

ĐÁNH GIÁ DỮ LIỆU NHÀ HÀNG ẨM THỰC

MỤC LỤC

I. Tổng quan dự án

II. Đánh giá dữ liệu

III. Chi tiết dự án

IV. Kết quả thu được

Tổng quan dự án

Mục đích dự án

- Thu thập dữ liệu đánh giá về các nhà hàng trong các khu vực
- Tạo ra một pipeline trên AWS nhằm xử lý dữ liệu
- Đưa ra các đánh giá thông qua biểu đồ giúp các nhà đầu tư có thể đưa quyết định kinh doanh



ĐÁNH GIÁ DỮ LIỆU

Zomato là một nền tảng trực tuyến liên quan đến ẩm thực

- Đề xuất các nhà hàng ẩm thực phổ biến
- Cung cấp đánh giá của người dùng
- Nơi để nhấn mạnh, quảng bá thương hiệu của các doanh nghiệp

Kích thước khoảng 600 MB

Dữ liệu gồm 17 trường

Bao gồm khoảng 550.000 dòng

	url	address	name	online_order	book_table	rate	votes	phone	location	rest_type	dish_liked	cuisines	approx_cost(for two people)	reviews.list	menu_item	listed_in(type)	listed_in(city)
	https://www.zomato.com/bangalore/jalsa-banashankari...	942, 21st Main Road, 2nd Stage, Banashankari, ...	Jalsa	Yes	Yes	4.1/5	775	080 42297555\n+91 9743772233	Banashankari	Casual Dining	Pasta, Lunch Buffet, Masala Papad, Paneer Laja...	North Indian, Mughlai, Chinese	800	[("Rated 4.0", "RATED\nA beautiful place to ...	0	Buffet	Banashankari
	https://www.zomato.com/bangalore/spice-elephant...	2nd Floor, 80 Feet Road, Near Big Bazaar, 6th ...	Spice Elephant	Yes	No	4.1/5	787	080 41714161	Banashankari	Casual Dining	Momos, Lunch Buffet, Chocolate Nirvana, Thai G...	Chinese, North Indian, Thai	800	[("Rated 4.0", "RATED\nHad been here for din...")]	0	Buffet	Banashankari
	https://www.zomato.com/SanchurroBangalore?cont...	1112, Next to KIMS Medical College, 17th Cross...	San Churro Cafe	Yes	No	3.8/5	918	+91 9663487993	Banashankari	Cafe, Casual Dining	Churros, Cannelloni, Minestrone Soup, Hot Choc...	Cafe, Mexican, Italian	800	[("Rated 3.0", "RATED\nAmbience is not that ...")]	0	Buffet	Banashankari
	https://www.zomato.com/bangalore/addhuri-udupi...	1st Floor, Annakuteera, 3rd Stage, Banashankari...	Addhuri Udupi Bhojana	No	No	3.7/5	88	+91 9620009302	Banashankari	Quick Bites	Masala Dosa	South Indian, North Indian	300	[("Rated 4.0", "RATED\nGreat food and proper...")]	0	Buffet	Banashankari
	https://www.zomato.com/bangalore/grand-village...	10, 3rd Floor, Lakshmi Associates, Gandhi Baza...	Grand Village	No	No	3.8/5	166	+91 8026612447\n+91 9901210005	Basavanagudi	Casual Dining	Pani Puri, Gol Gappe	North Indian, Rajasthani	600	[("Rated 4.0", "RATED\nVery good restaurant ...")]	0	Buffet	Banashankari

