Problem Setting

Smartphones are undeniably an integral part of everyone's daily life, with additional capabilities being added to them frequently their importance in our lives only solidifies further. It is crucial to analyze their influence in all sectors of business so as understand how it can used to their advantage. The analysis for the purpose of this project will be centered around how smartphone usage affects the tourism industry, specifically how they affect the behavior of international tourists. Usage can be grouped into several forms such as social media, internet access, mobile operator used, etc. The goal of the analysis is the understand the extent of impact that smartphone usage has on tourists.

Data Source & Description

The datasets that are used for this project are the following:
1. Questionnaire Data: Rusdi, Jack Febrian (2019), "Smartphone usage and International Tourist Behaviour", Mendeley Data, V1, doi: 10.17632/zwzb8hzc9j.1
The data in this data set was collected through the means of a questionnaires filled in by 302 participants from 52 countries who had travelled to the city of Bandung, Indonesia.

2. Trip Advisor Bandung, Indonesia "Things to Do" page: A web scraper tool was used to collect data about the 30 most popular things to do in the city of Bandung.

3. Trip Advisor Bandung Indonesia "Popular Hotels" page: A web scraper tool was used to collect data about the 30 most popular hotels in the city of Bandung.

Trip Advisor data was used as it was a source of information that the tourists looked at before making decisions about where they would stay and what they would do during their travel.
The questionnaire data is in .csv format where as the trip advisor data is in .json format.
The data can he found here.

Problem Definition

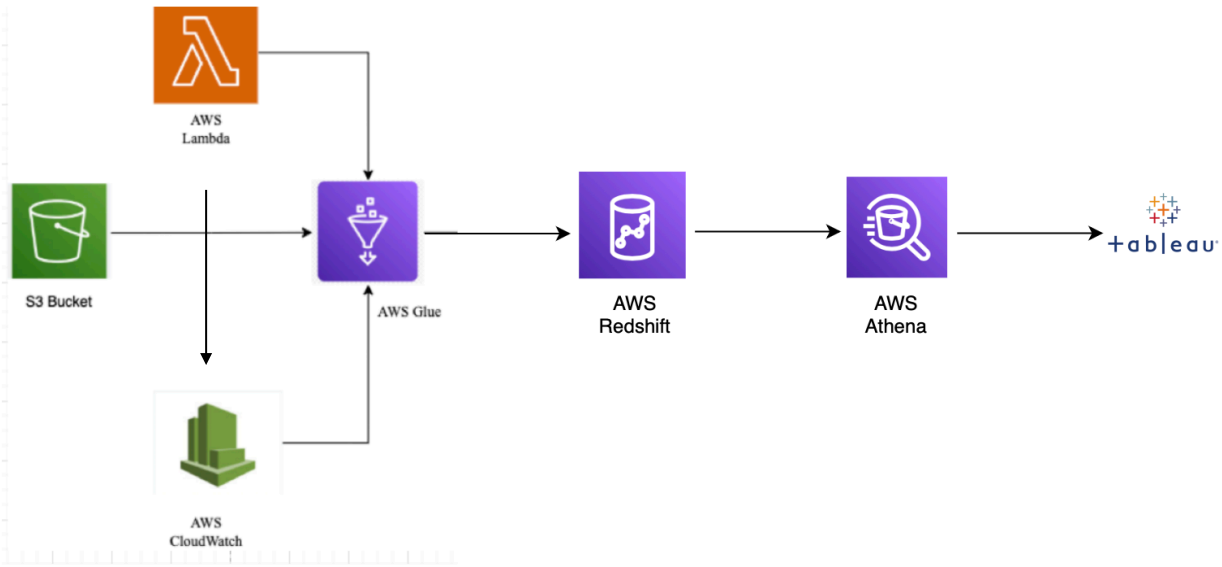
The raw data provides information about how tourists behave during the trip with respect to internet and smartphone feature usage as well as their behavior during the planning of the trip, specifically in relation to Trip Advisor reviews that may impact their decision. Dimensions derived from the data include:

- stat (**stat_id**, operator, number_of_smartphone_functions, number_of_locations, purpose)
- social_media (**scomed_id**, fb_usage, ig_usage, twitter_usage, youtube_usage)
- person (**person_id**, age, gender, education, country, *stat_id, scomed_id*)
- hotel (**hotel_id**, name, no_of_ratings, rating)

• place (**place_id**, tourist_attraction, type, no_of_ratings, rating)
• behavior (***person_id, hotel_id, place_id,*** internet_daily_usage, ratings_per_review_hotel, ratings_per_review_place)

