Machine Learning Platform

Content

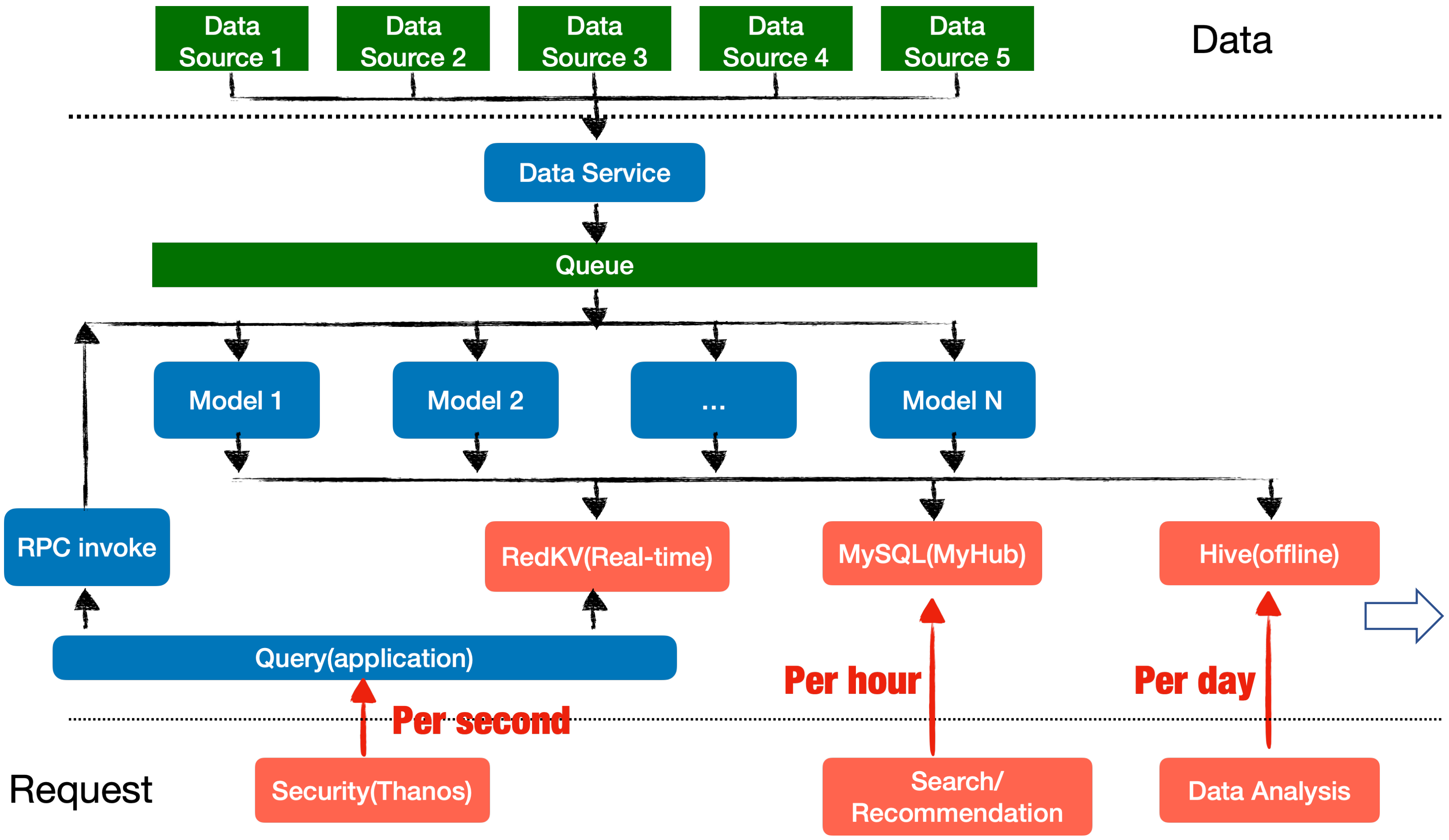
I. Data Service

II. Deployment Service

III. Model Zoo

I. Data Service

Unify data download method



Problem:

