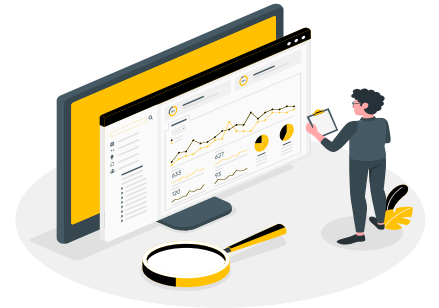


# Instacart Market Basket Analysis



# Table of Contents

01



Project Background &  
Objectives

02



Automated ETL pipeline  
for big data

03



Deployment

04



Future Work

# 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

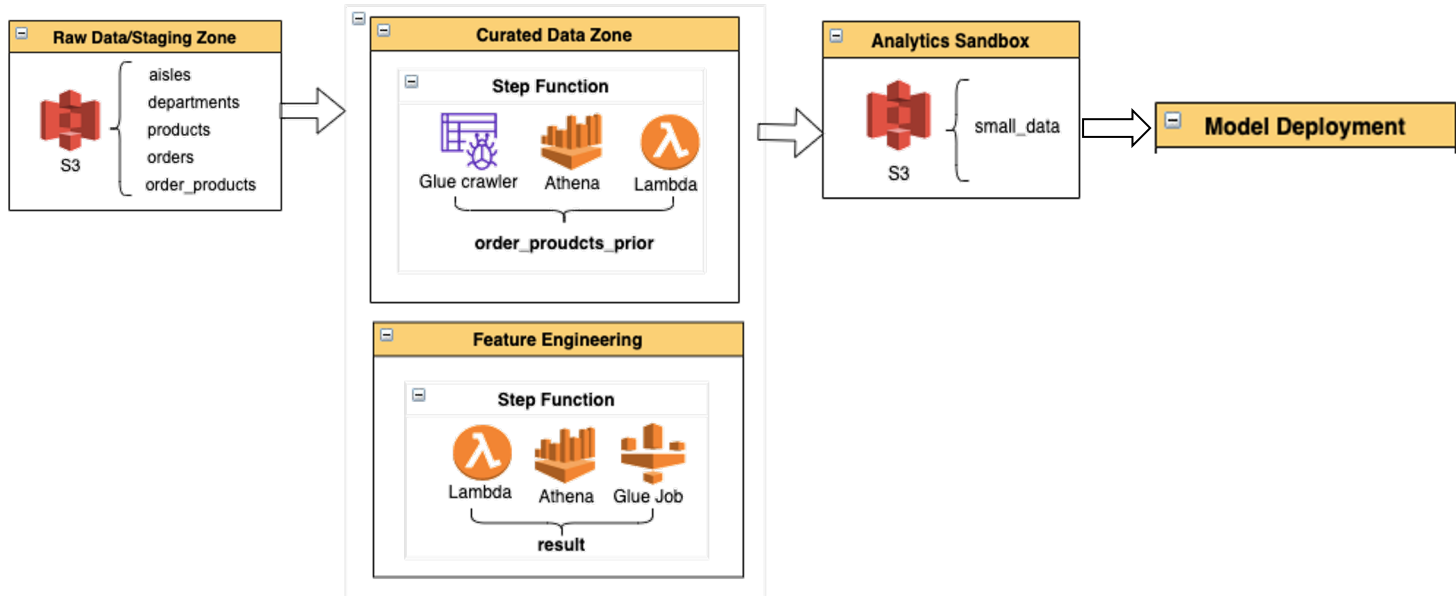
# 01 Project Background & Objectives

## Objectives

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



## 02 Automated ETL pipeline for big data



## 02 Automated ETL pipeline for big data

### Glue Crawler

[Crawlers](#) > imba

[Run crawler](#) [Edit](#)

Name	imba
Description	Create a single schema for each S3 path
Security configuration	false
Tags	-
State	Ready
Schedule	
Last updated	Sat Dec 12 12:32:20 GMT+1100 2020
Date created	Sat Dec 12 12:32:20 GMT+1100 2020
Database	prob
Service role	service-role/AWSGlueServiceRole-imbaRole
Selected classifiers	
Data store	S3
Include path	s3://imba4sophie/data
Connection	
Exclude patterns	

Configuration options

Schema updates in the data store	Update the table definition in the data catalog.
Object deletion in the data store	Mark the table as deprecated in the data catalog.

