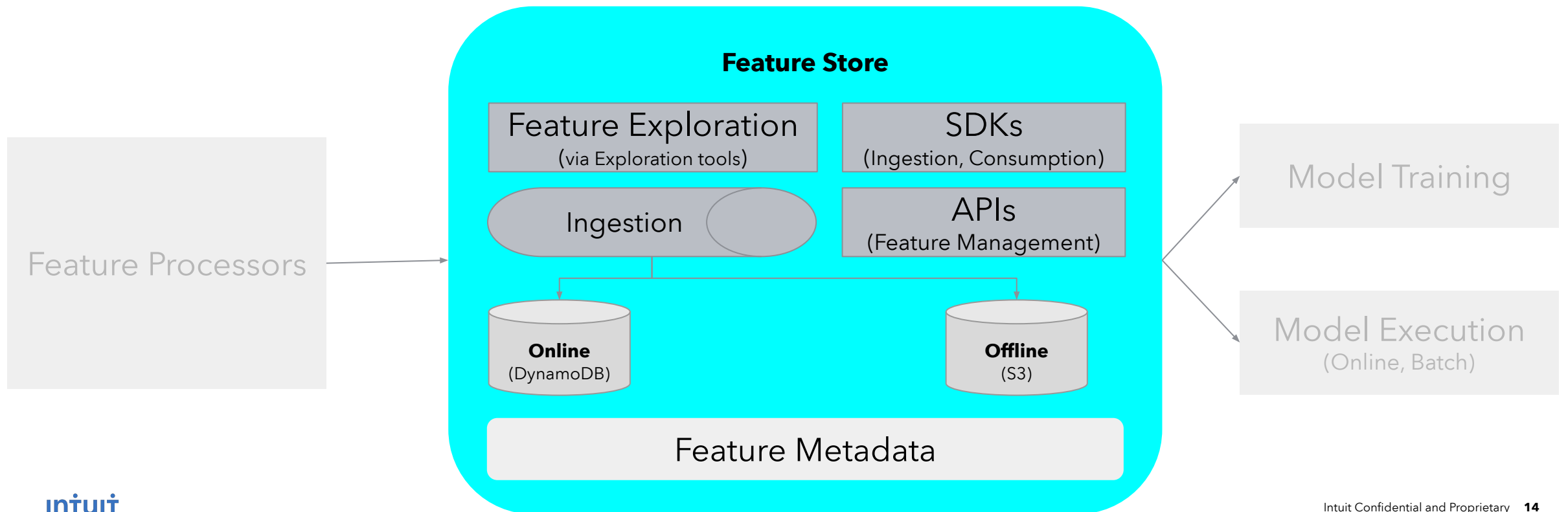


Argo

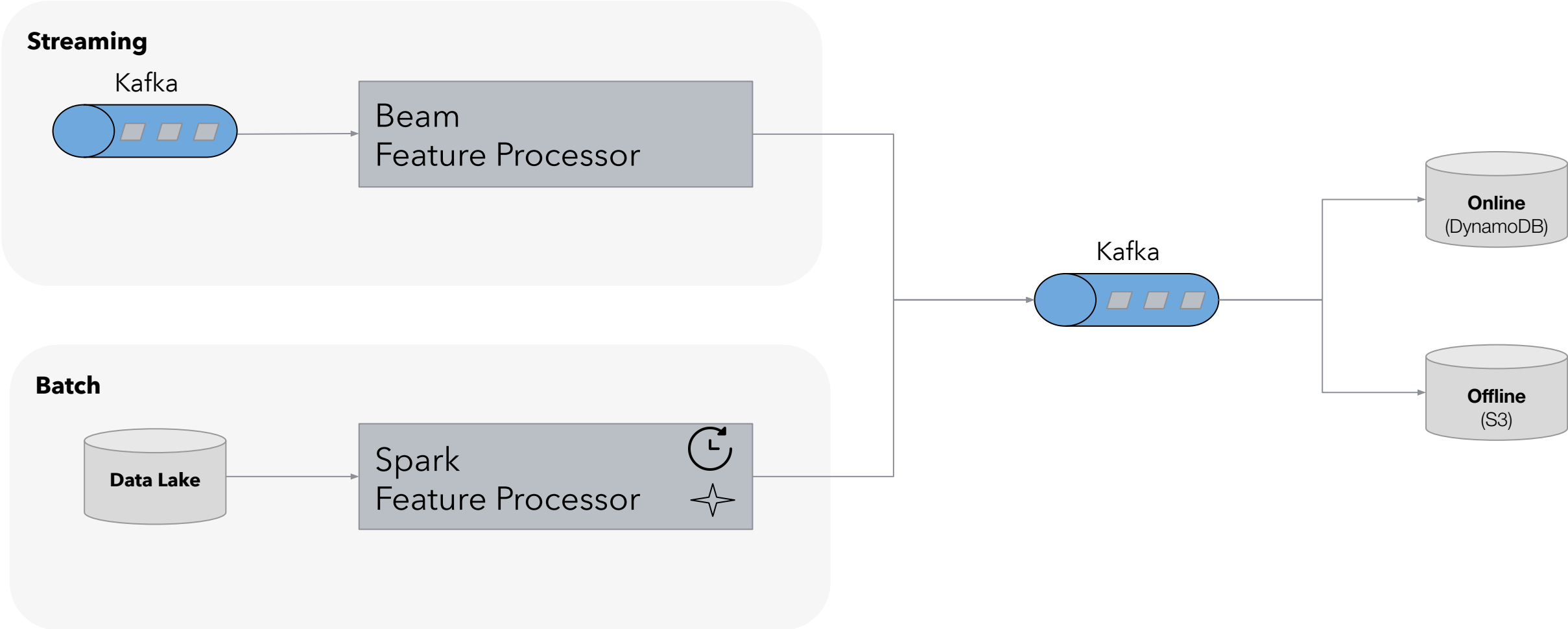


# Feature Store

