Scheduler as a Service

EA Digital Platform, Data & AI

Nitish Victor
Preethi Ganeshan
Xiaojin Zhu
Nitish Victor
Software Engineer II

nvictor@ea.com
WE EXIST TO
INSPIRE THE WORLD TO PLAY
Dataset in Petabytes

10s of Terabytes of data generated every day

Thousands of ETL Jobs

38 game studios spread across the world

Scale at EA
Data at EA
Data at EA

- Game Studios with Embedded Analyst Teams
- Highly sensitive datasets
- Multiple data pipelines from a central data warehouse.
- Varied SLA requirements
- 1000s of ETL jobs and ~50k ad-hoc queries everyday using our platform
## Scheduler at EA

<table>
<thead>
<tr>
<th>Before Airflow</th>
<th>With Airflow</th>
<th>Future of Airflow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oozie/Cron/Scripts</td>
<td>Central Airflow</td>
<td>Multiple Airflow scheduler</td>
</tr>
<tr>
<td>Custom Continuous Integration (CI)</td>
<td>Common Plugins</td>
<td>Self-Serve Orchestration System</td>
</tr>
<tr>
<td>In-House Deployment Tool</td>
<td>Standard Monitoring</td>
<td>Improved workload management</td>
</tr>
<tr>
<td>Basic monitoring</td>
<td>CI/CD</td>
<td></td>
</tr>
<tr>
<td>Basic Auth Model</td>
<td>SSO integrated</td>
<td></td>
</tr>
<tr>
<td>Standalone Systems</td>
<td>Multi-tenancy</td>
<td></td>
</tr>
</tbody>
</table>
Scheduler-as-a-Service Objectives

- Multi-tenancy with RBAC (SSO with Custom Auth Middleware)
- Ease of job deployment (Gitlab CI/CD)
- Ease of interfacing with the dataset (Custom Plugins with ACL)
- Monitoring and Alerts as part of service
Preethi Ganeshan
Software Engineer III

pganeshan@ea.com
Multi-Tenancy and RBAC
Multi-Tenancy

Airflow Clusters
- Airflow Web Pods
- Airflow Workers

Components:
- Nginx SSO Auth
- S3 Sync Sidecar
- Message Broker/Queue
- Backend DB
- Scheduler

Teams and Repositories:
- Team A
  - Repo A
- Team B
  - Repo B
- Team C
  - Repo C

S3 Bucket
RBAC as a Service
RBAC as a Service
Job Deployment
Auto Deployment – S3 Sync Sidecar

- Need a way to sync DAGs across multiple pods whenever there is new or updated DAG
- Sidecar runs alongside Web, Workers and Scheduler Pods
- DAG files and Plugins synced from S3 location periodically
Sandbox as a Service

Custom "Dev" Docker Image

Gitlab merge triggers CI/CD pipeline

Developers expected to test their jobs by running it on Airflow docker image locally. The DAG is tested on the "Dev" data warehouse

Custom Airflow Image with "Dev" data warehouse access and OAuth Integration provided to DAG developers
Xiaoqin Zhu
Software Engineer III

xzhu@ea.com
Custom Plugins

Ease of accessing the datasets
Plugin Library as a Service

Plugins are managed through a single repository

Requires merge request and approvals to contribute

Plugins are deployed through continuous deployment

In-house Plugins to interface with Presto Platform
Custom Plugins with ACL

Plugin access defined per team in JSON

Gitlab CI Tests to enforce access control

New approved plugins added to config
Monitoring and Alerts
Alerting as a Service

- Kubernetes provides a base level of fault tolerance
- Pod level metrics and alerts available with Prometheus
- Email based notifications
- Airflow Service level metrics and alerts
- Job/DAG level monitoring
- ELB alerts and monitoring
Monitoring as a Service
Monitoring as a Service
Architecture and Integration
Airflow Infrastructure

- Deployed on Kubernetes
- Celery Executor - Redis broker
- PostgreSQL backend
- Proxy Based Authentication
- Multi-tenant system
- Centralized plugin system to interface with multiple Presto cluster ecosystem.