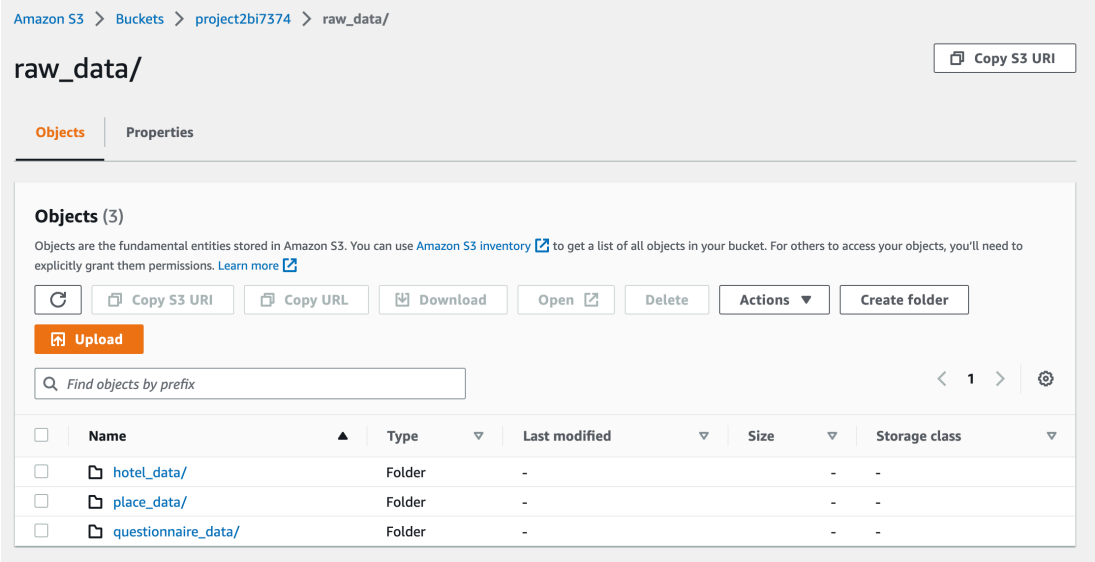
It is known that tourism can have a momentous effect on a country's economy and gaining insight into how it this be leveraged using smartphones as well as provide the tourist an enjoyable experience is ultimately a mutually beneficial endeavor.
The goal is to analyze the relationship between smartphone usage and international tourist behavior in the city of Bandung, Indonesia, and to understand the extent of the impact that smartphone usage has on tourists. This will be achieved by analyzing the questionnaire data and Trip Advisor data to gain insights into how tourists use their smartphones during their trip, how they use social media, how they plan their trip using online resources, and how they rate hotels and tourist attractions.

Architecture Diagram



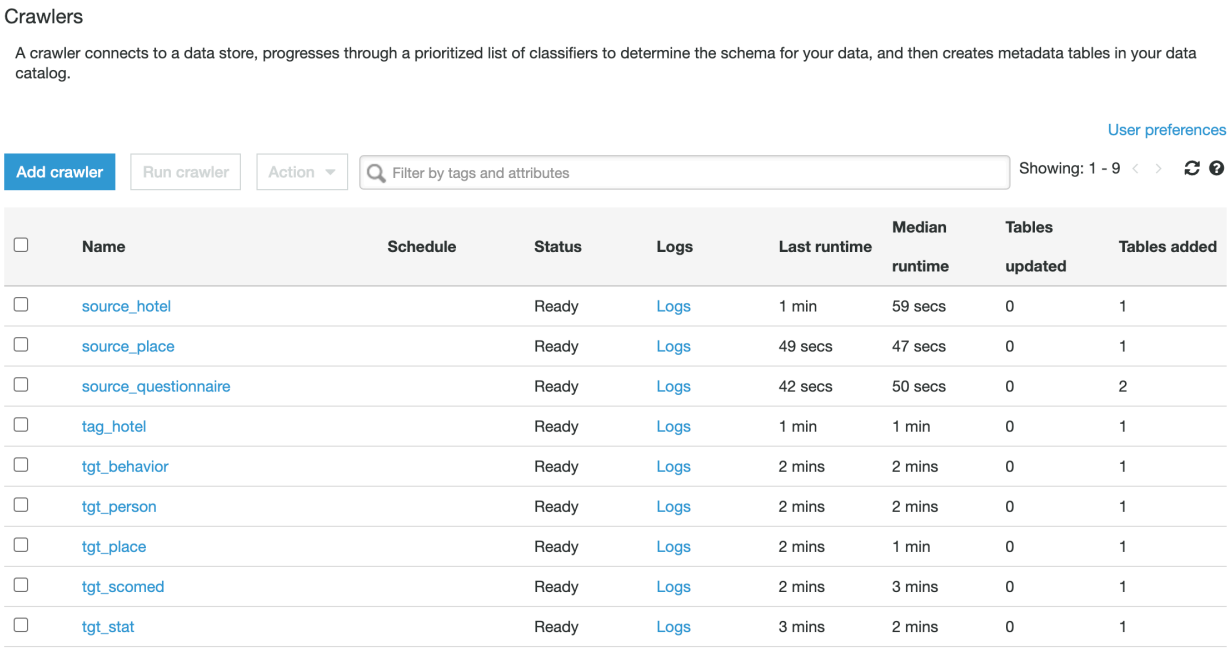
Storage

The raw data sources (questionnaire.csv, hotel.json, place.json) were stored in an S3 bucket:



Ingestion

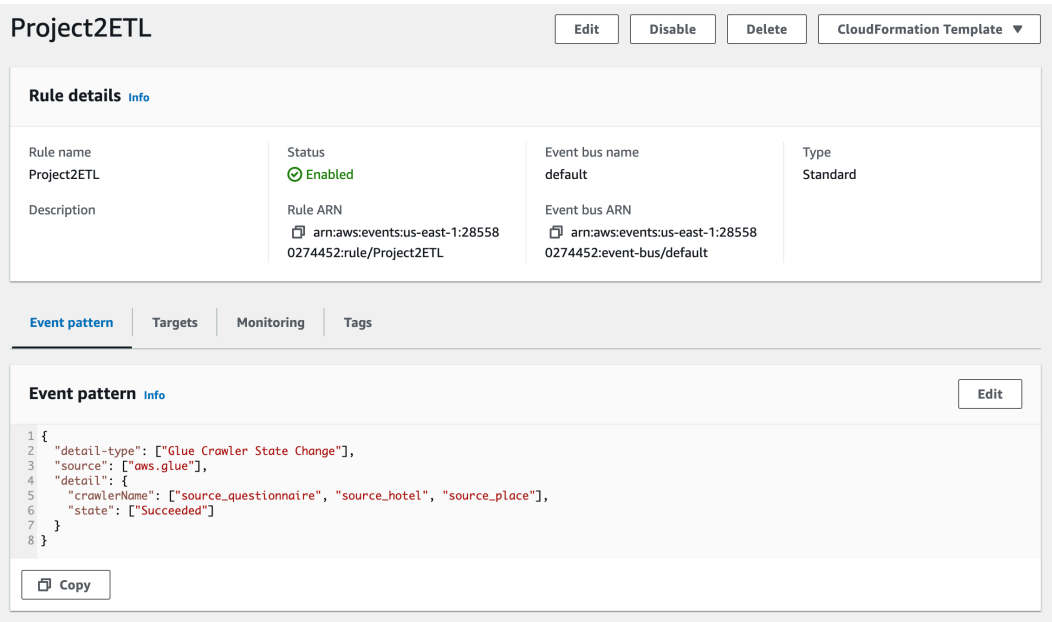
The source crawlers source_hotel, source_place, source_questionnaire are created to get the schema of the source data, target crawlers are created to get the schema of the warehouse tables.



Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
source_hotel		Ready	Logs	1 min	59 secs	0	1
source_place		Ready	Logs	49 secs	47 secs	0	1
source_questionnaire		Ready	Logs	42 secs	50 secs	0	2
tag_hotel		Ready	Logs	1 min	1 min	0	1
tgt_behavior		Ready	Logs	2 mins	2 mins	0	1
tgt_person		Ready	Logs	2 mins	2 mins	0	1
tgt_place		Ready	Logs	2 mins	1 min	0	1
tgt_scomed		Ready	Logs	2 mins	3 mins	0	1
tgt_stat		Ready	Logs	3 mins	2 mins	0	1

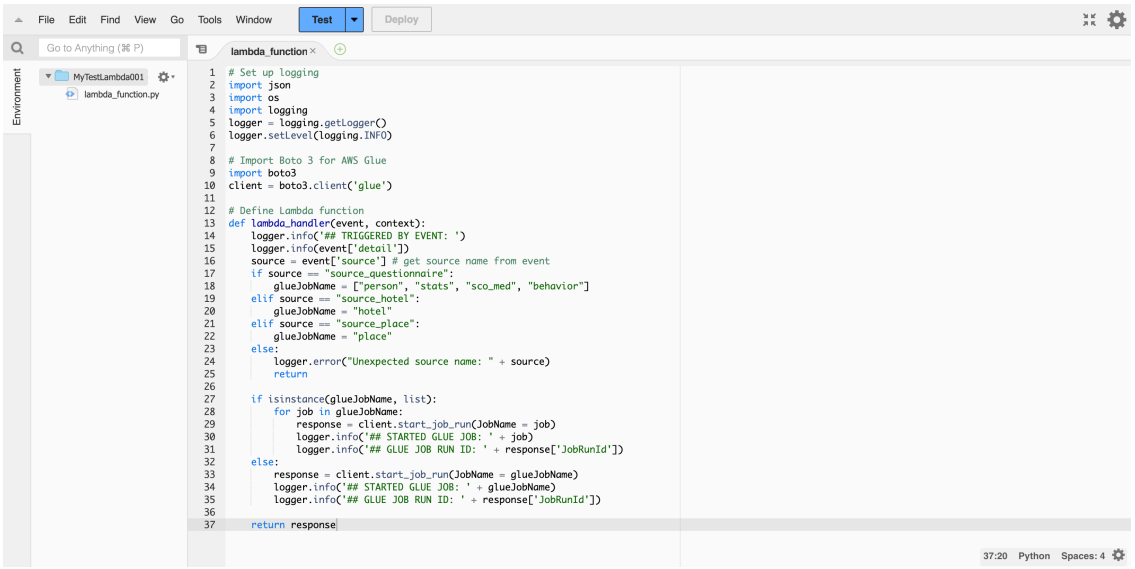
The source crawler names are used to create an event in CloudWatch to collect and log metric

data from the



different sources.