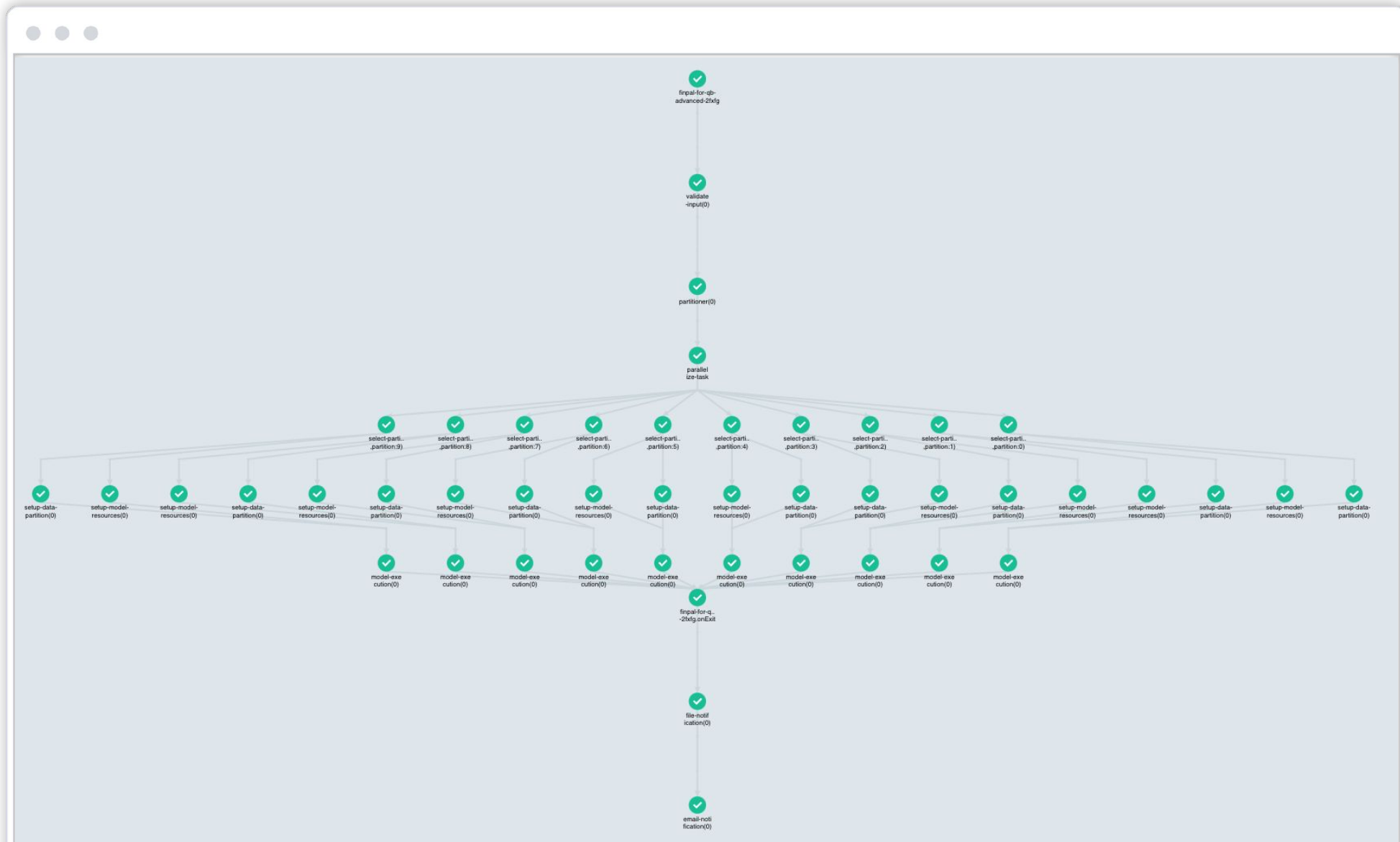
- Store Features in offline (durable) and online (transactional) stores
- Metadata about Features
- Find and re-use Features
- Feature access during model inference and model training