Name whatever you like for your database's name in catalog

Your data's S3 path

Make sure your role has permission for glue service, s3 and Athena

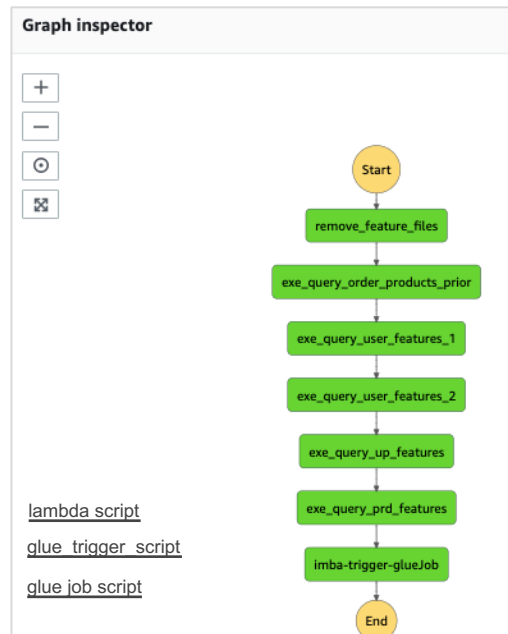
## 02 Automated ETL pipeline for big data

### Step Function

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

Details	Execution input	Execution output	Definition
1 +	{		
2	"bucket": "imba4sophie",		
3	"prefix": "features/",		
4	"database": "prob",		
5	"query_output": "s3://imba4sophie/query_results/"		
6	}		

Replace with your catalog and  
S3 bucket for data and output



## 02 Automated ETL pipeline for big data

### Glue Job Creating

#### Configure the job properties

Name

imba-glue

IAM role ⓘ

jr-part4-glue-s3

Make sure the role has permission for Athena, S3, Glue Console

Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job. [Create IAM role.](#)

Type

Spark

Glue version

Spark 2.4, Python 3 with improved job startup times (Glue Version 2.0)

