Scheduler as a Service

Apache Airflow

EA Digital Platform, Data & AI

Nitish Victor
Preethi Ganeshan
Xiaoqin Zhu

Nitish Victor
Software Engineer II

nvictor@ea.com

WE EXIST TO

INSPIRE THE WORLD TO PLAY

One Company
Multiple Game Studios

BioWare
Austin | Edmonton

chillingo

Criterion

DICE
Stockholm | Los Angeles

CAPITAL GAMES

Firemonkeys

FROSTBITE

INDUSTRIAL TOYS

EA KOREA STUDIO

maxis

MOTIVE
Vancouver

pogo
Shanghai | Redwood Shores

PopCap
Seattle | Shanghai | Hyderabad

RED CROW

REDWOOD STUDIOS

Respawn ENTERTAINMENT

TIBURON

TRACK TWENTY

EA Digital Platform

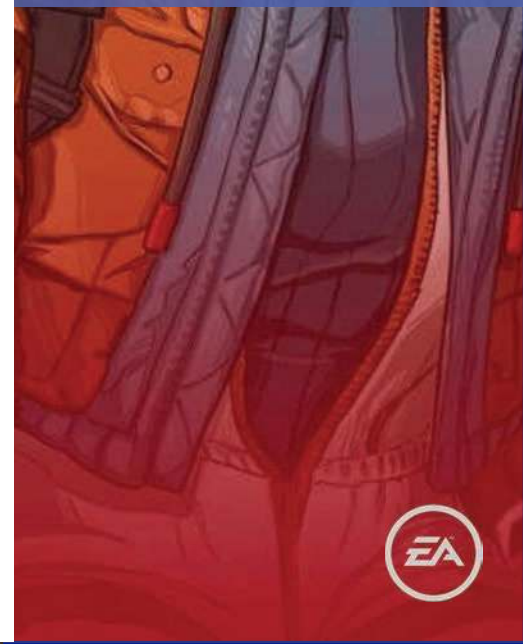
Data & AI

Player Network Framework

One Central Data Platform Team

Support Multiple Game Studios

Data Engineers & Scientists



WHO WE ARE

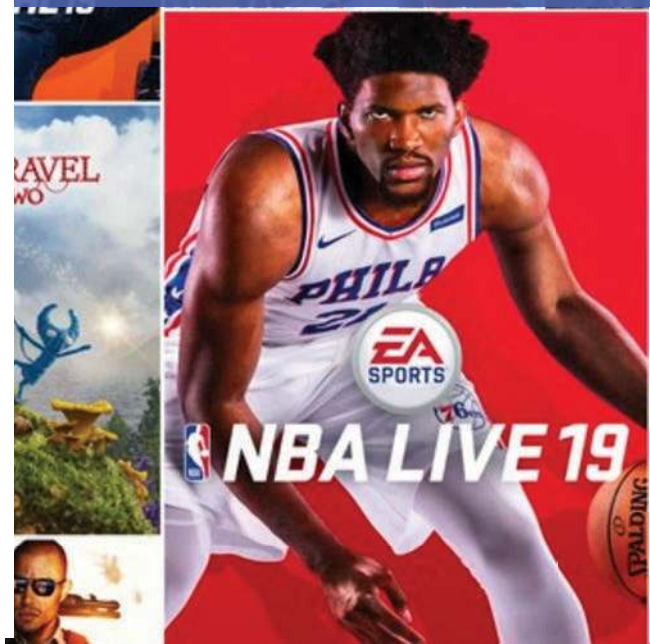
Dataset in Petabytes

10s of Terabytes of data generated every day

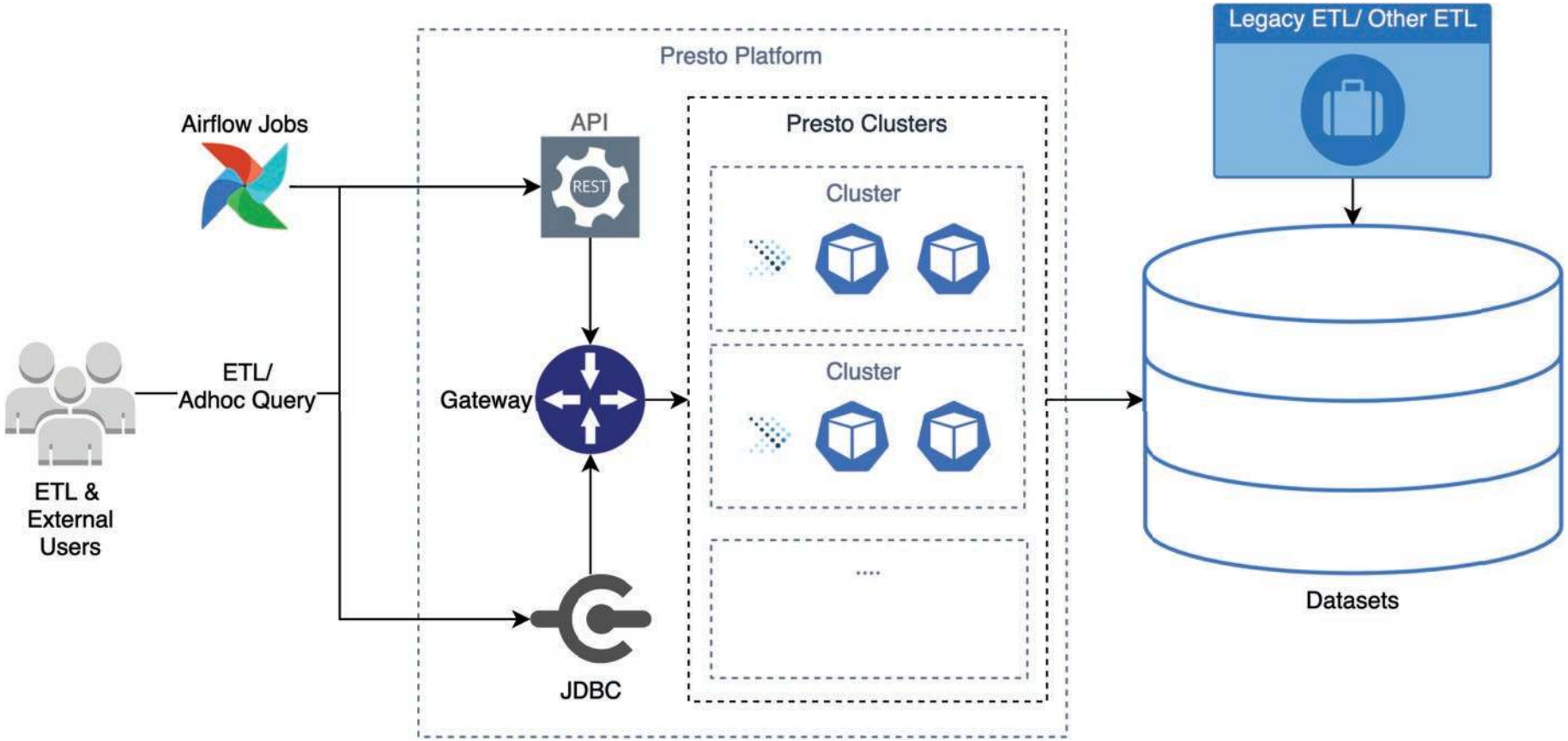
Thousands of ETL Jobs

38 game studios spread across the world

Scale at EA



Data at EA


Presto Platform
Airflow Jobs
API
REST
Presto Clusters
Cluster
Cluster
....
Legacy ETL/ Other ETL
ETL/
Adhoc Query
Gateway
ETL &
External
Users
JDBC
Datasets

Data at EA

Game Studios with Embedded Analyst Teams

Multiple data pipelines from a central data warehouse.

1000s of ETL jobs and ~50k ad-hoc queries everyday using our platform

Highly sensitive datasets

Varied SLA requirements

Scheduler at EA

Before Airflow	With Airflow	Future of Airflow
Oozie/Cron/Scripts	Central Airflow	Multiple Airflow scheduler
Custom Continuous Integration (CI)	Common Plugins	Self-Serve Orchestration System
In-House Deployment Tool	Standard Monitoring	Improved workload management
Basic monitoring	CI/CD	
Basic Auth Model	SSO integrated	
Standalone Systems	Multi-tenancy	

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)