# Feature Processing



# Model Training

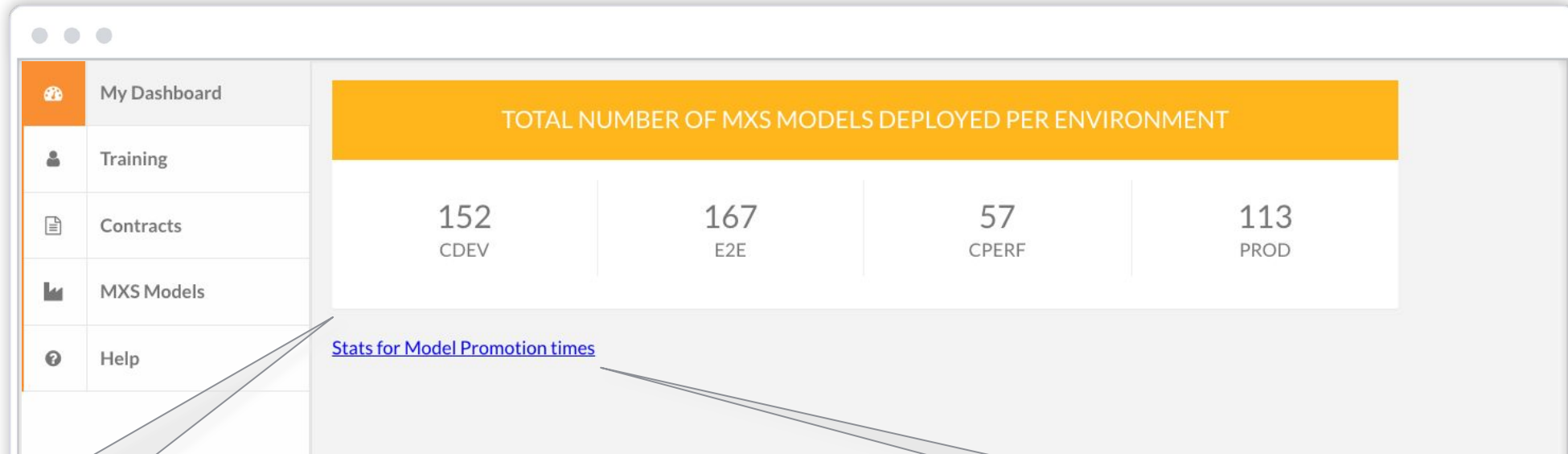




# Model Training

```
1 pipeline_type: sagemaker-training
2 pipeline_version: master
3 artifact_url: test.docker.repo/creditcard-approval-model/
4 project_name: creditcard-approval-model
5 kms_key_arn: arn:aws:kms:us-west-2:1234567890:key/key
6 email: firstNameLastName@intuit.com
7 training_data_config:
8   data_uri: s3://creditcard-approval-input/data/training/
9   content_type: csv
10 validation_data_config:
11   data_uri: s3://creditcard-approval-input/data/validation/
12   content_type: csv
13 additional_data_configs:
14   - data_config_name: test
15     data_uri: s3://creditcard-approval-input/data/validation/
16     content_type: csv
17 output_bucket_uri: s3://creditcard-approval-output/
18 tags:
19   - Key: intuit:billing:fp
20     Value: 3po4uih3-9084h49iu3fn3-894n
21 resource_config:
22   training:
23     instance_count: 1
24     instance_type: ml.m5.xlarge
25     max_run_time_sec: 86400
26     ebs_volume_gb: 1
27
```