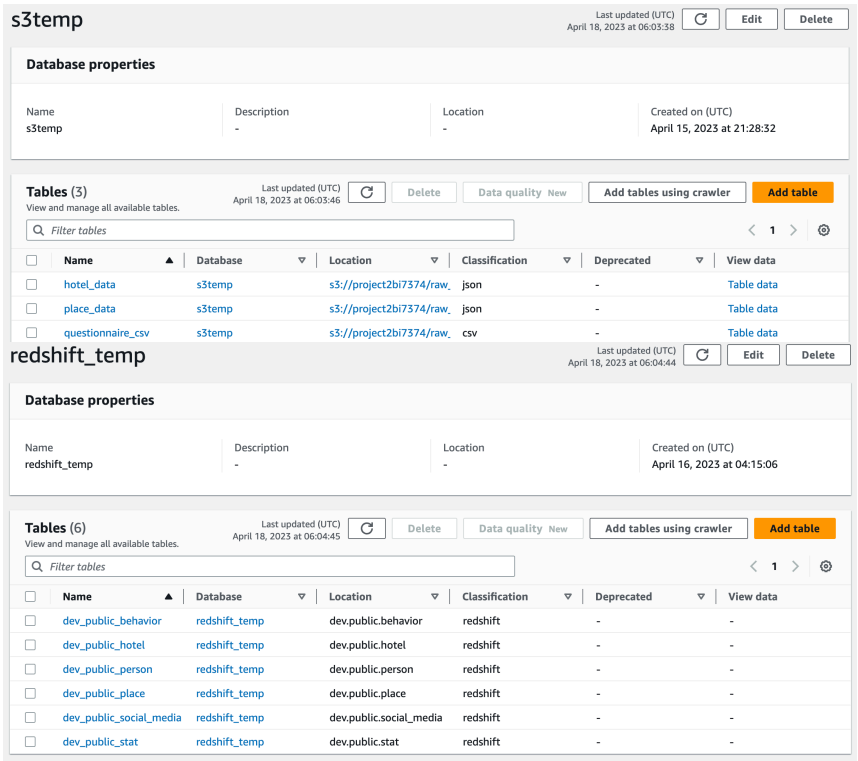
The CloudWatch event triggers an AWS Lambda function



This lambda function will invoke the jobs “person”, “stats”, “sco_med”, and “behavior” (jobs whose data source is the questionnaire data file) if source_questionnaire crawler is run as it implies that changes have been made to the source questionnaire data. Similarly, changes to hotel data triggers the job “hotel” and changes to place data triggers to job “place”

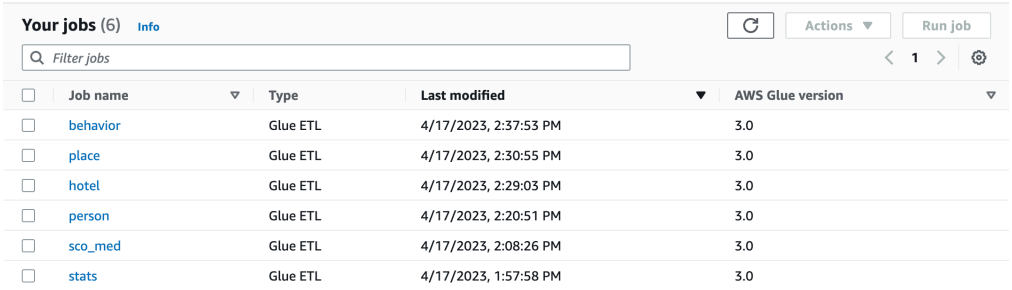
ETL Jobs

To load data from the AWS Glue Data Catalog, data source crawlers are run so as the get the schema of the source data, then the required data tables are warehouse which is created in Redshift

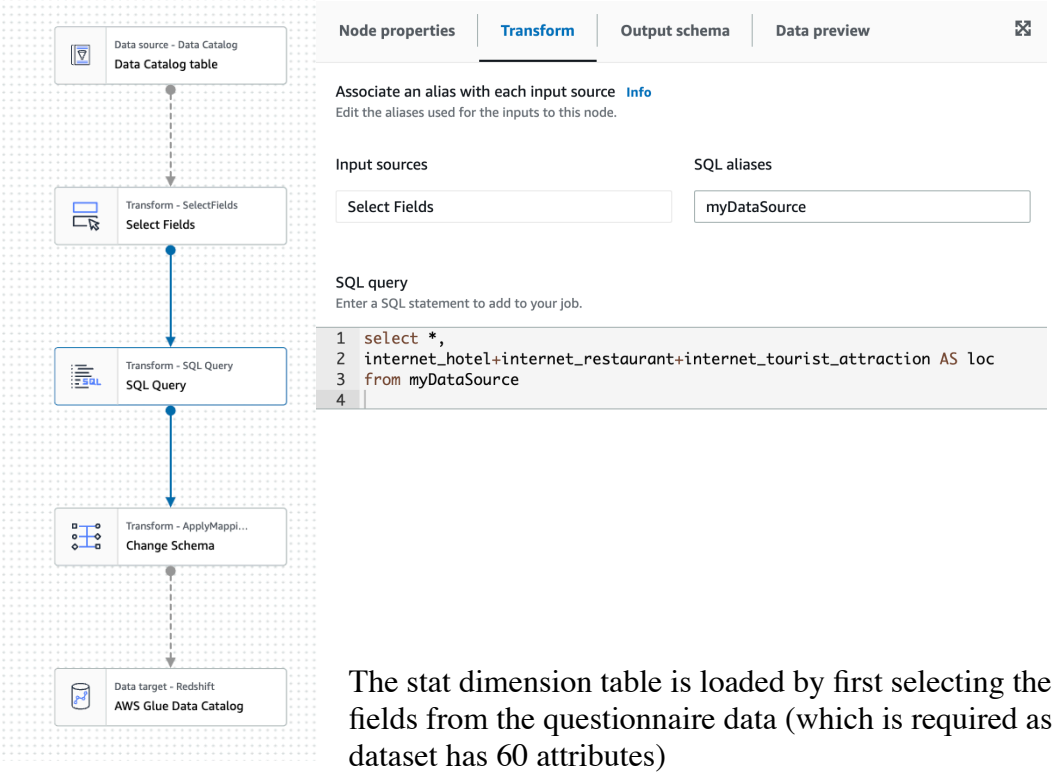


then crawled as well. The S3 bucket data is stored in a AWS Glue database “s3temp” and the Redshift table schema is store in another Glue database “redshift_temp”