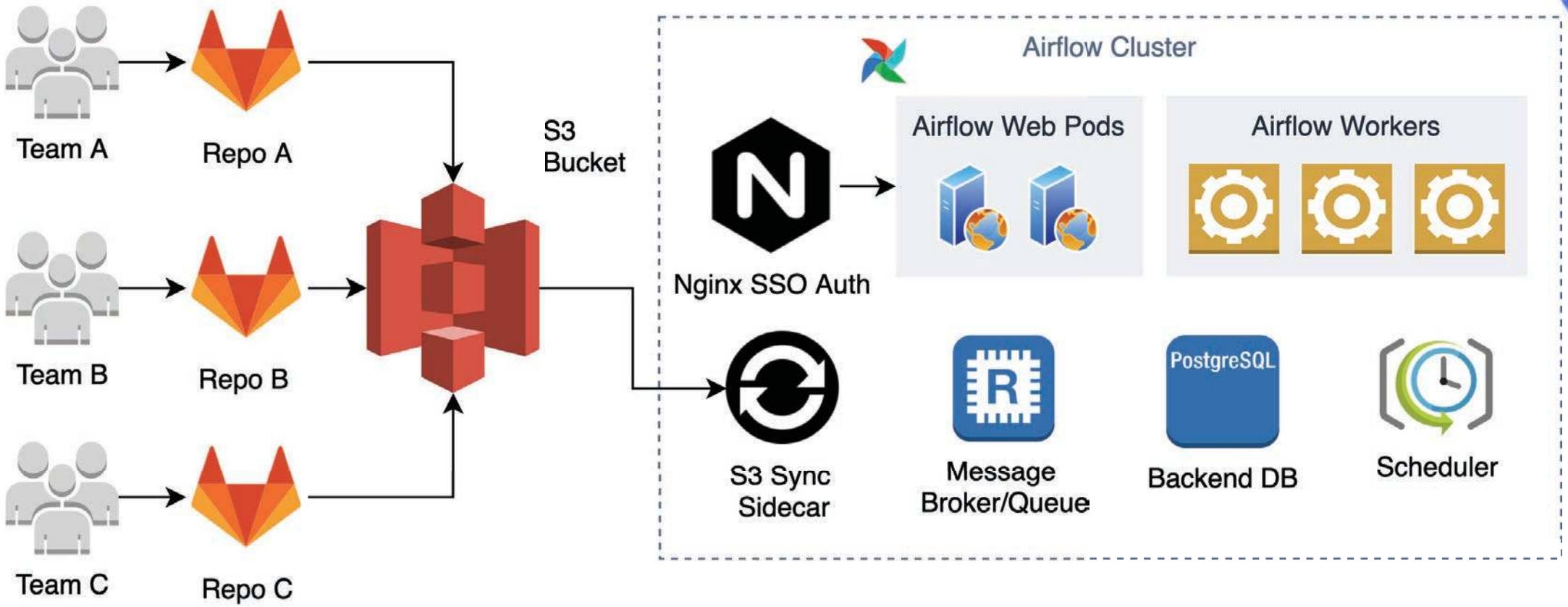
Monitoring and Alerts as part of service