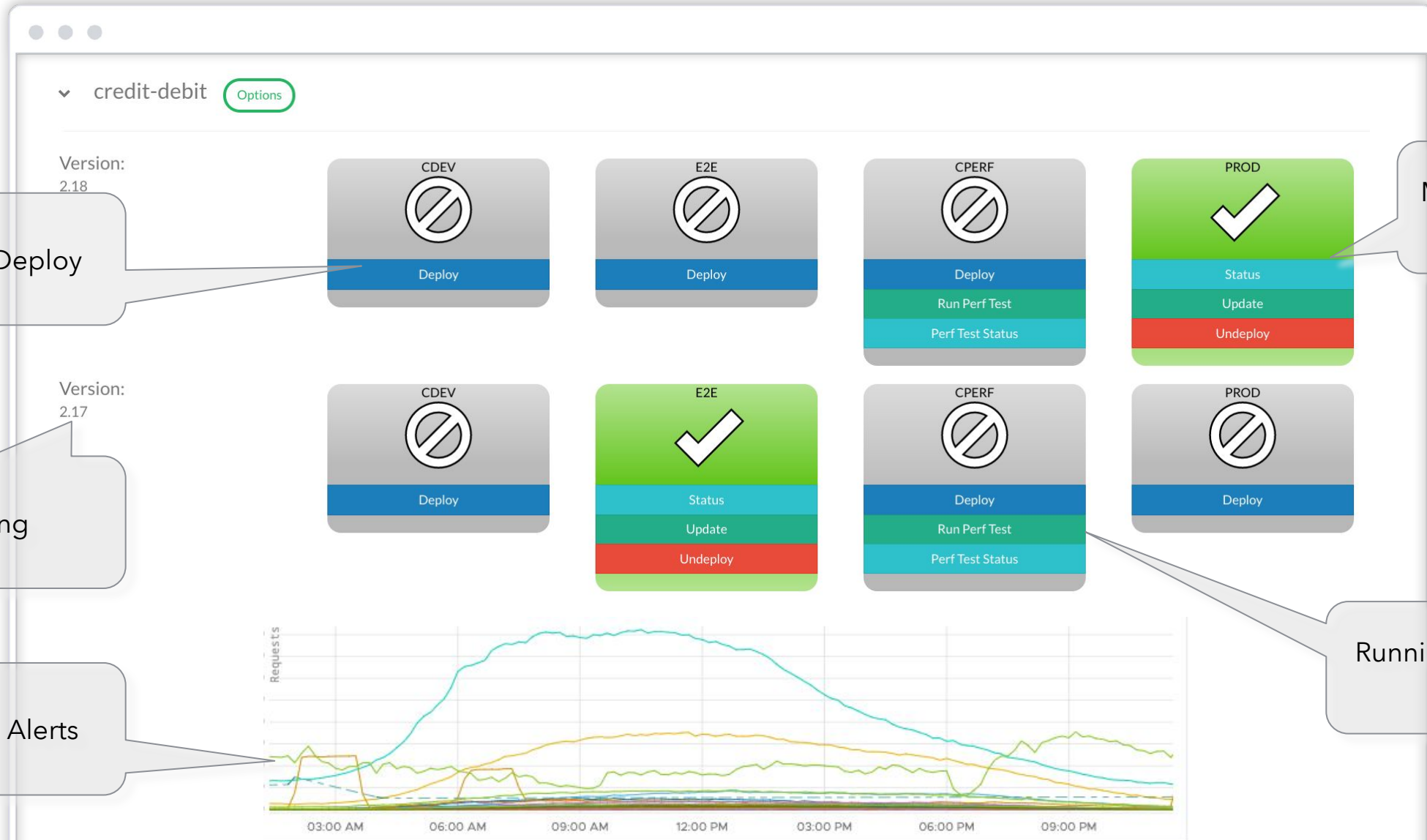
# Self Service Interfaces



non-prod environments for customers

Statistics about productivity increases

# Automate the simple things



Single-click Deploy

Versioning

Live Metrics, Alerts

Model's Runtime Status

Running Performance Tests

# With great speed, comes cost

Visibility and Self-service

Platform success = Lots of experiments

## Cost Transparency

- Upfront pricing
- Cost dashboards

## Cost Assignment

- Tagging
- Business Unit assignments
  - Platform ensures correct BU chargebacks

The screenshot displays a management interface for test environments. On the left is a table with a 'Name' column and a dropdown arrow. The table lists ten environments, each with a radio button for selection. On the right, configuration panels are shown for the selected environment, 'E2E-test-rlgnjb-6'. The 'Deploy Options' section shows 'Fixed Scale' selected with a radio button. The 'Fixed Scale Setting' section includes an 'Instance Count' input field set to '10' and a yellow badge indicating a 'Monthly Cost : 5,418.00'. Below this, another 'Deploy Options' section shows 'Auto Scale' selected. The 'Auto Scale Setting' section includes 'Min Instances' (input '1') and 'Max Instances' (input '10'), with yellow badges for 'Min Monthly Cost : 541.80' and 'Max Monthly Cost : 5,418.00'.

Name
<input type="radio"/> E2E-testitay-1
<input type="radio"/> E2E-simtranstest1-1
<input type="radio"/> E2E-test-rlgnjb-6
<input type="radio"/> CDEV-ke-ml-testdeploy-1
<input type="radio"/> CDEV-testflight-1
<input type="radio"/> E2E-smartlist-test-1
<input type="radio"/> CDEV-smartlist-test-1
<input type="radio"/> E2E-test-content-classifier-1
<input type="radio"/> CDEV-or-test-1
<input type="radio"/> CDEV-new-test-for-iks-1
<input type="radio"/> CDEV-coa-configurator-test-2

