



# Ingesting Game Telemetry in near Real-time dynamically into Redshift with Airflow

Karthik Kadiyam





# About Me

Karthik Kadiyam

<https://www.linkedin.com/in/karthik-kadiyam-1a4491bb/>  
kkadiyam@wbgames.com

Lead Big Data Engineer @ WB Games Analytics

Hobbies : Video games, Basketball and Cricket





# Data Landscape Requirements

50+

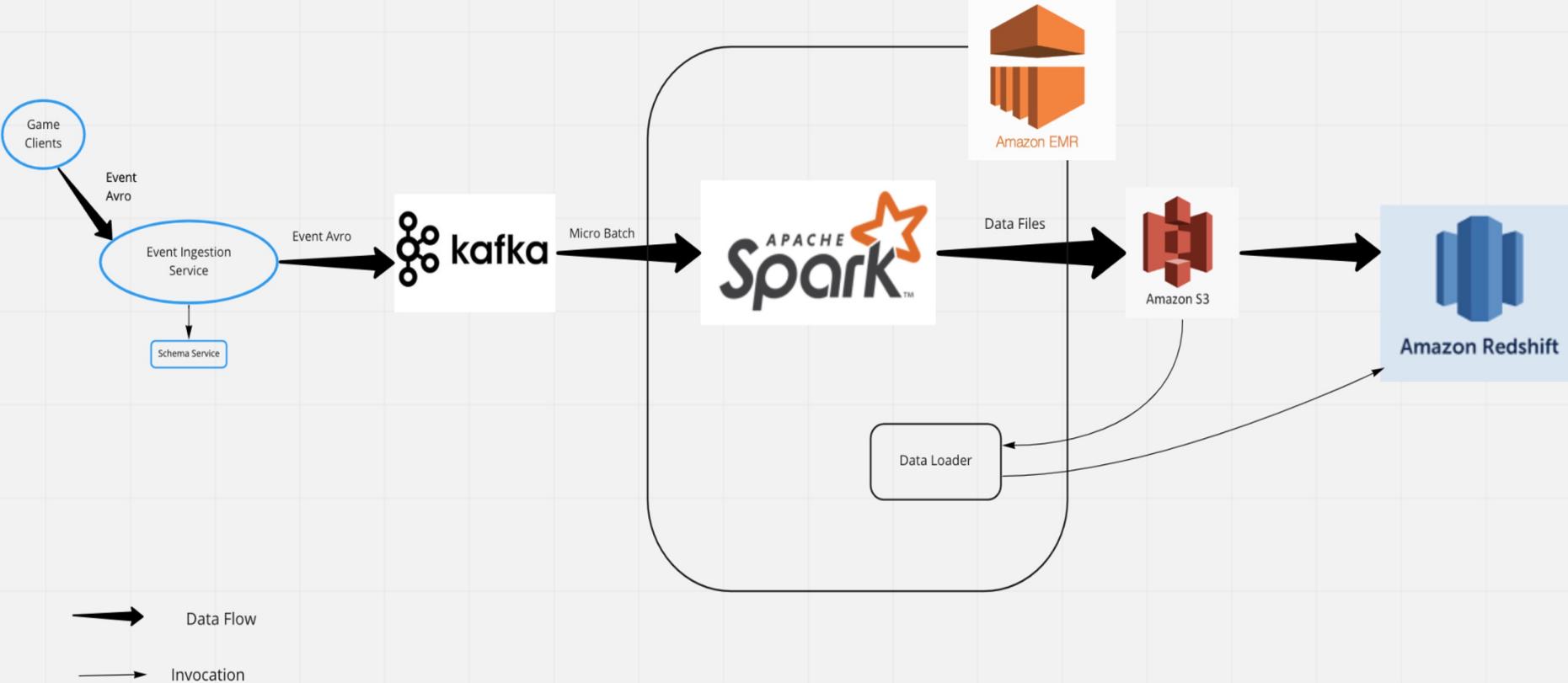
Game/  
Environments

GigaBytes

Data daily

Faster Data Availability  
Game Dev/Analyst/Data Scientist

# Game Telemetry Data Pipeline



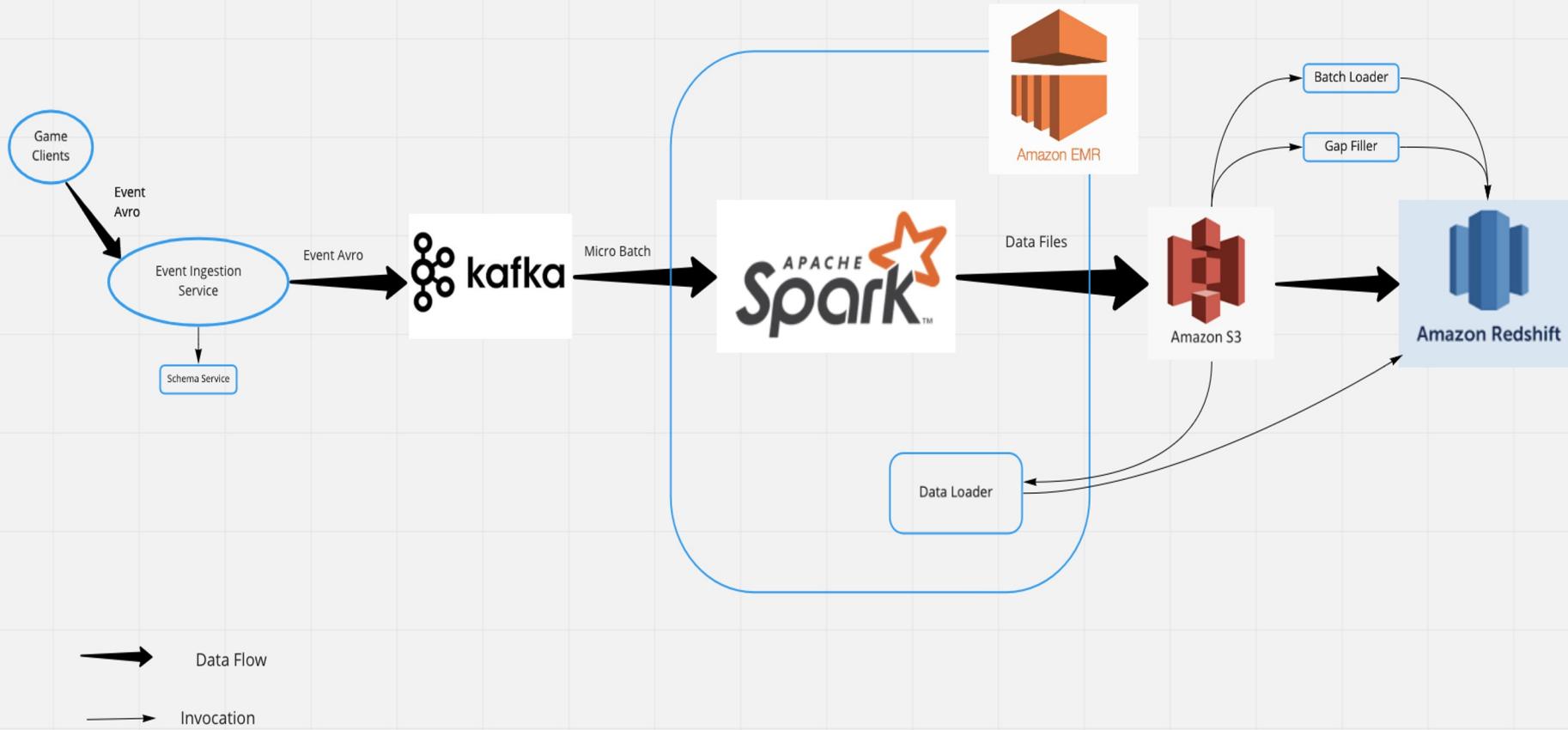
# Ingesting data into Redshift

- What is a Copy command in Redshift?
- S3 Copy command syntax:

```
copy table1  
(column1, column2, etc.)  
from 's3://demobucket/table1/datafile_1.txt'  
iam_role 'arn:aws:iam:::role/'  
region 'us-west-2';
```

- S3 manifest   's3://demobucket/table1/datafiles.manifest'
- S3 Prefix :   's3://demobucket/table1/datafile\_'

# Game Telemetry Data Pipeline





# Challenges

- Data loading delays into Data Warehouse
- Heavy Operational overhead
- Data Quality Issues



