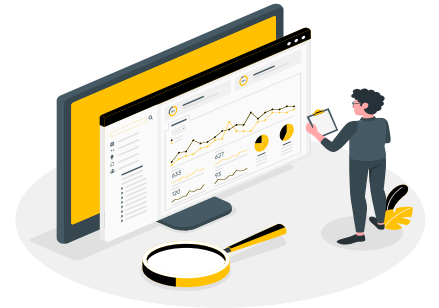


# Instacart Market Basket Analysis



# Table of Contents

01



Project Background &  
Objectives

02



Automated ETL pipeline to process  
big data

03



Data Modelling

04



Deployment

# 01 Project Background & Objectives

## Background

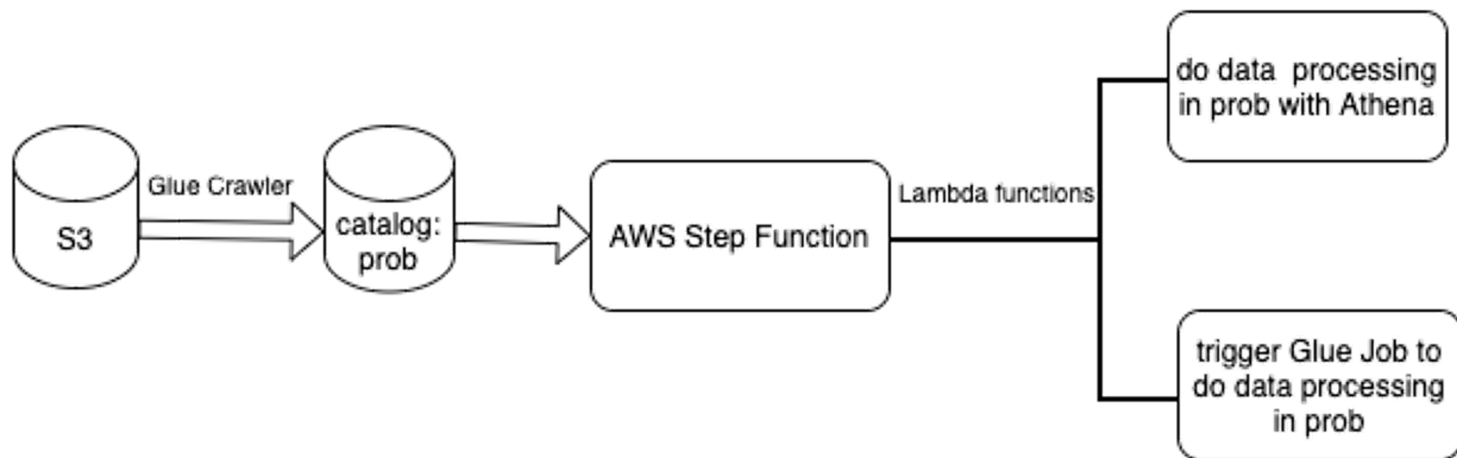
- Data comes from [Kaggle](#)
- To provide a delightful shopping experience by using customer orders over time to predict which previously purchased products will be in a user's next order



## Objectives

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

## 02 Automated ETL pipeline to process big data



## 02 Automated ETL pipeline to process big data

### Glue Crawler

[Crawlers](#) > imba

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

Name	imba
Description	
Create a single schema for each S3 path	false
Security configuration	
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.

Annotations:

- 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

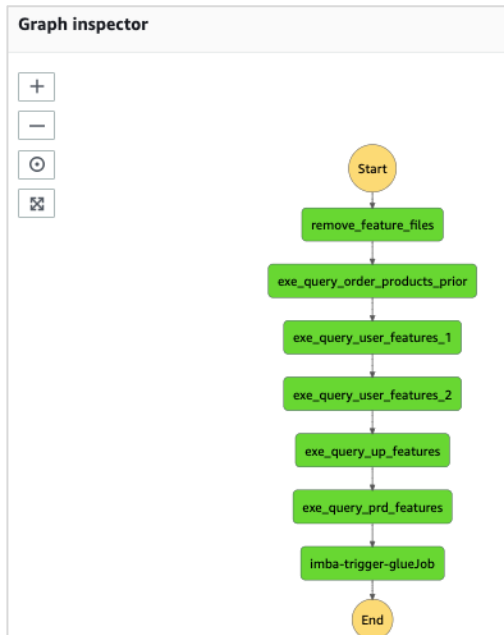
## 02 Automated ETL pipeline to process big data

### Step Function

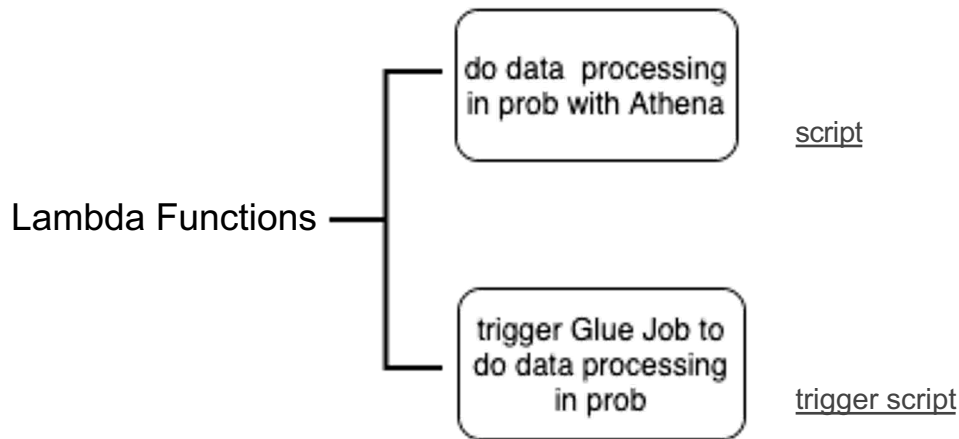
- Create a state machine with script in definition and execution input as below figure showing
- Make sure it has Lambda permission