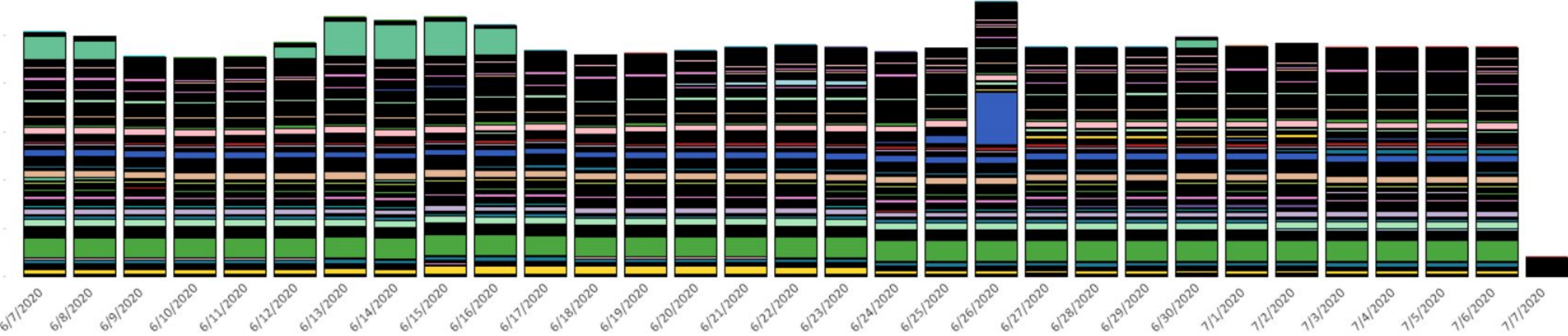
Deploy Options: Fixed Scale  Auto Scale

Fixed Scale Setting: Instance Count: 10 Monthly Cost : 5,418.00

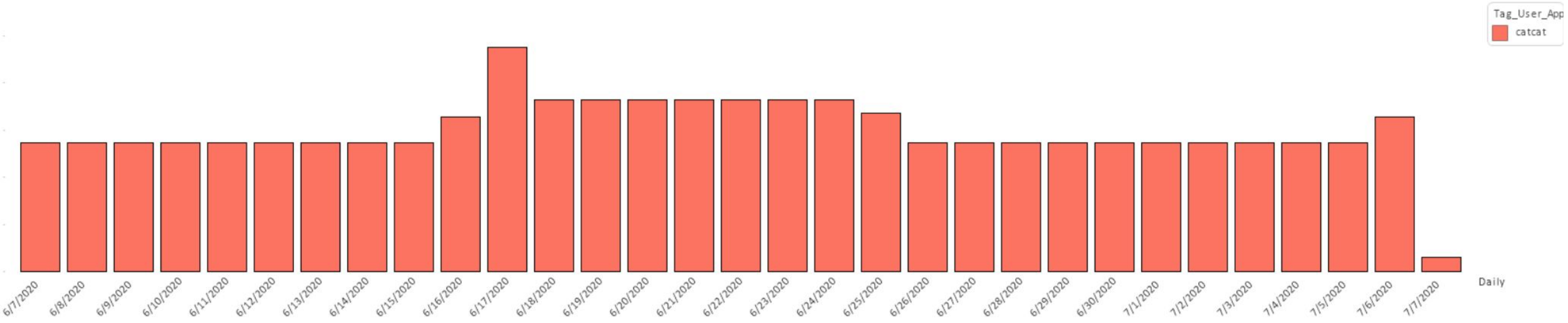
Deploy Options: Fixed Scale  Auto Scale

Auto Scale Setting: Min Instances: 1 Max Instances: 10 Min Monthly Cost : 541.80 Max Monthly Cost : 5,418.00