# Data Ingestion Into Redshift using Airflow

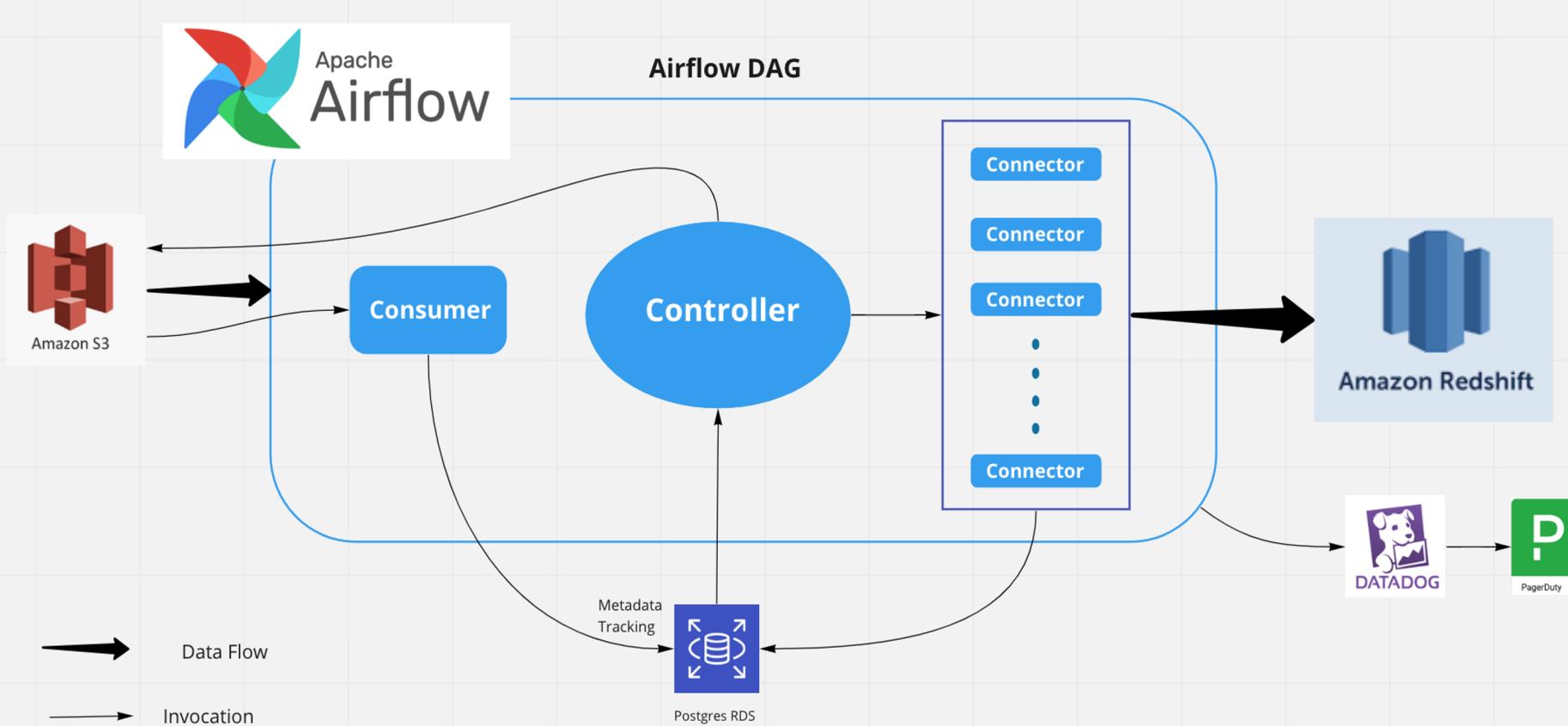
# Features

- Faster data backfills when data delayed in Redshift
- Data quality and Observability
- Unified framework for Redshift data loading

# Why Airflow?

- Scheduler based Dag workflow
- Flexible API to design Idempotent task in Dag
- Highly Scalable
- Built-in Features (Web UI, Variables, Connections, Retry, Alerting)
- Easy integration with external services (AWS, Datadog)