Preethi Ganeshan
Software Engineer III

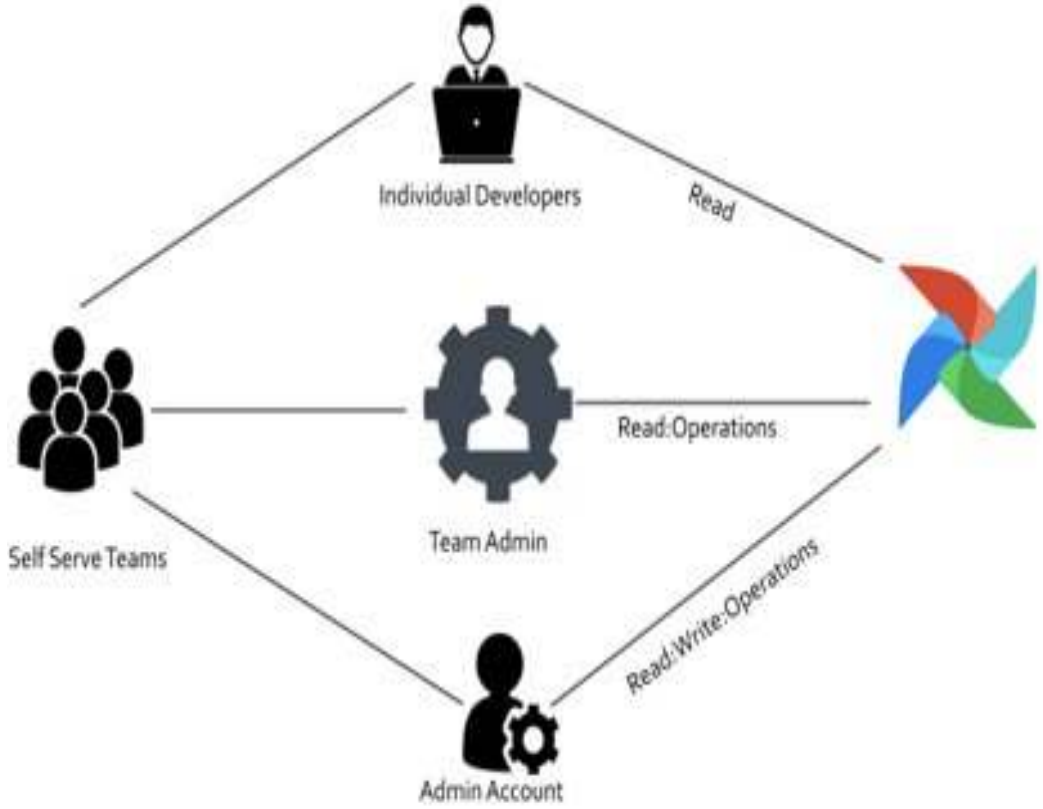
pganeshan@ea.com

Multi-Tenancy and RBAC

Multi-Tenancy



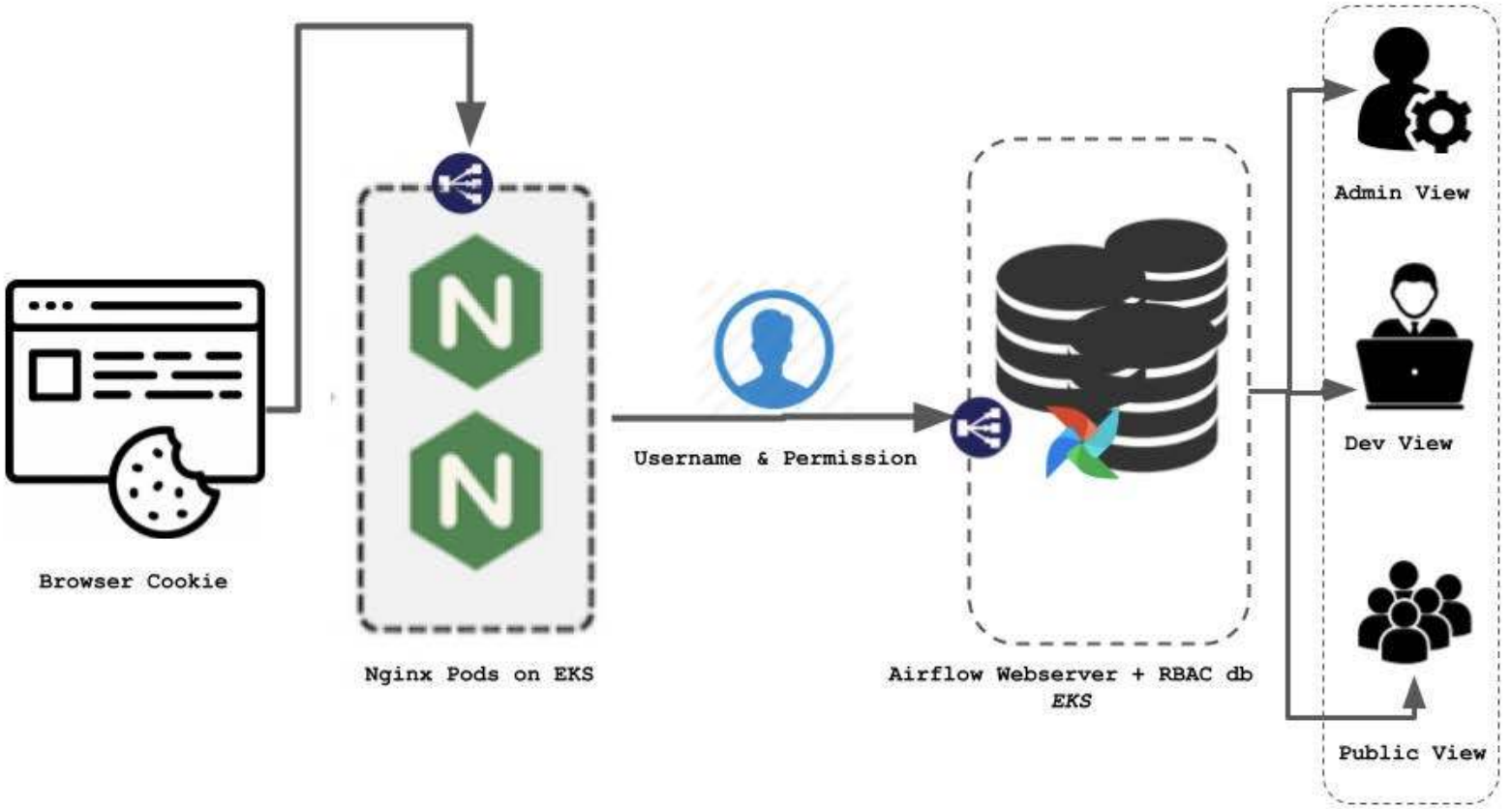
RBAC as a Service



Schedule	Owner	Recent Tasks	Last Run	DAG Runs	Links
@monthly	avatar-ci				
None	avatar-ci	5	2020-05-12 17:36	1	
@daily	avatar-ci				
@daily	avatar-ci				
*****	avatar-ci				
@daily	avatar-ci				