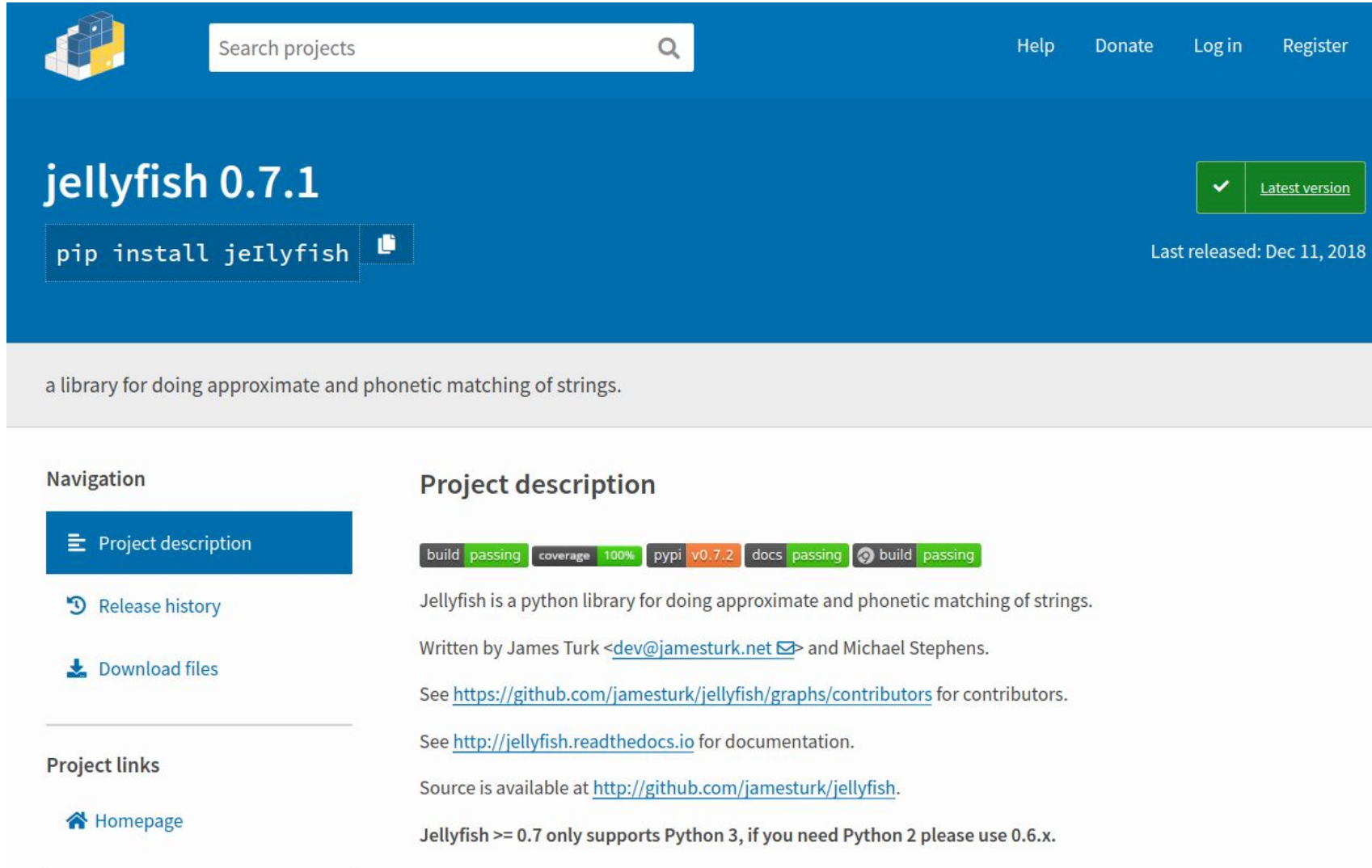
# Cost transparency



# Cost transparency



# What can go wrong?



The screenshot shows the PyPI page for the 'jellyfish' package. At the top, there is a search bar and navigation links for Help, Donate, Log in, and Register. The package name 'jellyfish 0.7.1' is prominently displayed, along with a 'Latest version' button and the release date 'Last released: Dec 11, 2018'. A code block shows the installation command 'pip install jeIlyfish'. Below this, a description states it is 'a library for doing approximate and phonetic matching of strings.' The page is divided into sections: 'Navigation' with links for Project description, Release history, and Download files; 'Project links' with a link to the Homepage; and 'Project description' which includes a status bar (build passing, coverage 100%, pypi v0.7.2, docs passing, build passing), a description of the library, author information (James Turk and Michael Stephens), contributor links, documentation link, source code link, and a note that version 0.7 only supports Python 3.