# Redshift Loader



# Redshift Loader Components

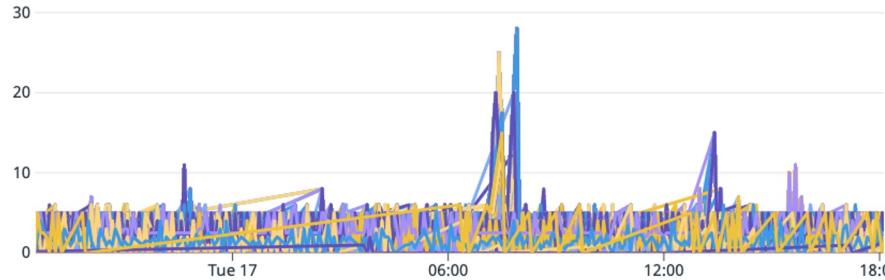
- **Consumer** : Captures S3 metadata for the s3 files and adds it to metadata audit tracking table.
- **Controller** : Pulls Metadata from audit. Dynamically generates copy commands for each table version and sends it to Connector via Xcom variable. In manifest scenario it saves the manifest file in s3.
- **Connector**: Connects to Redshift and executes Copy command statement.

# Data Loading Scenarios

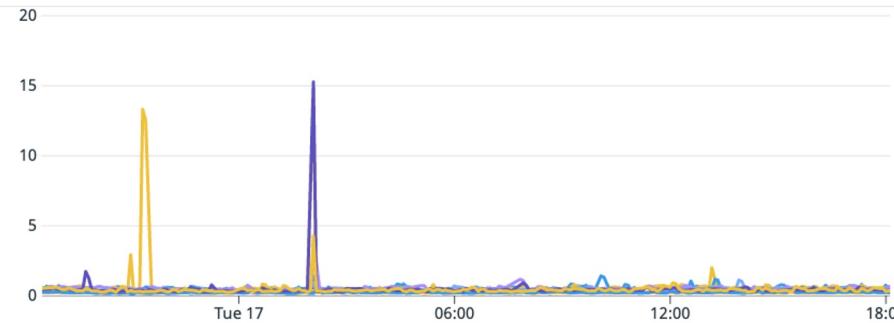
- **Scenario 1** : No resource constrains on Redshift. Default copy command execution.
- **Scenario 2:** Redshift is under heavy query workloads or maintainence and data is not loading for more than an hour. We let the current batch complete and in the next batch collect all the s3 files and create a manifest file. Execute single Copy command with S3 manifest specified.

# Datadog Metrics

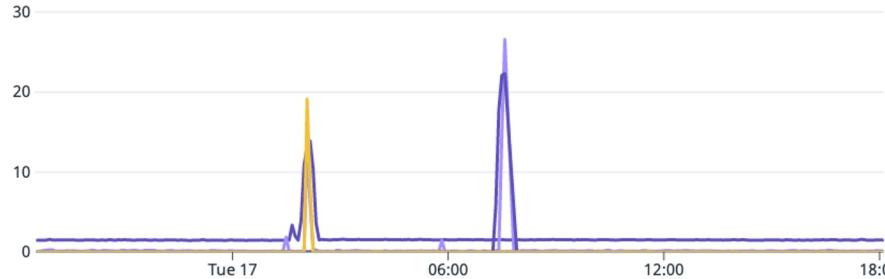
rsloader.data.loading.lag.from.s3 (min)



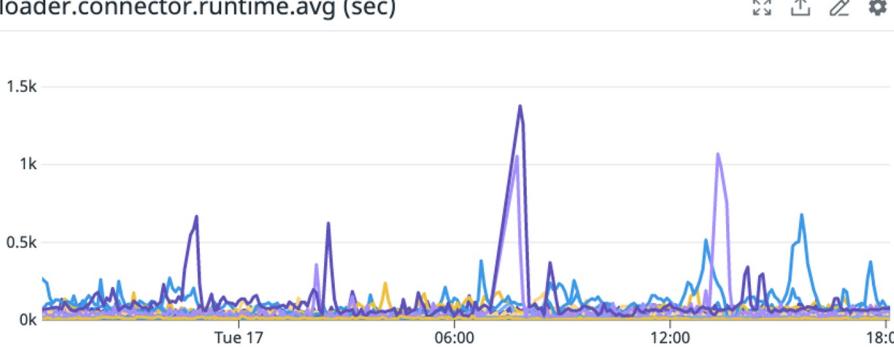
rsloader.consumer.runtime.avg (sec)



rsloader.controller.runtime.avg (sec)



rsloader.connector.runtime.avg (sec)