Xử lý dữ liệu



Loại bỏ các trường
không cần thiết



Format các kiểu dữ
liệu hợp lý



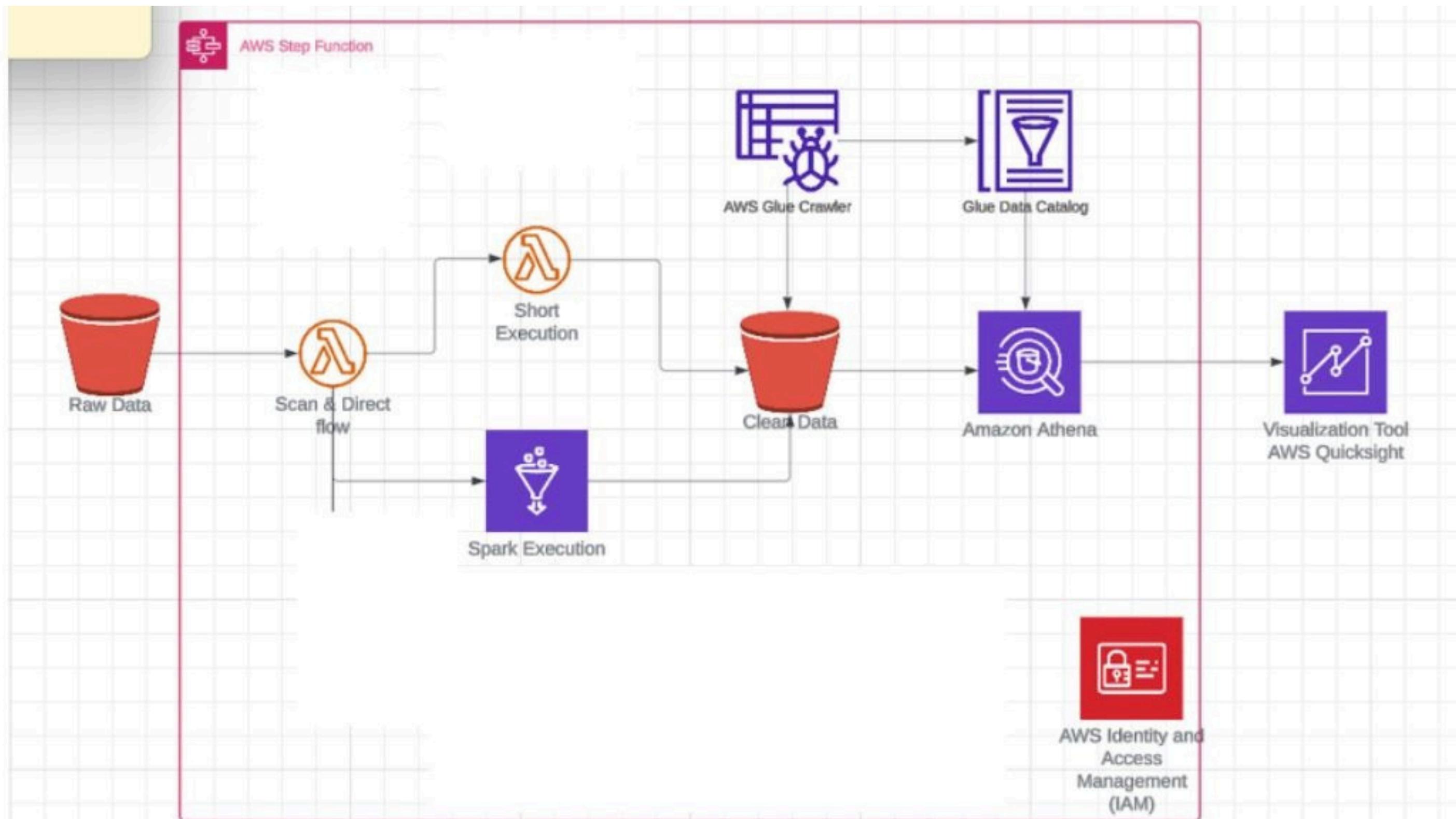
Xử lý các dữ liệu
trống, trùng lặp



CHI TIẾT DỰ ÁN



Sơ đồ xử lý dữ liệu



Tải dữ liệu lên S3

Sao chép giấy
ghi chú rồi thêm
suy nghĩ hoặc
cảm xúc của bạn.