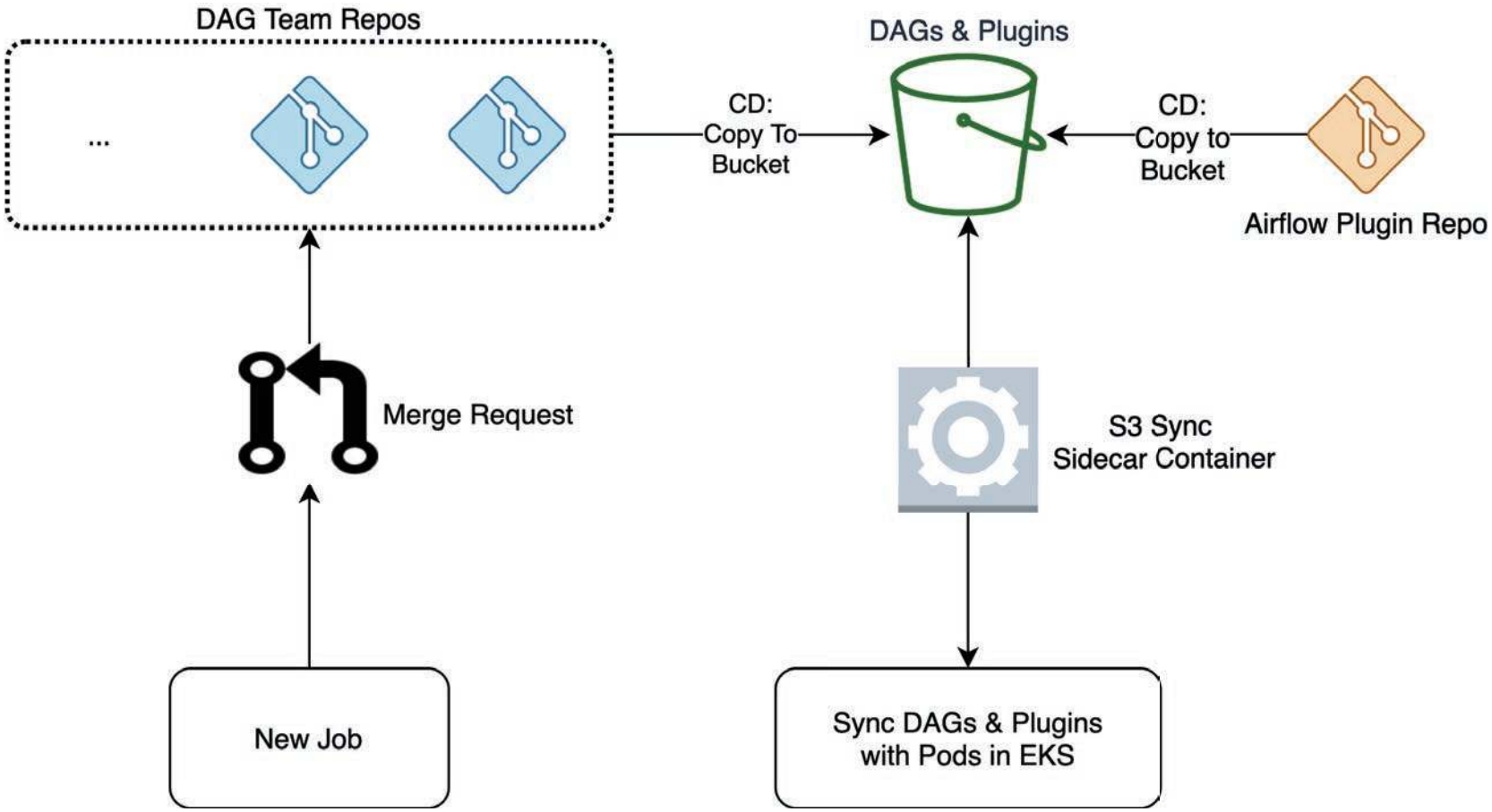
Apache Airflow

RBAC as a Service



Job Deployment

Auto Dev-Ops as a Service



Auto Deployment – S3 Sync Sidecar

```
echo "Syncing DAGs"
output=$(aws s3 sync ${DAG_S3_LOCATION} /tars --region us-xxxxx --delete)
if [ -z "$output" ]; then
    echo "No changes"
else
    echo "Extracting changes"
    mkdir -p dags_staging
```

- 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

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

Gitlab merge triggers CI/CD pipeline

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

```
{
    "description": "Airflow operator access control list",
    "policyDefinitions": {
        "default": {
            "operators": [