This job runs

☐ A proposed script generated by AWS Glue ⓘ

☒ An existing script that you provide

☐ A new script to be authored by you

S3 path where the script is stored

s3://imba4sophie/scripts/glue\_job.py

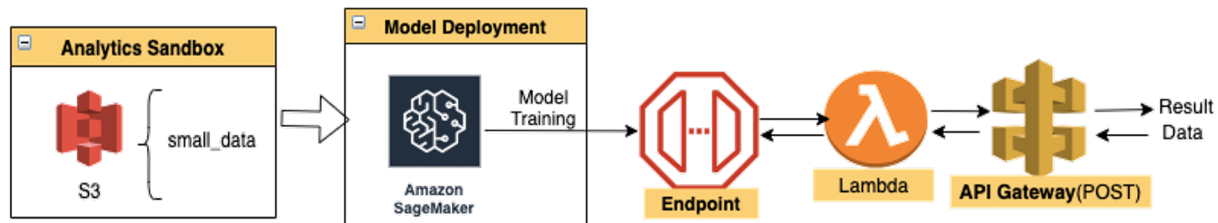
glue job script

Temporary directory ⓘ

s3://imba4sophie/root



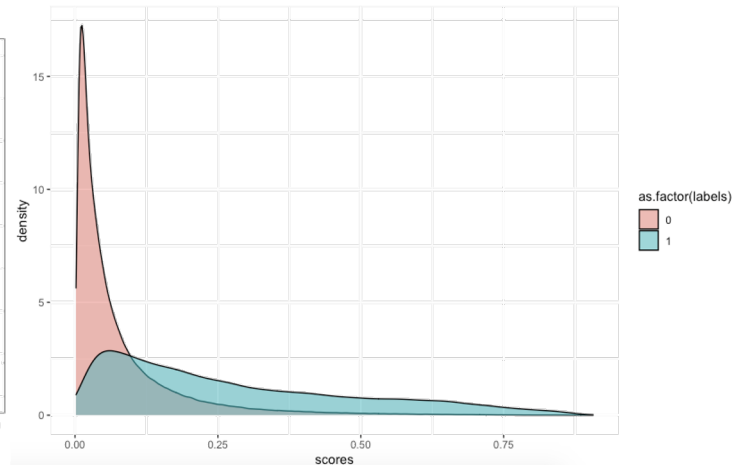
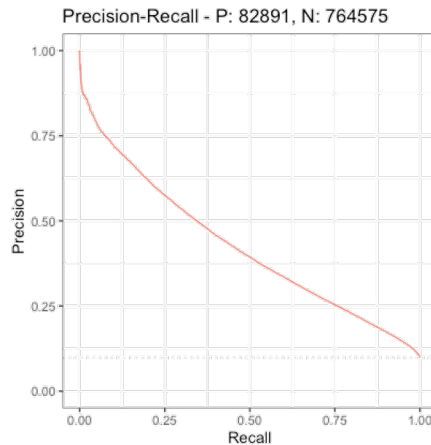
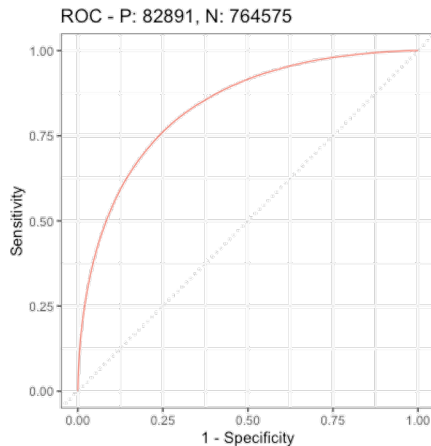
## 03 Deployment



# 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



## 03 Deployment – Sage Maker

- Create a notebook instance
- make sure it has permission for S3 and Sage Maker

Notebook instances

🔄

Actions ▾

Create notebook instance

🔍

Search notebook instances

<

1

>

⚙️

	Name ▾	Instance	Creation time ▾	Status ▾	Actions
<input type="radio"/>	imba-deployment	ml.t2.medium	Dec 30, 2020 03:39 UTC	🟢 InService	<div>Open Jupyter</div> <div>Open JupyterLab</div>

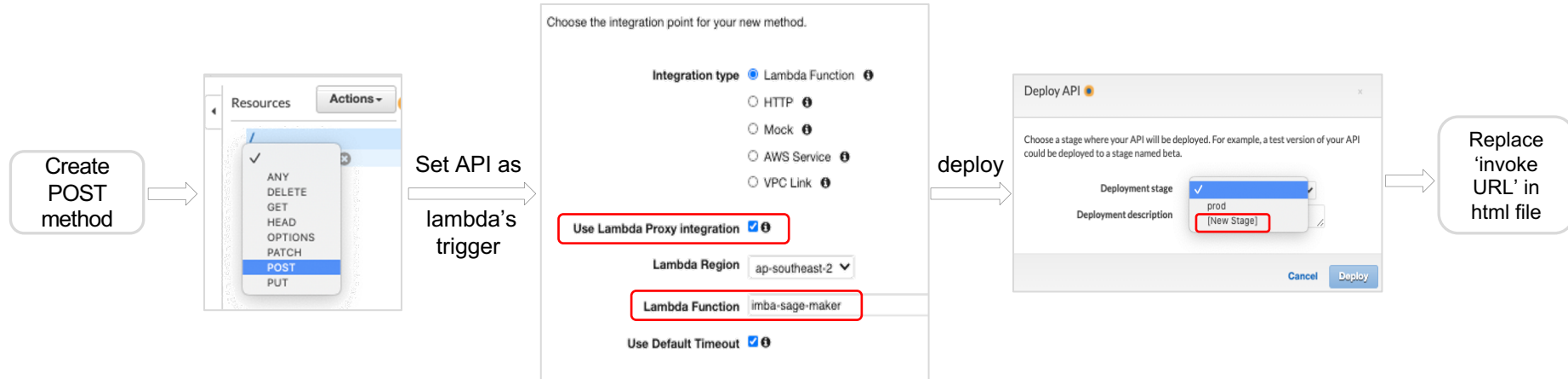
- After attaching trained estimator object, we can get endpoint via deploy method



lambda script

## 03 Deployment – API Gateway

Setting up a new API



## 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