Instacart Market Basket Analysis

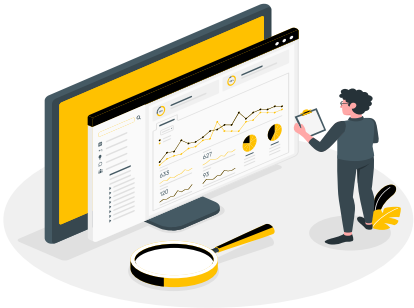


Table of Contents

01 Project Background & Objectives

Background

- Data comes from Kaggle

aisles	134 × 2	aisle_id(int), aisle(chr)
departments	21 × 2	department_id(int), department(chr)
products	49,688 × 4	product_id(int), product_name(chr), aisle_id, department_id
orders	3,421,083 × 7	order_id(int), user_id(int), eval_set(chr), order_number(int), order_dow(int), order_hour_of_day(int), days_since_prior_order(num)
order_products	33,819,106 × 4	order_id(int), product_id(int), add_to_cart_order(int), reorder(int)

- To provide a delightful shopping experience by using customer orders over time to predict which previously purchased products will be in a user's next order

Instacart Market

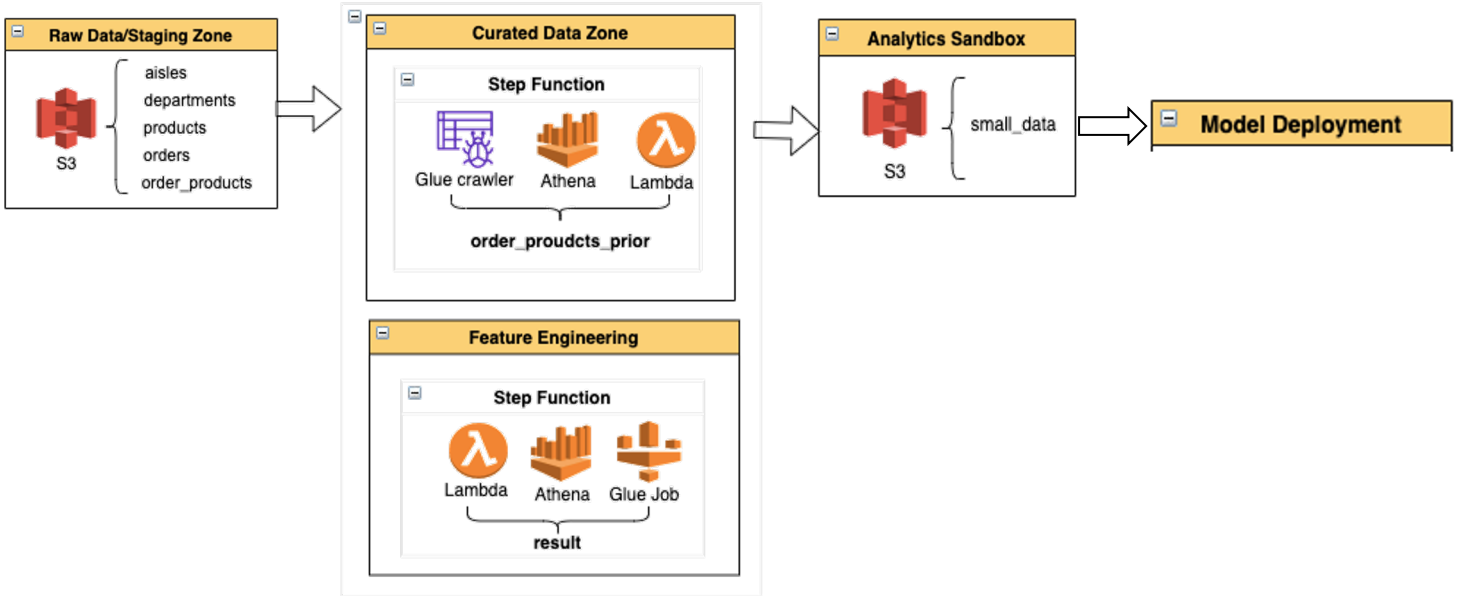
01 Project Background & Objectives

Objectives

- Build automated ETL pipeline to process big data
- Build model to do the prediction
- Deploy