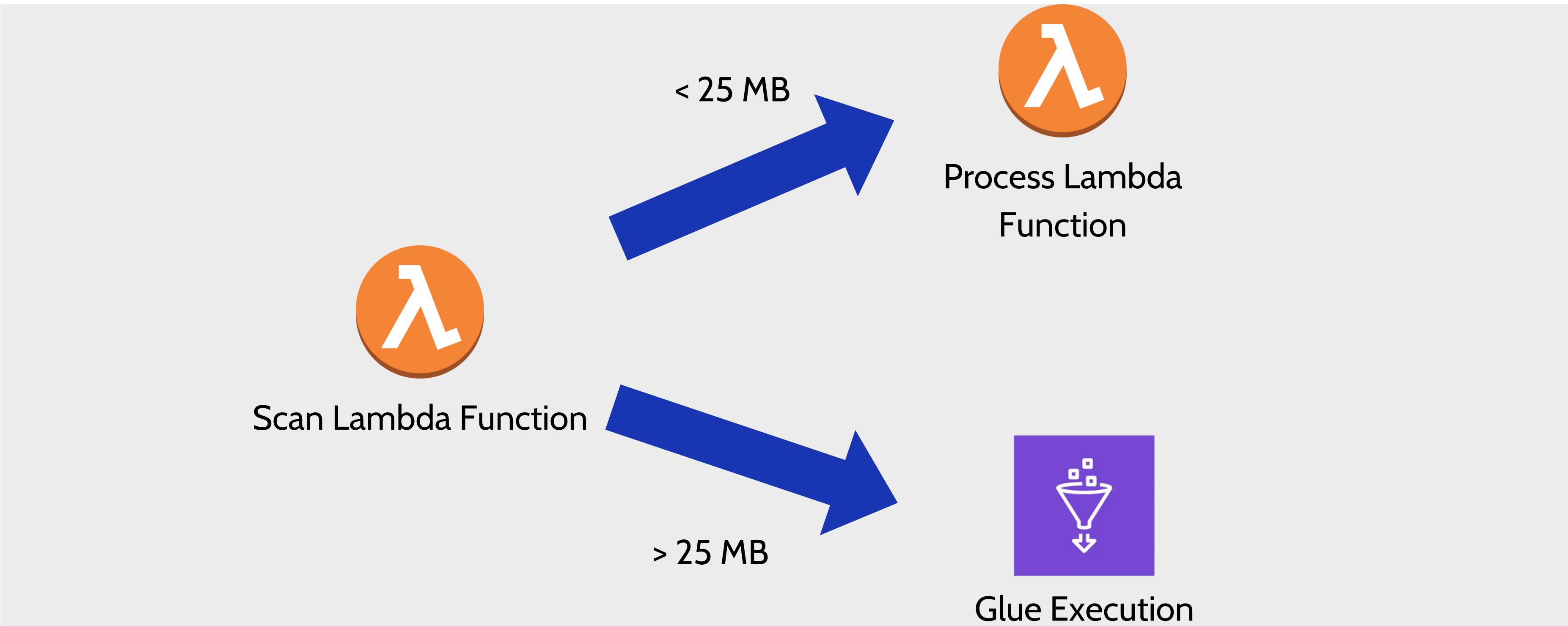
Sao chép giấy
ghi chú rồi thêm
suy nghĩ hoặc
cảm xúc của bạn.

Sao chép giấy
ghi chú rồi thêm
suy nghĩ hoặc
cảm xúc của bạn.

The screenshot shows the AWS S3 console interface. On the left, there's a sidebar with 'Amazon S3' selected. The main area shows a bucket named 'hoanglht2-bucket' with a folder 'raw_data/'. The 'Objects' tab is active, displaying four CSV files: 'zomato_1.csv', 'zomato_2.csv', 'zomato_3.csv', and 'zomato_4.csv'. Each file is 75.2 MB, 91.3 MB, 129.9 MB, and 129.9 MB respectively, all stored in the 'Standard' storage class. The 'Actions' dropdown menu is open, with 'Upload' highlighted in orange. A tooltip for 'Upload' says: 'Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory [?] to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more [?]'.

Name	Type	Last modified	Size	Storage class
zomato_1.csv	csv	September 10, 2024, 15:40:46 (UTC+07:00)	75.2 MB	Standard
zomato_2.csv	csv	September 10, 2024, 15:40:46 (UTC+07:00)	91.3 MB	Standard
zomato_3.csv	csv	September 10, 2024, 16:27:19 (UTC+07:00)	129.9 MB	Standard
zomato_4.csv	csv	September 11, 2024, 00:00:00 (UTC+07:00)	129.9 MB	Standard

Tạo Scan Lambda Function



Tạo scan lambda function

The screenshot shows the AWS Lambda Functions overview page for a function named 'hoanglht2-lambda'. The top navigation bar includes the AWS logo, Services, a search bar, and account information for 'HoangLHT2 @ fsoft-cta' in the 'N. Virginia' region. The main content area displays the function's name, a 'Function overview' section with tabs for 'Diagram' (selected) and 'Template', and a 'Code source' section with an 'Upload from' button. On the right side, detailed information is provided: Description (empty), Last modified (2 days ago), Function ARN (arn:aws:lambda:us-east-1:666243375423:function:hoanglht2-lambda), and Function URL (Info). Below the main content, there are tabs for 'Code', 'Test', 'Monitor', 'Configuration', 'Aliases', and 'Versions'.