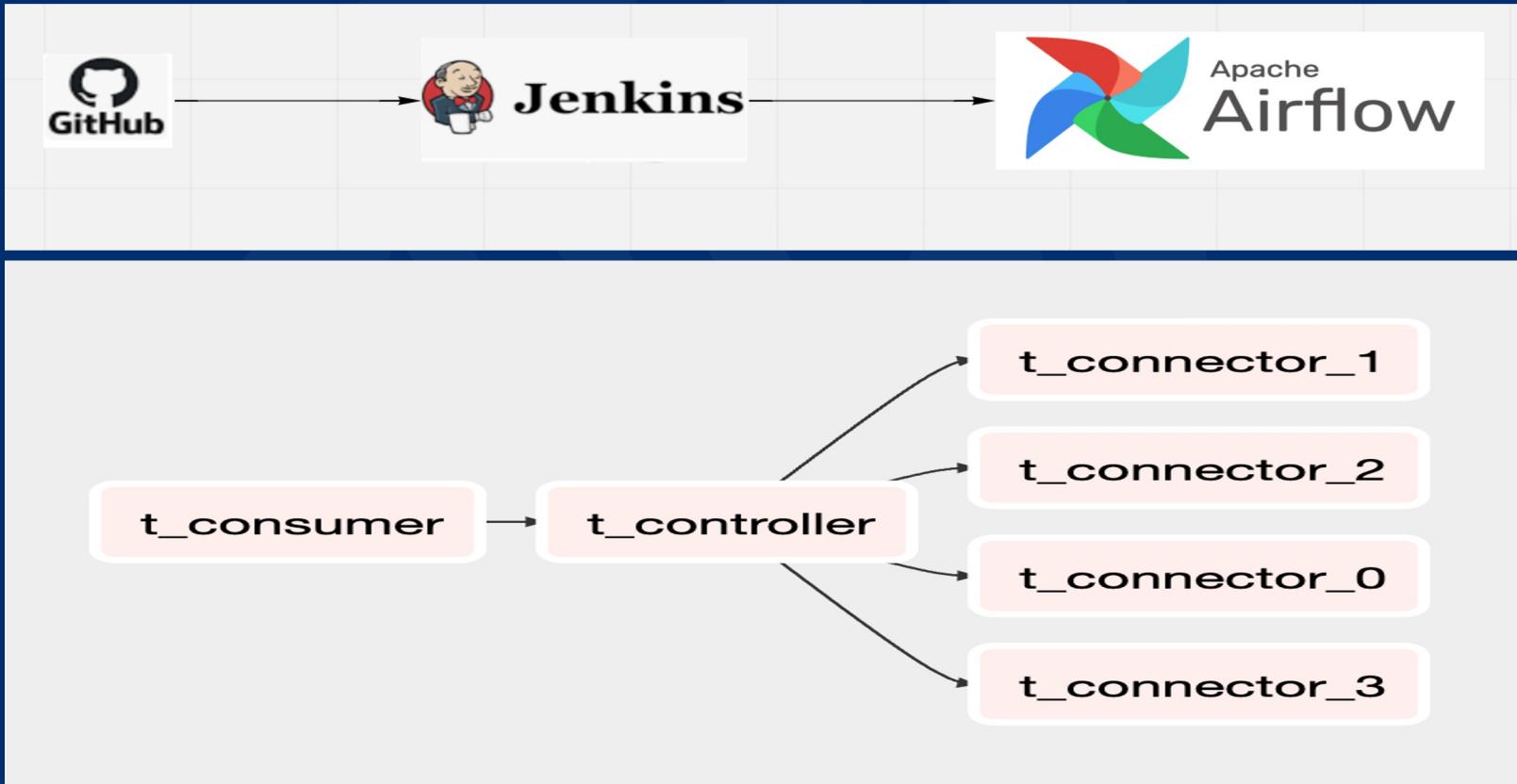
# Declarative Dag Deployment on Airflow

## Easy Onboarding New Game:

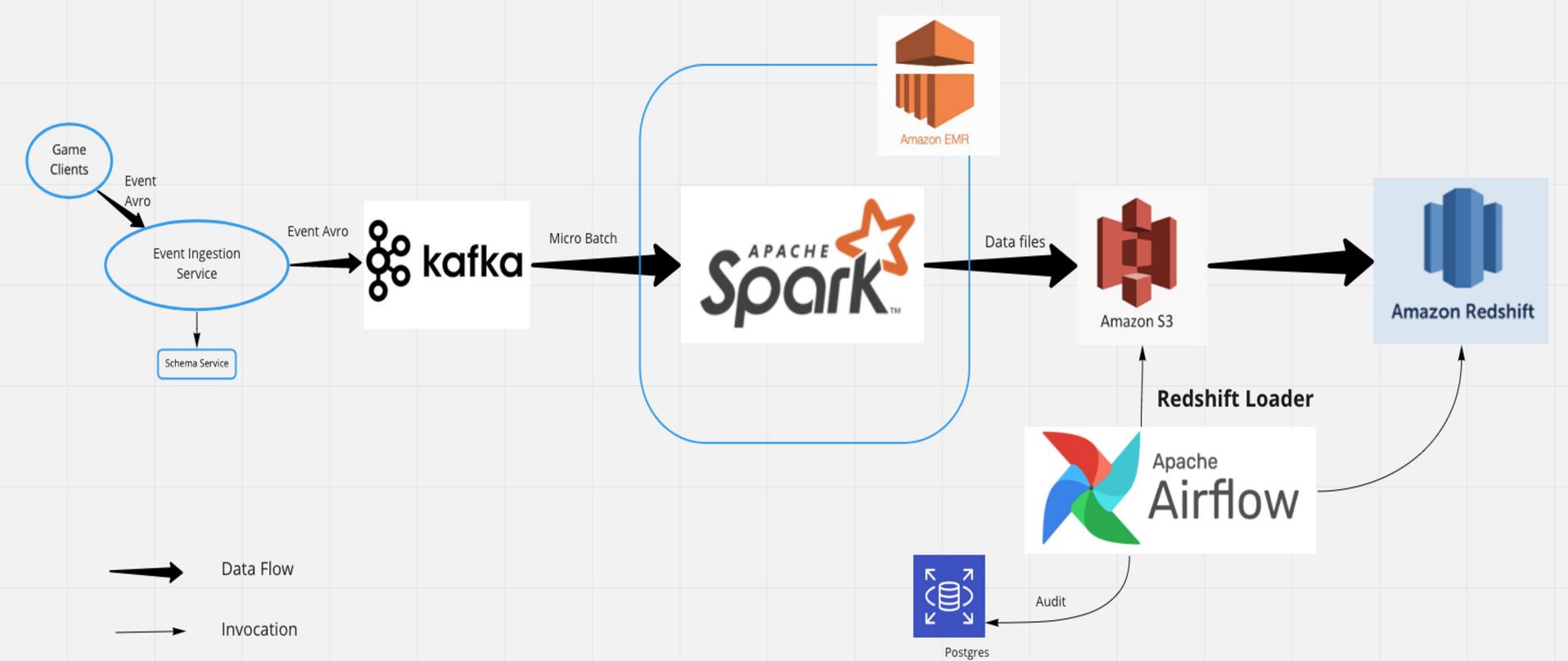
- Create a Json variable with name as <gameName>-<environment>-<eventType>.conf

```
{  
    "schedule": "*/5 * * * *",  
    "daysback": 1,  
    "redshifthostname": "xxxxxxxxxxxx",  
    "connectorqueuecount": 4,  
    "batchwindow": 30,  
    "manifestmin": 3,  
    "exclude.tables": "table1,table4",  
    "email": "xxxxxxxxxx",  
    "batchsize": 1,  
    "retry": 2  
}
```

# Declarative Dag Deployment on Airflow



# WBA Game Telemetry Data Pipeline



# Review

---

---

- Data loading delays into Data Warehouse.
- Heavy Operational overhead
- Data Quality Issues

- Dynamic data loading from Airflow Dag by switch between manifest /prefix / full path of s3 in Copy command.
- Unified framework for data loading via Airflow Dags on a multi-tenant Airflow Celery Cluster with custom metric insights.
- Idempotent tasks in Airflow and maintain transactional audit.



# We are hiring!

<https://careers.wbgames.com/careers/>



# Questions ?