                "PostgresOperator",
                "MySqlOperator"
            ]
        },
        "etlWorkflow": {
            "operators": [
                "PostgresOperator",
                "JdbcOperator",
                "HiveOperator",
                "HivePartitionSensor",
                "SlackAPIPostOperator"
            ]
        }
    },
    "teams": {
        "etl": "etlWorkflow",
        "team-demo": "default",
        "internal-team" : "*"
    }
}
```

Plugin access defined per team in JSON

Gitlab CI Tests to enforce access control

New approved plugins added to config

Apache Airflow

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

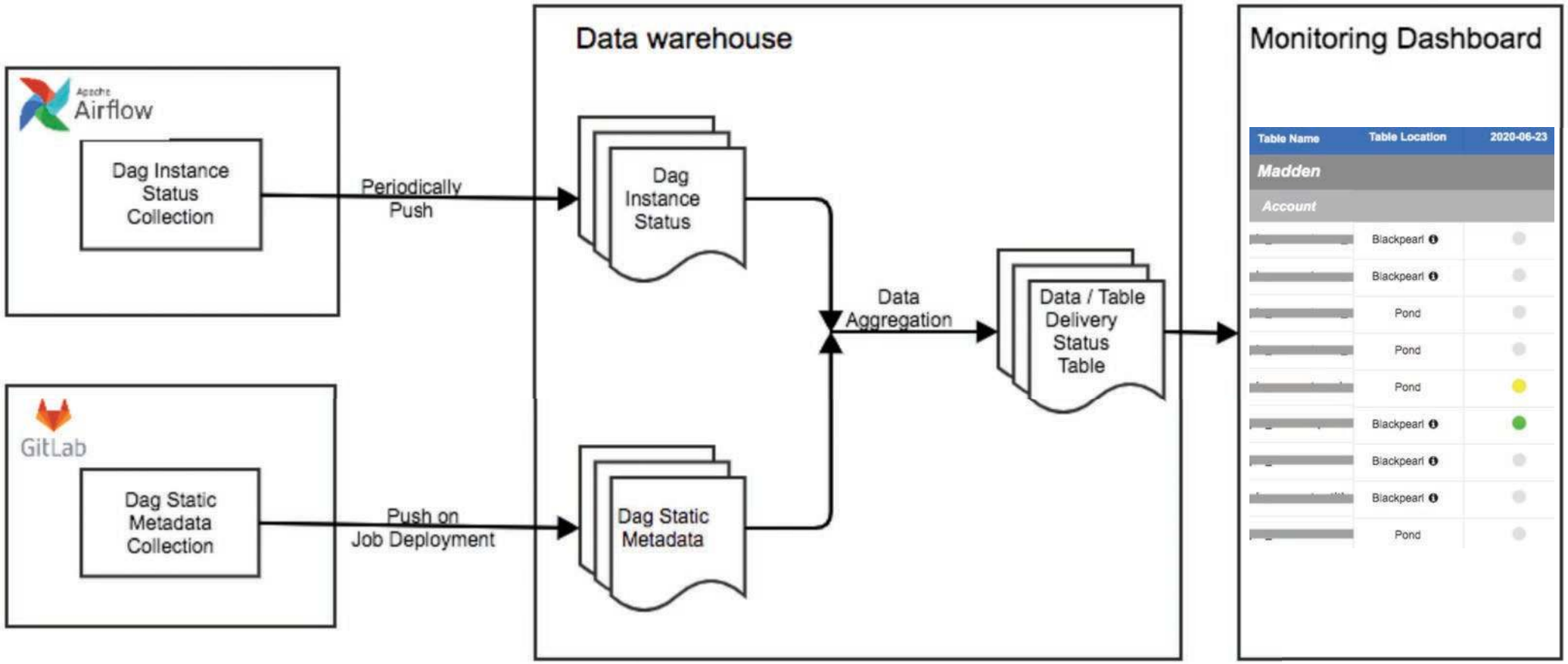
Airflow Service level metrics and alerts