Lambda > Functions > hoanglht2-lambda

hoanglht2-lambda

Function overview Info

Diagram Template

hoanglht2-lambda

S3

Add destination

Add trigger

Description

Last modified

2 days ago

Function ARN

arn:aws:lambda:us-east-1:666243375423:function:hoanglht2-lambda

Function URL Info

Code Test Monitor Configuration Aliases Versions

Code source Info Upload from

CloudShell Feedback © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences



Tạo Execution Lambda Function

The screenshot shows the AWS Lambda Functions console. The top navigation bar includes the AWS logo, Services dropdown, a search bar, and user information (HoangLHT2 @ fsoft-cta). The left sidebar shows the Lambda > Functions hierarchy. The main content area displays the details for the function 'hoanglht2-lambda-process-data'. The 'Function overview' tab is selected, showing a diagram of the function, which consists of a single layer named 'hoanglht2-lambda-process-data' (1 version). Below the diagram are buttons for '+ Add trigger' and '+ Add destination'. To the right, the function's ARN is listed as arn:aws:lambda:us-east-1:666243375423:function:hoanglht2-lambda-process-data. The bottom navigation bar includes tabs for Code, Test, Monitor, Configuration, Aliases, and Versions, with Configuration currently selected. A 'Code source' section at the bottom allows for uploading code from a local file.



Tạo Glue Execution

Screenshot of the AWS Glue console showing the creation of a new ETL job named "hoanglht2-glue-execution".

The left sidebar shows navigation options like "Getting started", "ETL jobs", "Visual ETL", "Notebooks", "Job run monitoring", "Data Catalog tables", "Data connections", "Workflows (orchestration)", "Data Catalog", "Databases", "Tables", "Stream schema registries", "Schemas", "Connections", "Crawlers", "Classifiers", "Catalog settings", "Data Integration and ETL", and "Legacy pages".

The main area displays the "Script" tab of the job configuration. The script content is as follows:

```
11 from awsglue.context import GlueContext
12 from awsglue.job import Job
13 from pyspark.sql.functions import col, regexp_replace, when, avg, first, mean, lit
14
15 # Initialize Spark and Glue contexts
16 sc = SparkContext()
17 glueContext = GlueContext(sc)
18 spark = glueContext.spark_session
19 args = getResolvedOptions(sys.argv, ['JOB_NAME', 'bucket_name', 'object_key'])
20 job = Job(glueContext)
21 job.init(args['JOB_NAME'], args)
22
23 bucket_name = args['bucket_name']
24 object_key = urllib.parse.unquote_plus(args['object_key'])
25
26 # Read CSV file directly from S3 into a DataFrame
27 zomato_df = spark.read.option("multiline", True).csv(f"s3://{bucket_name}/{object_key}", header=True, inferSchema=True, escape='''')
28
29 # Drop Unnecessary Columns
30 zomato_df = zomato_df.drop('url', 'phone', 'address', 'reviews_list', 'menu_item')
31
32 # Replace all commas
```

The status bar at the bottom indicates "Python" and "Ln 1, Col 1". Error and warning counts are shown as 0.

At the bottom, there are links for "CloudShell", "Feedback", "Privacy", "Terms", and "Cookie preferences".

Tạo Glue Crawler

AWS Glue

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

Crawlers

Databases

Tables

Stream schema registries

Schemas

Connections

Classifiers

Catalog settings

Data Integration and ETL

Legacy pages

What's New

Advanced settings

Crawler runs

Schedule

Data sources

Classifiers

Tags

Crawler runs (9)

The list of crawler runs for this crawler.