We use the visual job editor to configure nodes. We are reading data from source location, then applying a transform to the data and loading the data for AWS Redshift. The following ETL jobs were used to load data into the warehouse tables:



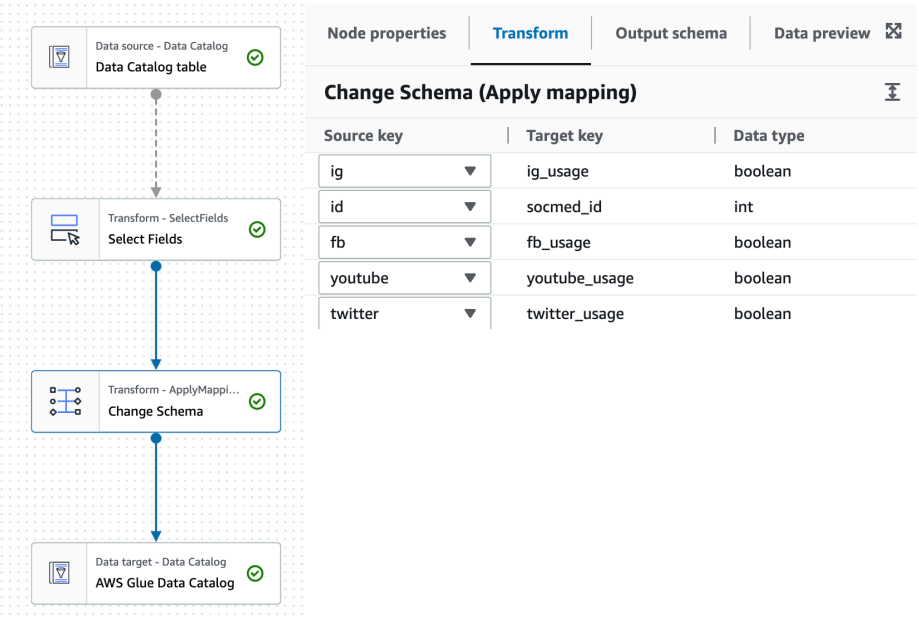
stat:



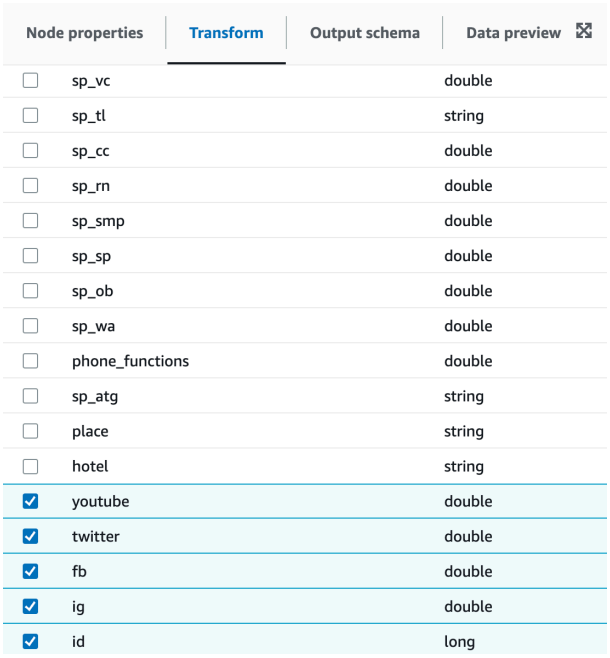
The stat dimension table is loaded by first selecting the required fields from the questionnaire data (which is required as this dataset has 60 attributes)

Then, an SQL statement is used to add the boolean tables internet_hotel/internet_restaurant/internet_tourist_attraction. These are the various locations where the tourist uses data, by adding them we get count of the number of locations a particular tourist uses internet.

sco_med:

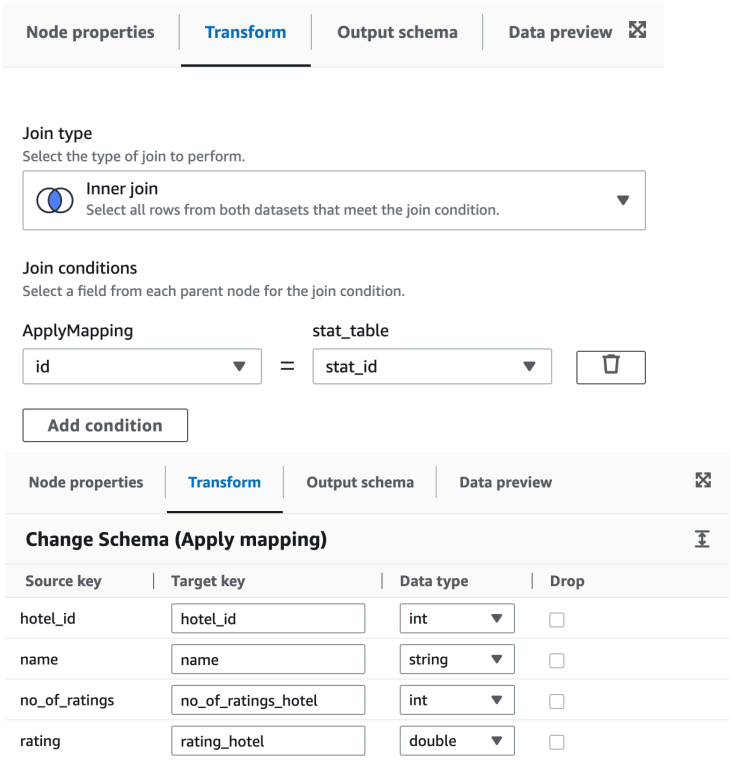


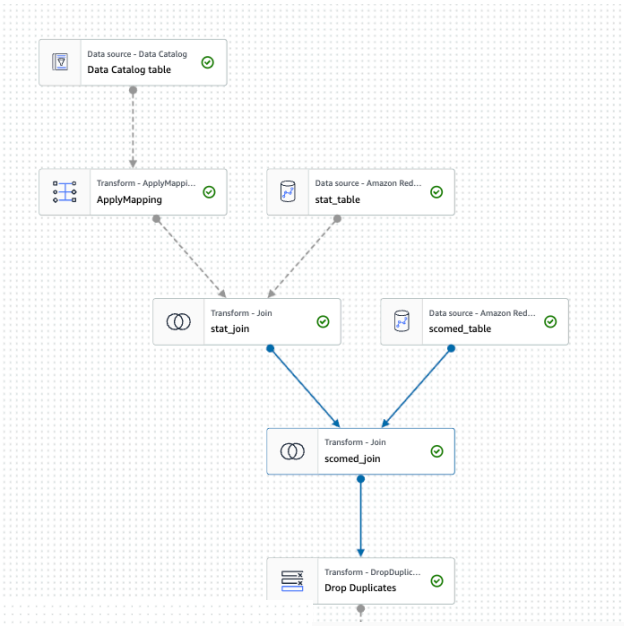
The social_media table is loaded by mapping the schema to the appropriate columns and data type in the warehouse.



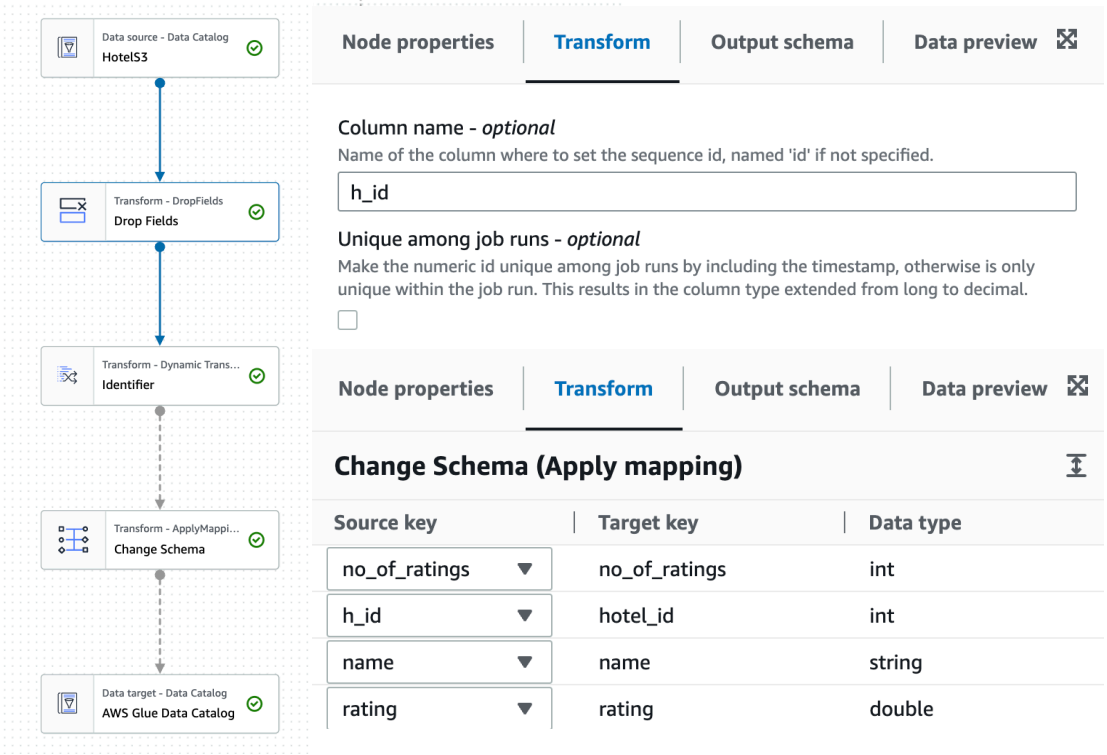
person:

For the dimension person, we are reading the date, then transform it by mapping the data and performing two join to combine multiple tables. We then drop the fields that are not required, mapping the changed schema and then loading the data. The person table contains a hierarchy to the stat and social_media tables, hence it is joined with the respective tables from redshift so as to match each person to a stat_id and scomed_id



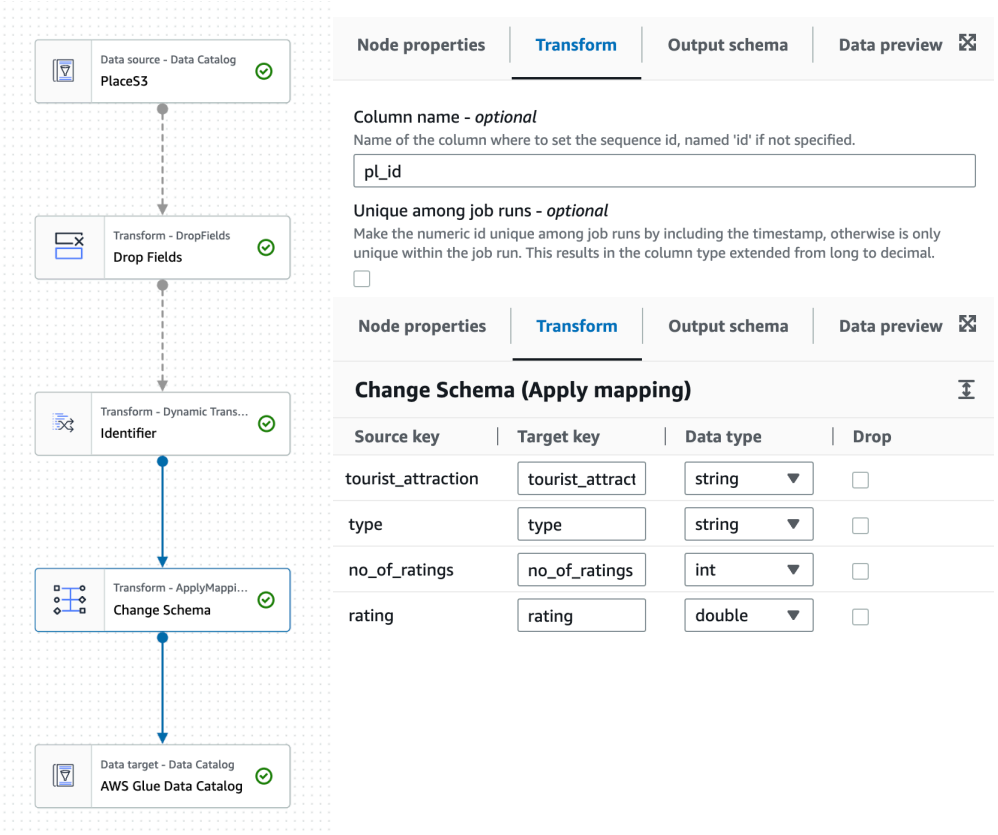


hotel:



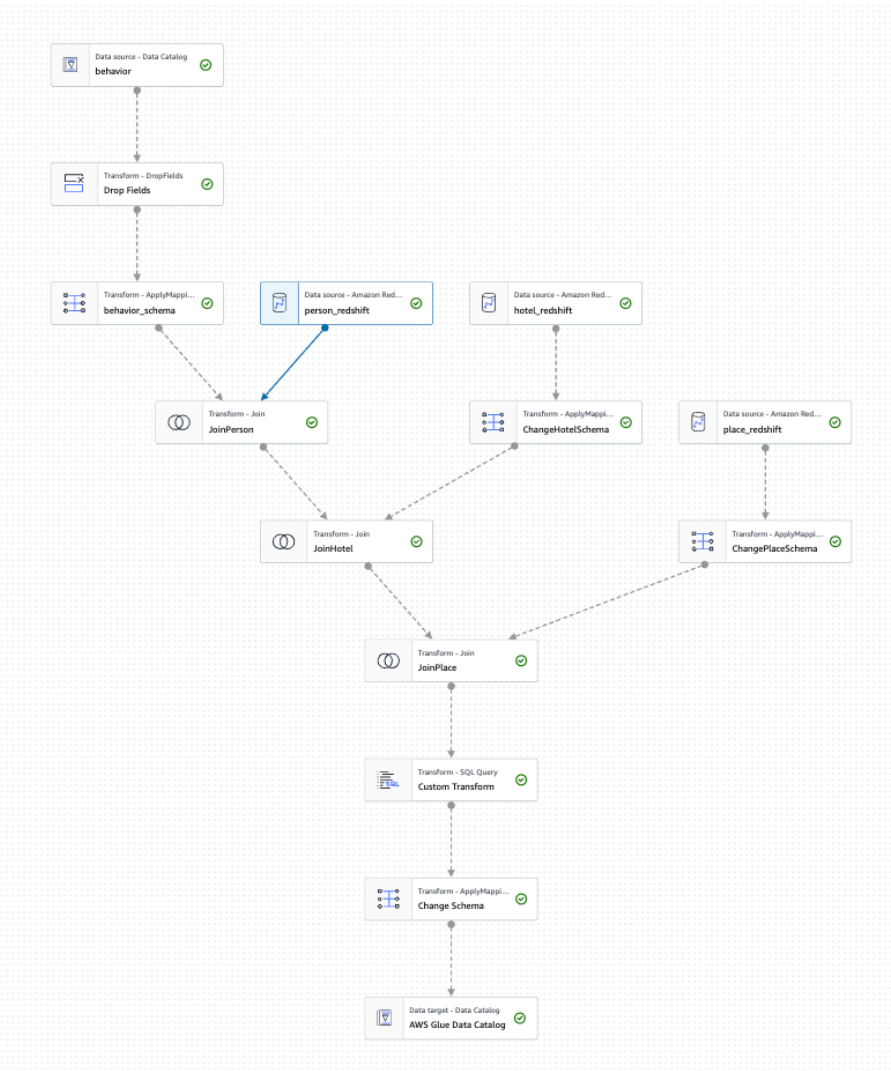
Hotel data is added to the redshift table first by dropping unnecessary field and then adding a unique identifier to each hotel using the transform ‘Identity’ then applying mapping.