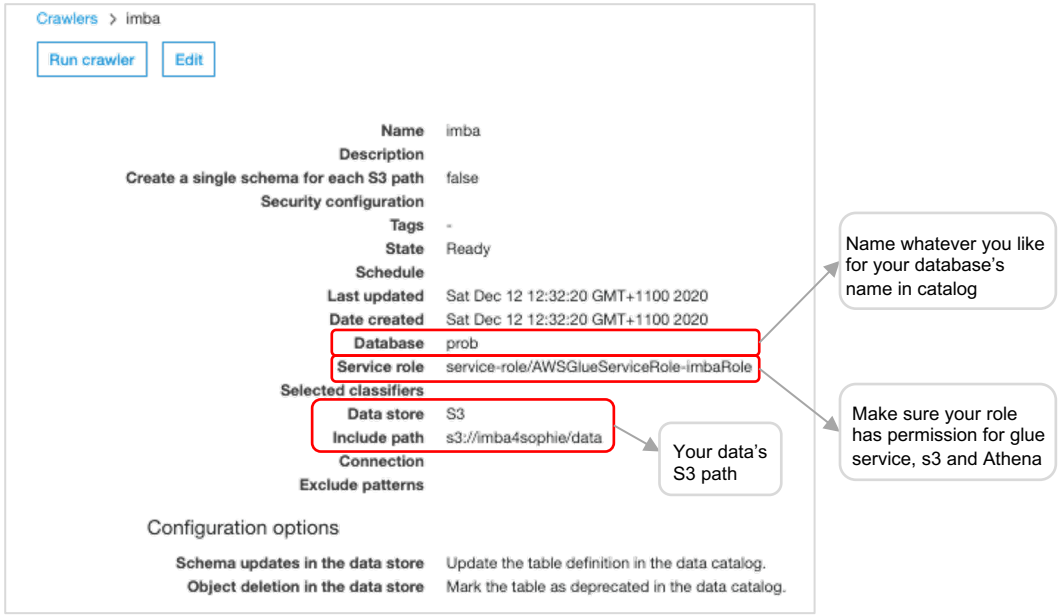
Instacart Market

02 Automated ETL pipeline for big data



02 Automated ETL pipeline for big data

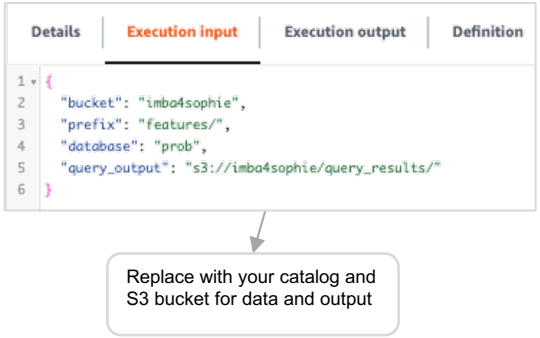
Glue Crawler

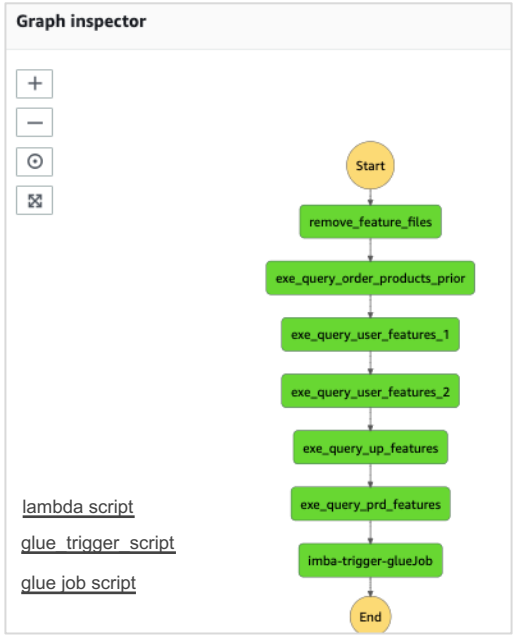


02 Automated ETL pipeline for big data

Step Function

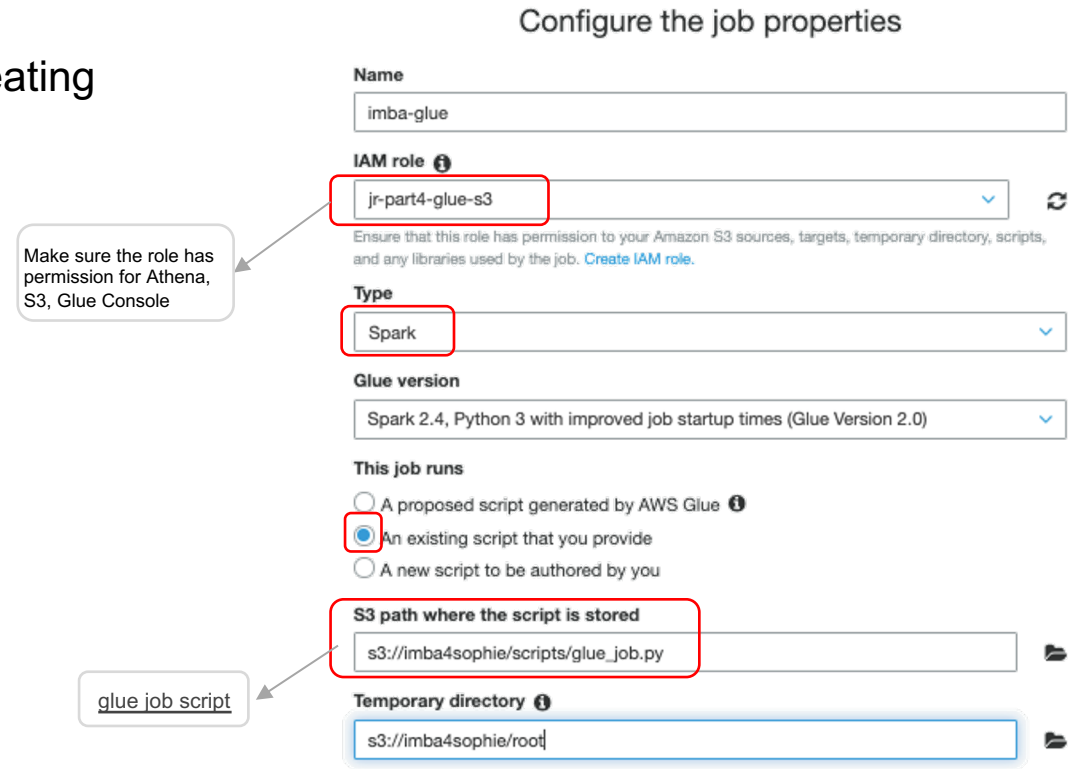
- Create a state machine with script in definition
- Give execution input as below figure showing
- Make sure it has Lambda permission