Filter data

Filter by a date and time range

Start time (UTC)

End time (UTC)

Current/last duration

Status

DPU hours

Table changes

	Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
1	September 10, 2024 at 18:11	September 10, 2024 at 18:12	01 min 34 s	Completed	0.041	1 table change partition change
2	September 10, 2024 at 18:04	September 10, 2024 at 18:05	01 min 30 s	Completed	0.041	-
3	September 10, 2024 at 17:51	September 10, 2024 at 17:52	01 min 20 s	Completed	0.037	1 table change partition change
4	September 10, 2024 at 09:34	September 10, 2024 at 09:35	01 min 12 s	Completed	0.044	1 table change partition change
5	September 10, 2024 at 08:44	September 10, 2024 at 08:45	01 min 02 s	Completed	0.047	1 table change partition change
6	September 10, 2024 at 08:15	September 10, 2024 at 08:16	01 min 14 s	Completed	0.039	1 table change partition change
7	September 10, 2024 at 07:55	September 10, 2024 at 07:56	01 min 16 s	Completed	0.040	1 table change partition change
8	September 10, 2024 at 07:52	September 10, 2024 at 07:53	01 min 03 s	Completed	0.043	1 table change partition change
9	September 10, 2024 at 07:26	September 10, 2024 at 07:26	01 min 06 s	Completed	0.038	1 table change partition change

CloudShell Feedback © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

AWS Athena và Quicksight

The image displays two screenshots illustrating the integration of AWS Athena and Amazon QuickSight.

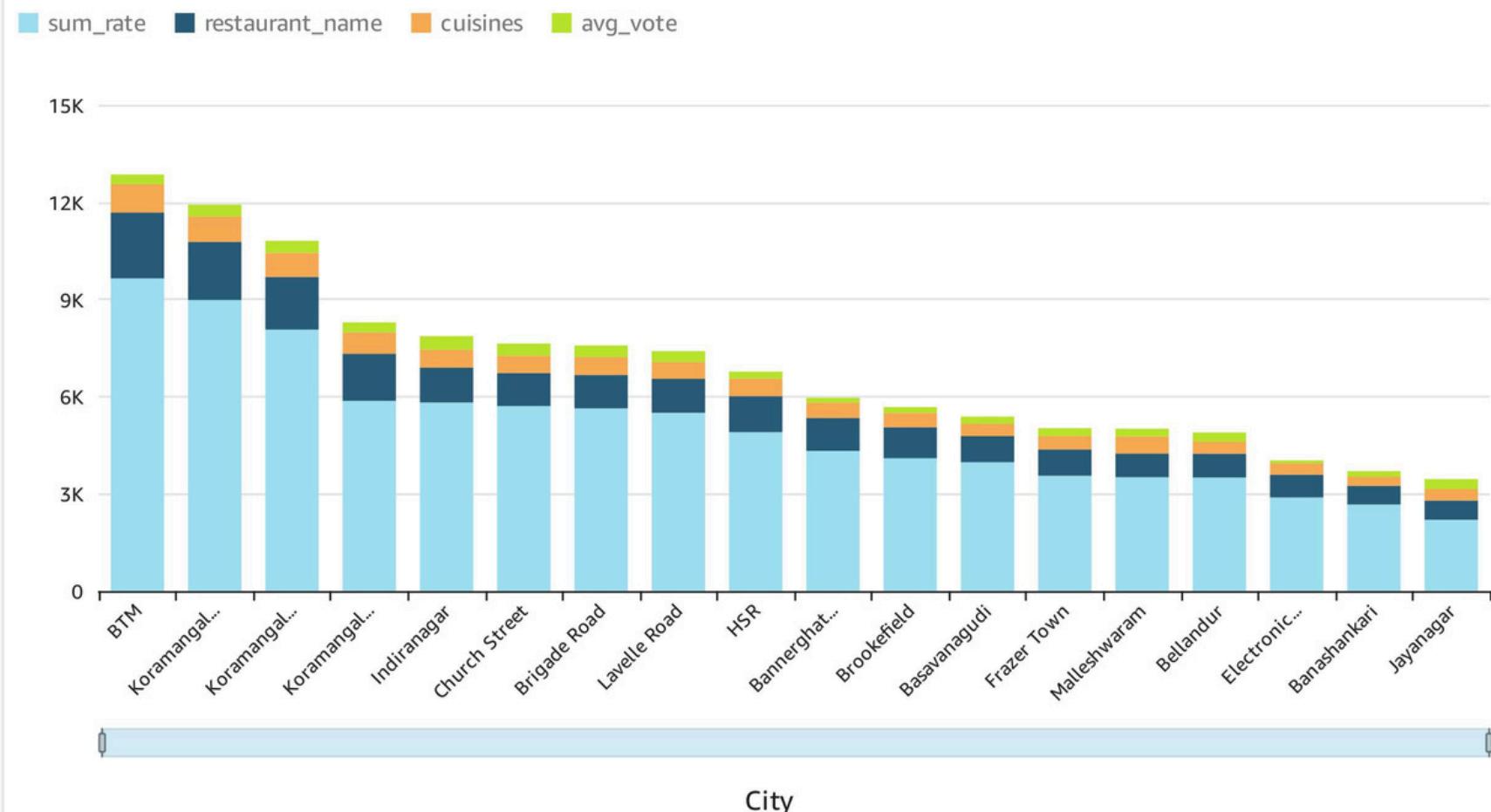
AWS Athena Screenshot: Shows the 'number_custom_each_city' dataset with a completed query. The results table contains 10 rows of data, including columns like name, online_order, book_table, rate, votes, location, and rest_type. The data is as follows:

#	name	online_order	book_table	rate	votes	location	rest_type
1	Juglee	1	0	0	New BEL Road	Quick Bites	
2	Chai Resto	1	0	0	New BEL Road	Beverage	
3	Mojo Pizza - 2X Toppings	1	0	4.1	289	Rajajinagar	Takeaway
4	Amma's Pastries	0	0	4.2	217	Malleshwaram	Bakery
5	La Heaven	1	0	3.8	87	Malleshwaram	Bakery Deli
6	Frozen Coast	0	0	3.9	107	Malleshwaram	Dessert Parlour
7	Smoor	1	0	4.2	287	Malleshwaram	Cafe Dessert
8	Natural Ice Cream	0	0	4.3	188	Malleshwaram	Dessert Parlour
9	Don World	0	0	2.0	122	Mallikarjuna	Bakery

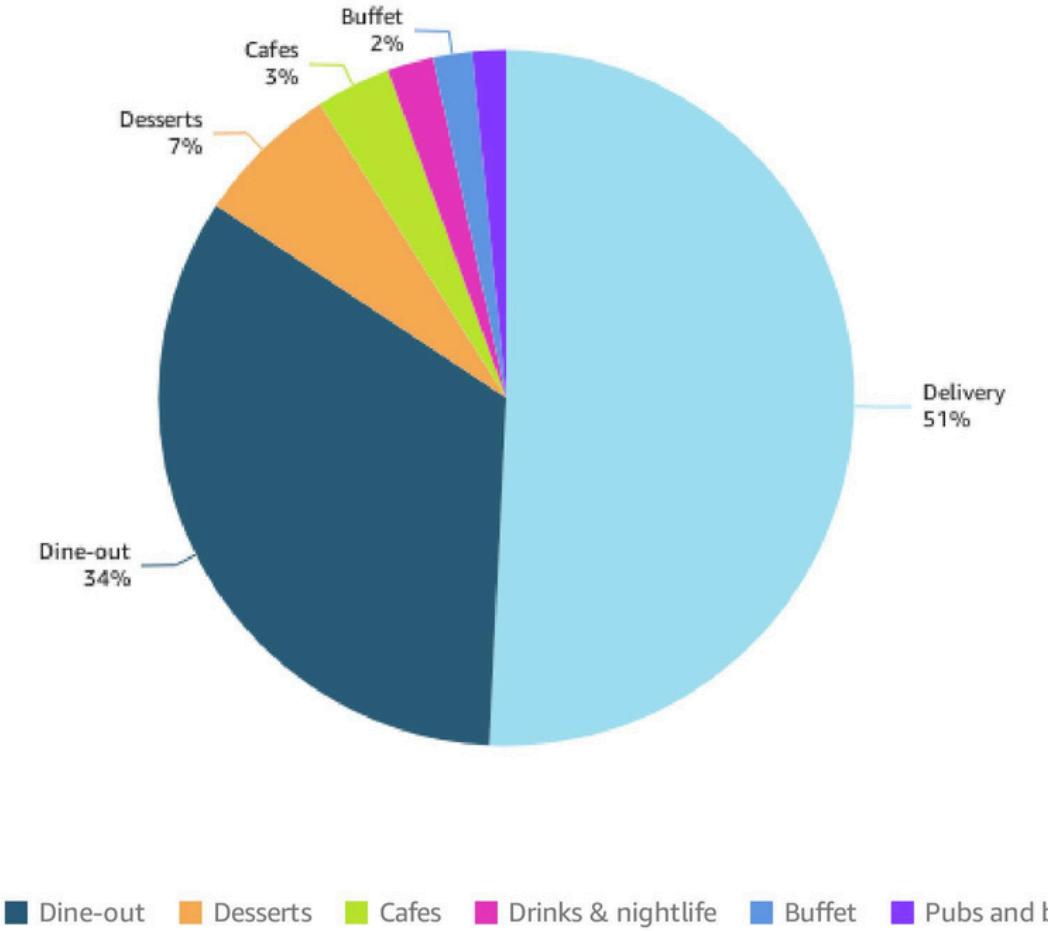
Amazon QuickSight Screenshot: Shows a dashboard titled 'number_custom_each_city analysis' with three visualizations:

- Pie chart:** Distribution of Restaurant Category. The largest category is Delivery (51%), followed by Dine-out (34%), Desserts (7%), Cafes (2%), and others.
- Bar chart:** Overview of each City's Cuisine. The chart shows the average vote for different cuisines across various cities.
- Line chart:** Popularity of Online Order and Book Table in each City. The chart tracks the number of online orders and book tables over time.

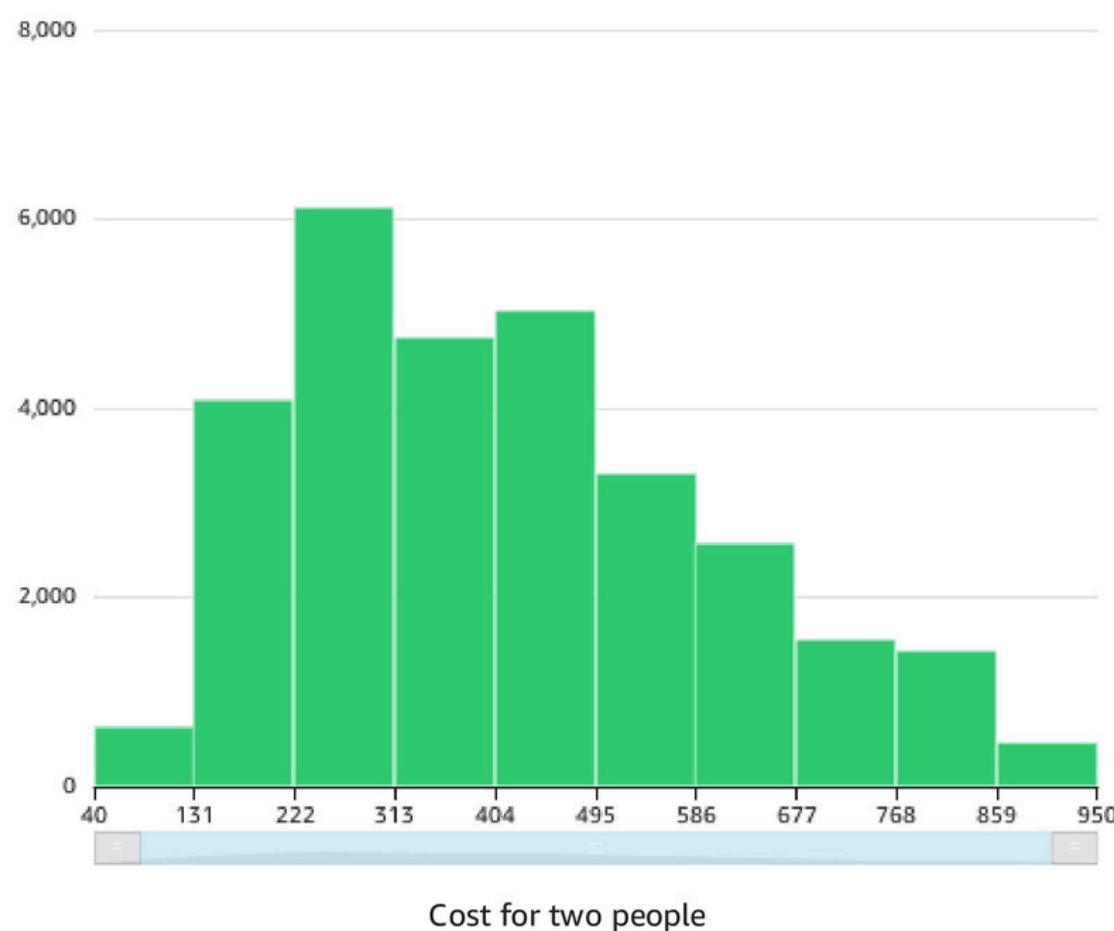
Overview of each City's Cuisine



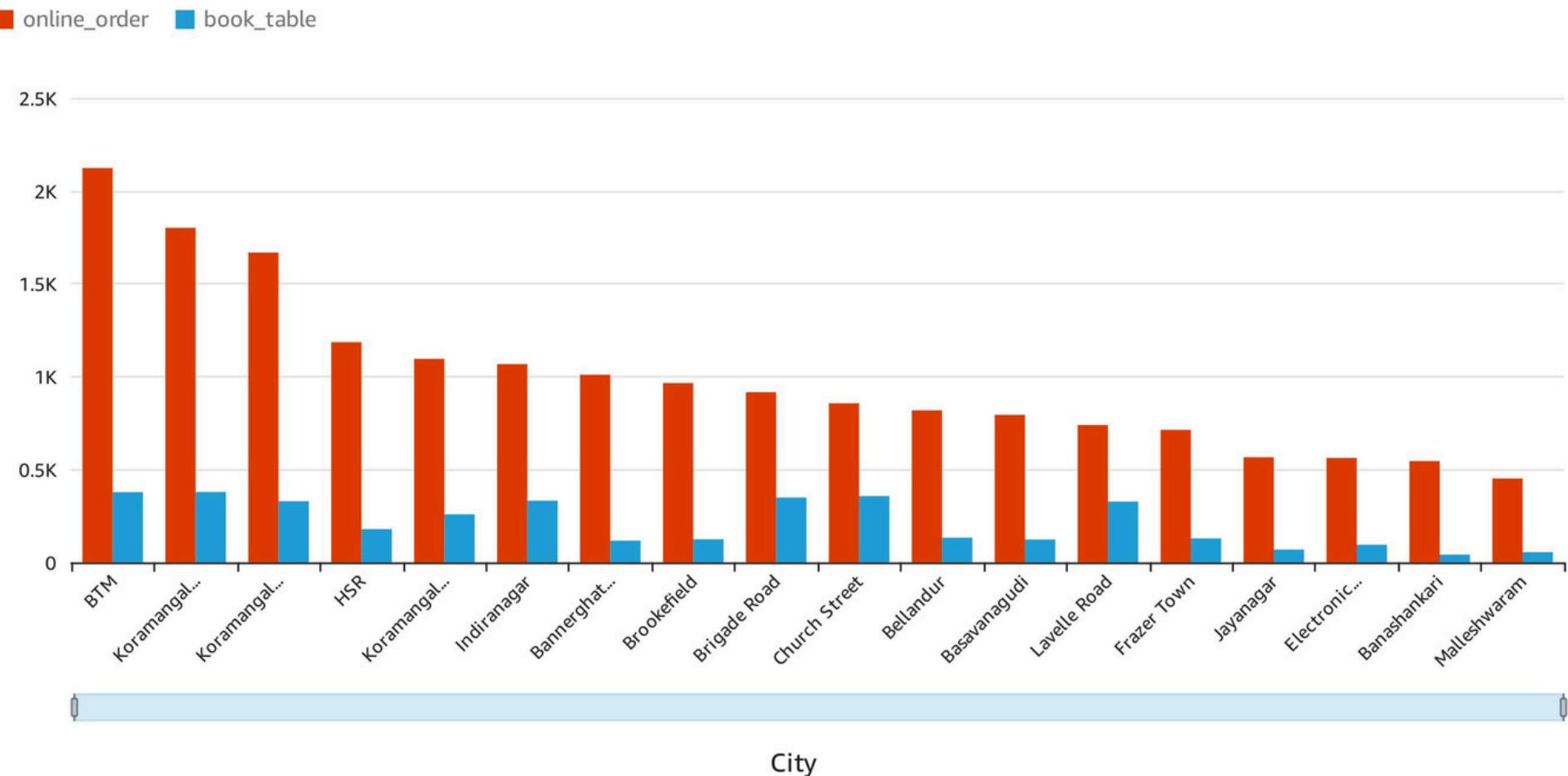
Distribution of Restaurant Category



Distribution of Price for two people



Popularity of Online Order and Book Table in each City





**THANK
YOU**

Cảm ơn các bạn
đã lắng nghe!