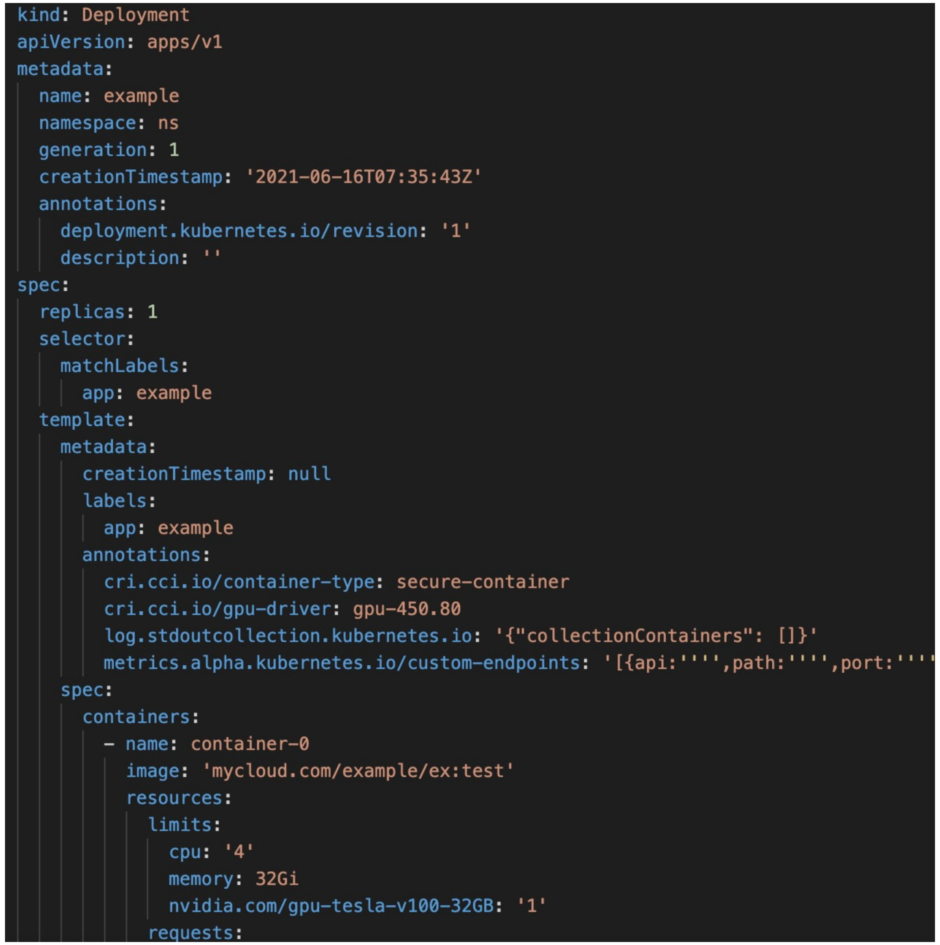
- Different data source
- JSON messages with different format
- Duplicated work among models

Solution:

- Models only need to subscribe queues they needed
- Unify JSON fields

II. Deploment Service

Simplify services deployment and management process



```yaml
kind: Deployment
apiVersion: apps/v1
metadata:
  name: example
  namespace: ns
  generation: 1
  creationTimestamp: '2021-06-16T07:35:43Z'
  annotations:
    deployment.kubernetes.io/revision: '1'
    description: ''
spec:
  replicas: 1
  selector:
    matchLabels:
      app: example
  template:
    metadata:
      creationTimestamp: null
      labels:
        app: example
      annotations:
        cri.cci.io/container-type: secure-container
        cri.cci.io/gpu-driver: gpu-450.80
        log.stdoutcollection.kubernetes.io: '{"collectionContainers": []}'
        metrics.alpha.kubernetes.io/custom-endpoints: '[{api:'''',path:'''',port:''''
    spec:
      containers:
        - name: container-0
          image: 'mycloud.com/example/ex:test'
          resources:
            limits:
              cpu: '4'
              memory: 32Gi
              nvidia.com/gpu-tesla-v100-32GB: '1'
            requests:
```

A simple example of deployment configuration

Problems:

- Complex deployment and management process
- Inconsistent configuration files

Solutions:

- Provide a single model deployment interface
- Uniform online configuration
- Provide services upgrade/rollback interfaces
- Visualize resources utilization

III. Model Zoo

- Provide a single COS interface to upload / download models
- Upload models after training to prepare the following inference service