Search projects

Help Donate Log in Register

## jellyfish 0.7.1

✓ Latest version

```
pip install jeIlyfish
```

Last released: Dec 11, 2018

a library for doing approximate and phonetic matching of strings.

### Navigation

- Project description
- Release history
- Download files

### Project links

- Homepage

### Project description

build passing coverage 100% pypi v0.7.2 docs passing build passing

Jellyfish is a python library for doing approximate and phonetic matching of strings.

Written by James Turk <dev@jamesturk.net > and Michael Stephens.

See <https://github.com/jamesturk/jellyfish/graphs/contributors> for contributors.

See <http://jellyfish.readthedocs.io> for documentation.

Source is available at <http://github.com/jamesturk/jellyfish>.

Jellyfish >= 0.7 only supports Python 3, if you need Python 2 please use 0.6.x.

# Security

ML Models are as vulnerable as any other piece of software

## ML Model Vulnerabilities

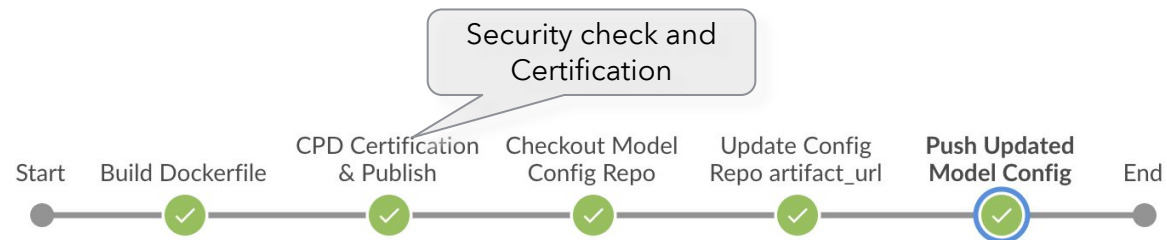
- ML models vulnerable for attacks
- Need security gates at every stage of development

## Quality Gates

- Source image validations
- Standardized framework for container images
- Image hardening, certification and signing

## Security Gates in our AWS setup

- Secure VPC endpoints
- VPC flow logs - monitor traffic
- Security groups - restrict outbound access
- KMS encryption - secure data at rest
- Least privilege IAM roles - secure management of the service





# Compliance

## Complex regulations

- The well known
  - CCPA, GDPR
- More common
  - PCI
- The arcane (very specific to tax compliance)
  - INDOR, NIST

## Platform enforces compliance

- Any customer data that is used in the ML model lifecycle is managed centrally
- Achieving compliance for all models reduces complexity in the process for applications using them
- Automating the necessary controlling actions for the platform ensures that future models remain compliant

# Learnings



## Models as Software

- Accelerated deployment through self service
- Using a GitOps approach allows for declarative development of models
- Production monitoring and alerting
- Security, compliance built-in



## Automated Workflows

- Workflows for standard ML operations help with complexity
- Hosting a workflow engine allows for easy extensibility
- Small threshold for customers starting out with the platform through templates with short configuration



## Curated Feature Store

- Model quality  $\infty$  Feature quality
- Curated feature store offers library of meaningful features
- Search, explore, share features across models enables acceleration of model development

## In Future...

- **Custom ML Operator** to orchestrate and manage ML Resources (Spark, SageMaker Training/Deploy, Feature Store)
- **Declarative Management** of MDLC such as re-training of models, monitoring of features etc.
- **Managed Notebook Service**

# Contact Us



**Tobias Wenzel**

[tobias\\_wenzel@intuit.com](mailto:tobias_wenzel@intuit.com)



**Srivathsan Canchi**

[srivathsan\\_canchi@intuit.com](mailto:srivathsan_canchi@intuit.com)