Details	Execution input	Execution output	Definition
1	<pre>{</pre>		
2			
3			
4			
5			
6			

Replace with your catalog and S3 bucket for data and output



## 02 Automated ETL pipeline to process big data



*Note: make sure your lambda function role has permission for Glue Console, Athena and Lambda*

## 02 Automated ETL pipeline to process big data

### Glue Job Creating

#### Configure the job properties

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

Name

imba-glue

IAM role ⓘ

jr-part4-glue-s3

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

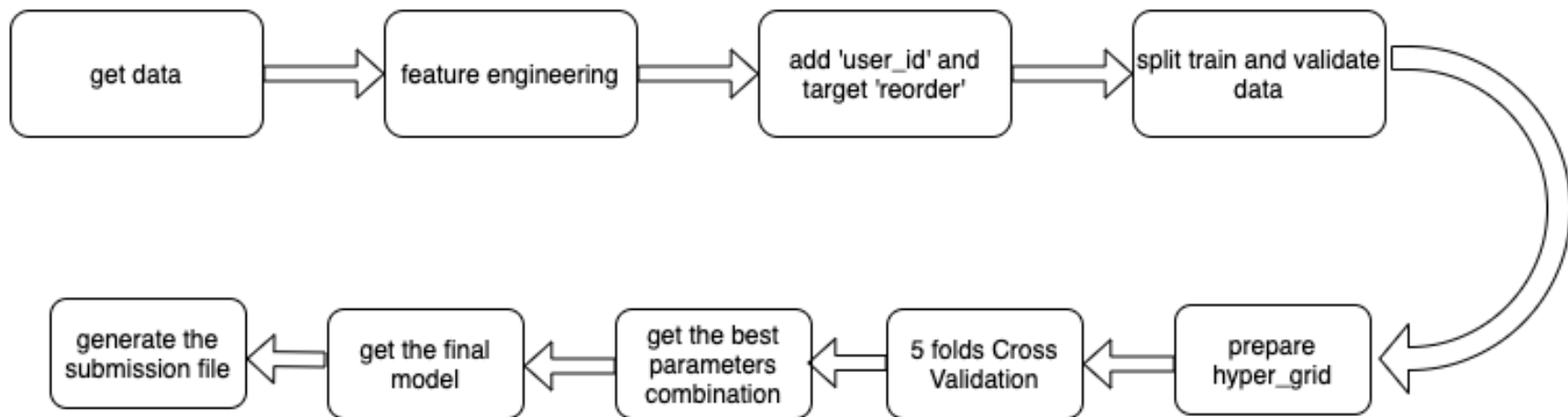
[glue job script](#)

Temporary directory ⓘ

s3://imba4sophie/root



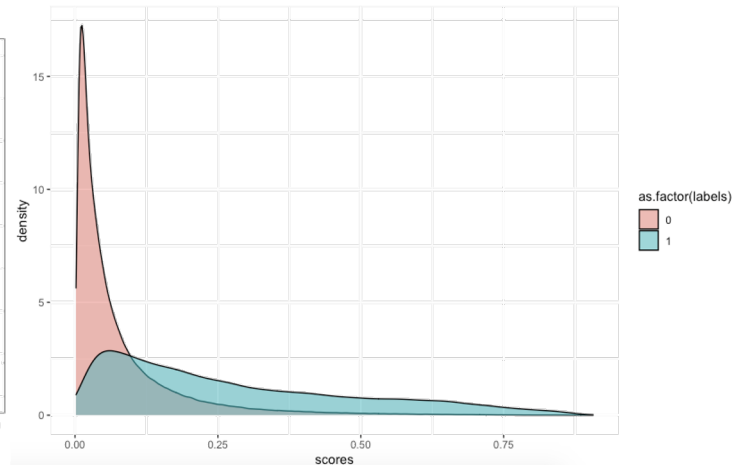
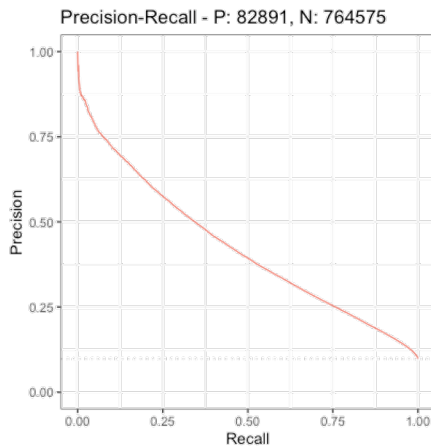
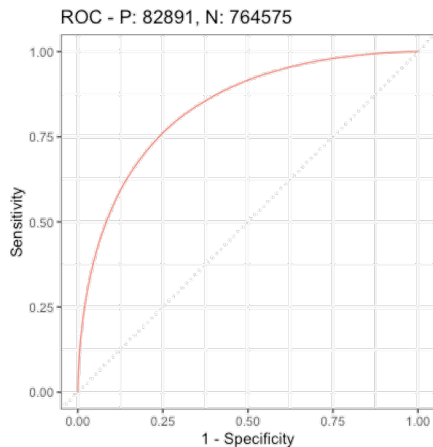
## 03 Data Modelling



## 03 Data Modelling

### Products Reordered Prediction [R script](#)

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



# 04 Deployment

To be added...