Job/DAG level monitoring

ELB alerts and monitoring

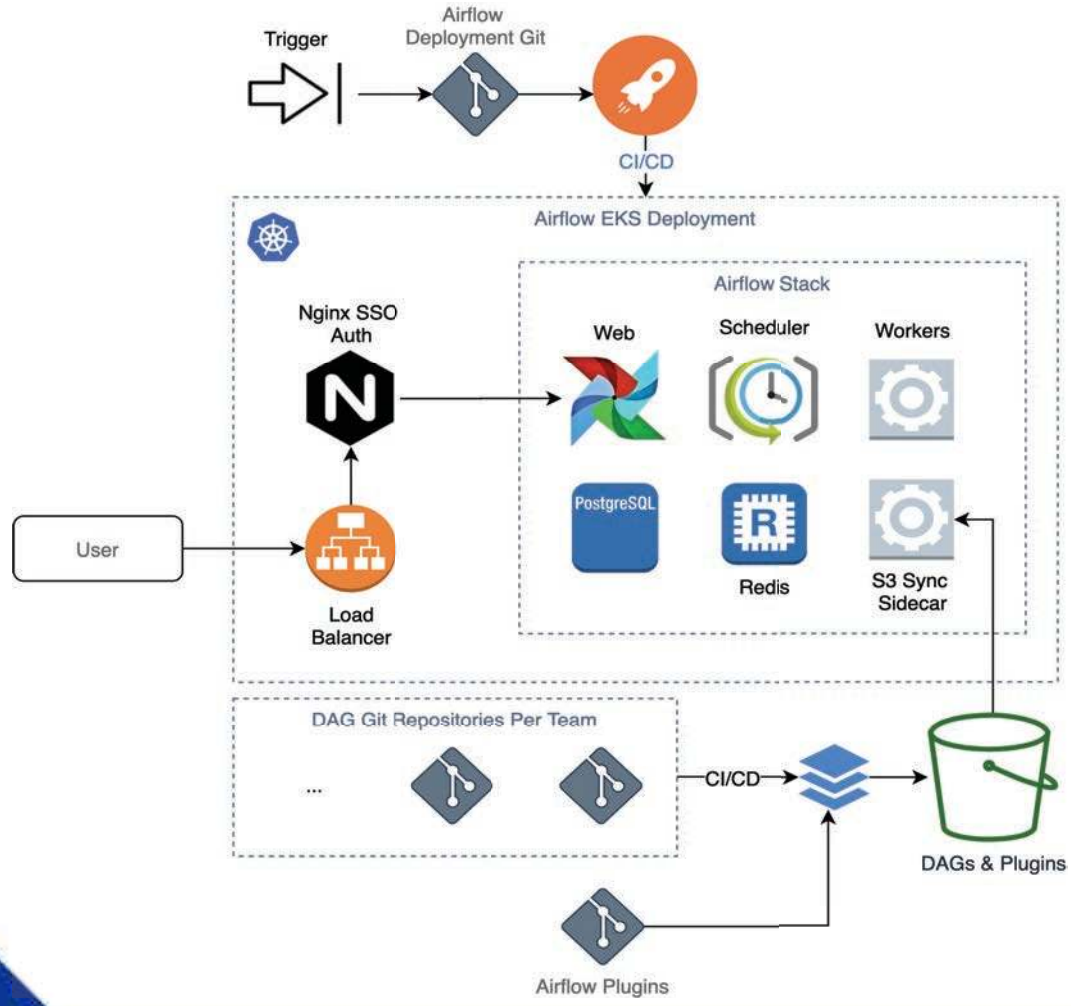
Monitoring as a Service

Monitoring as a Service



Architecture and Integration

Airflow Infrastructure

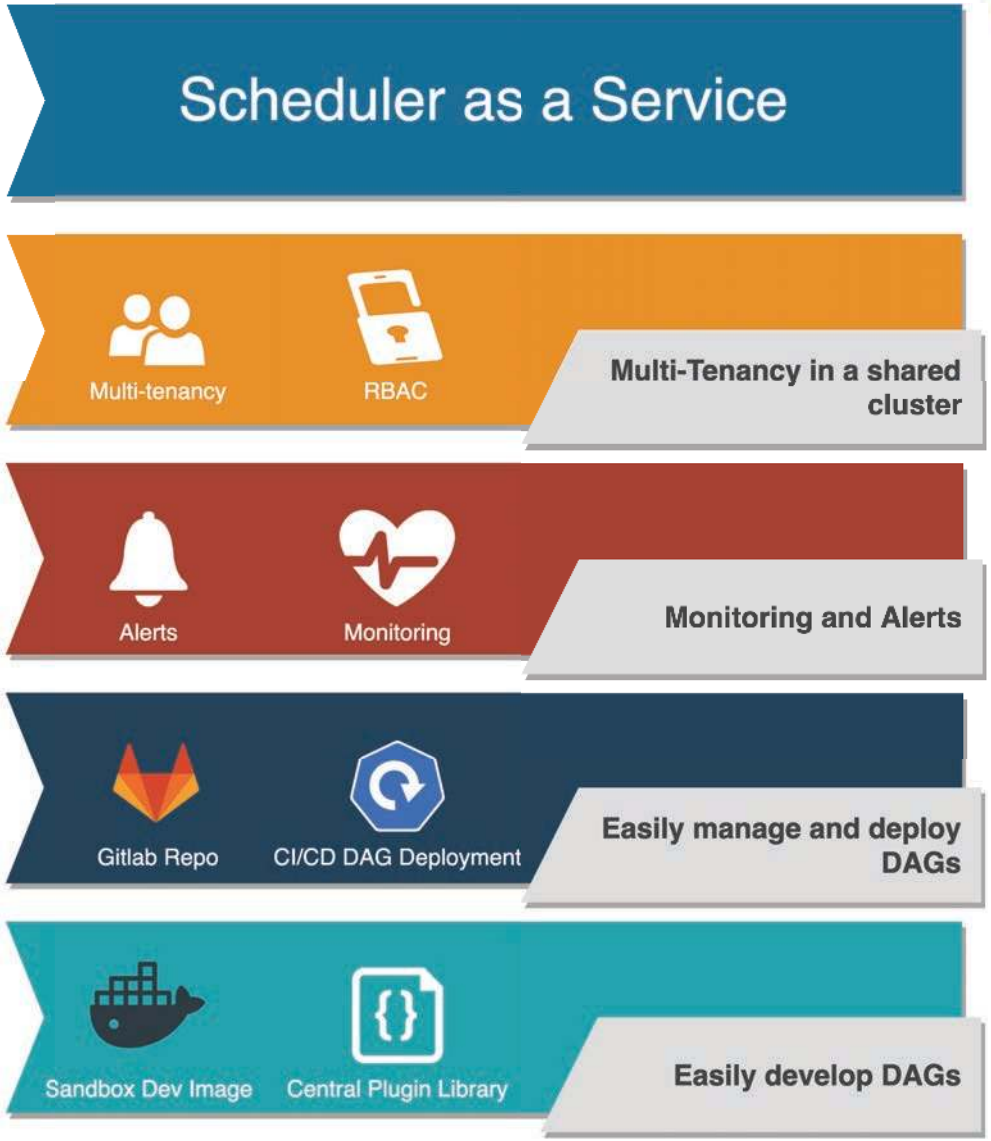


- Deployed on Kubernetes
- Celery Executor - Redis broker
- PostgreSQL backend
- Proxy Based Authentication
- Multi-tenant system
- Centralized plugin system to interface with multiple presto cluster ecosystem.

Team requests access

Gitlab Setup

Provision ACLs




Q & A