EMR

orders.csv — S3

Job1

B1

Job3

payment.cs

B2

Job2

A

B

C

C — Redshift

Azkaban?

Airflow

Workflow

Airflow DAG

T1

Job1

T2

Job2   } upstream

T3

Job3   } downstream

# Web App

## Frontend

UI

## Backend

Actual Process

data/
Metadata → Database

(EC 2)

Docker
Airflow docker image

Airflow → S3 → dags

S3
raw // day 1
      day 2

Airflow → Spark on EMR → day 1 → Hive Table
                         day 1
                          +
                         day 2

day 1