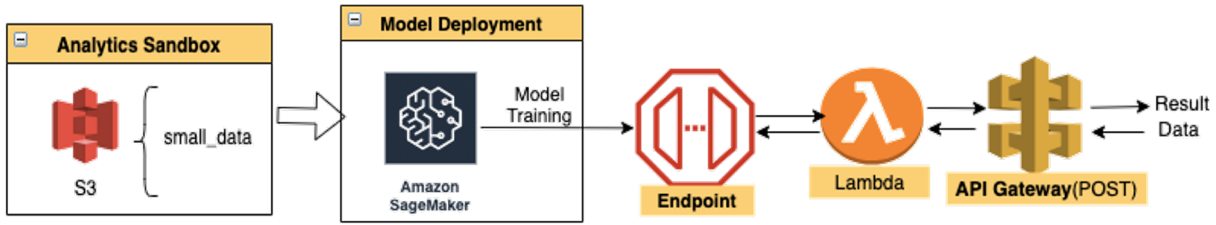


02 Automated ETL pipeline for big data

Glue Job Creating



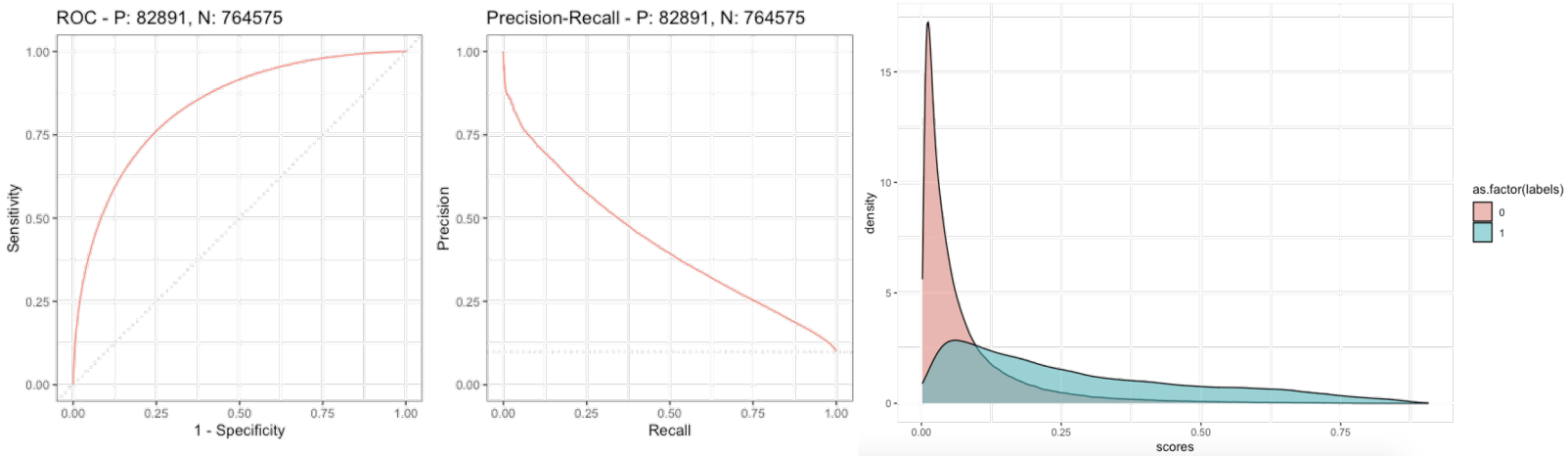
03 Deployment



Data Modelling

Products Reordered Prediction

- The goal is to predict if the purchased product will be ordered again
- The model was built using Xgboost
- The model achieved a test AUC of 0.832
- R Libraries used: ProjectTemplate, tidyverse, xgboost, pROC, precrec



04 Future Work

- Temp Zone: We can add temp zone before staging zone to do data validation
- Partition: When generating curated data, we can partition by a specific column to drastically cut the processing time and cost
- Streaming: We can set the interval to minute in Glue crawler