place:



Place data is loaded in a similar way to hotel.

behavior:



Node properties | Transform | Output schema | Data preview

Change Schema (Apply mapping)

Source key	Target key	Data type
internet_daily_usage	internet_daily_usage	double
rpr_hotel	rating_per_review_hotel	double
hotel_id	hotel_id	int
rpr_place	rating_per_review_place	double
place_id	place_id	int
person_id	person_id	int

SQL query
Enter a SQL statement to add to your job.

```
1  select *,
2      round(rating_hotel / no_of_ratings_hotel, 4) AS rpr_hotel,
3      round(rating_place / no_of_ratings_place, 4) AS rpr_place
4  from myDataSource
```

Change Schema (Apply mapping)

Source key	Target key	Data type	Drop
hotel_id	hotel_id	int	
name	name	string	
no_of_ratings	no_of_ratings_hotel	int	
rating	rating_hotel	double	

Node properties | Transform | Output schema | Data preview

Join type
Select the type of join to perform.

Inner join
Select all rows from both datasets that meet the join condition.

Join conditions
Select a field from each parent node for the join condition.

JoinPerson		ChangeHotelSchema
hotel	=	name

Add condition

For the fact table behavior, data from Hotel, Place, and Person is read and an SQL statement is created to convert the data into columns “reviews_per_rating_hotel”/ “reviews_per_rating_place”

Data in their respective tables in Redshift:

Rows returned (30)

Export

Search rows

hotel_id	name	no_of_ratings	rating
0	TheGaiaHotelBandung	632	5
1	ThePapandayan	3194	5
2	GrandTjokroPremiereBandung	4156	4.5
3	POPHotelFestivalCitylink	827	4
4	ibisStylesBandungGrandCentral	773	5
5	SwissBelresortDagoHeritage	1044	4.5
6	JanevallaBandung	1206	4.5
7	BelviuHotel	1929	5
8	TheTransLuxuryHotel	2173	4.5
9	TheNaripanHotel	1117	5

Rows returned (30)

Search rows

place_id	tourist_attraction	type	no_of_ratings	rating
0	BragaStreet	Neighborhoods	1936	4
1	GeologyMuseum	ScienceMuseums	799	4
2	TransStudioBandung	AmusementampThemeParks	2189	4
3	ParisVanJava	ShoppingMalls	1283	4
4	GedungSate	ArchitecturalBuildings	527	4
5	NuArtSculpturePark	ArtGalleriesParks	227	4.5
6	MuseumofTheAsianAfricanConference	SpecialityMuseums	428	4
7	DusunBambuFamilyLeisurePark	PointsofInterestampLandmarks	792	4
8	BandungGrandMosque	PointsofInterestampLandmarksArchitecturalBuildings	616	4
9	RumahModeFactoryOutlet	FactoryOutlets	1271	4

Rows returned (302)

Search rows

stat_id	operator	no_of_smartphone_func	no_of_locations	purpose
1	local	1	0	
2	local	9	3	work
3	local	10	3	
4	local	10	2	
5	local	9	2	
6	local	4	2	
7	local	10	1	
8	local	4	3	
9	local	8	2	
10	local	7	1	

Rows returned (302)

Search rows

socmed_id	fb_usage	ig_usage	twitter_usage	youtube_usage
1		true		true
2	true	true	true	
3	true	true	true	
4	true	true		
5	true	true		true
6	true	true		
7	true			
8	true			
9	true	true		
10	true			

Rows returned (302)

Search rows

person_id	age	gender	education	country	socmed_id	stat_id
18	34	F	D	Malaysia	18	18
38	60	M	PG	Netherlands	38	38
46	23	F	PG	Kazakhstan	46	46
73	34	F	ETC	Singapore	73	73
172	25	F	D	Australia	172	172
186	27	F	D	US	186	186
263	67	M	D	Mauritius	263	263
282	43	F	D	Malaysia	282	282
28	28	M	L	Australia	28	28
58	28	F	D	UK	58	58

Rows returned (262)

Export

Search rows

person_id	place_id	hotel_id	rating_per_review_hotel	rating_per_review_place	internet_daily_usage
241	18	15	0.0079	0.0205	11
302	18	14	0.0101	0.0205	
74	18	6	0.0037	0.0205	4
192	18	4	0.0065	0.0205	5
153	18	4	0.0065	0.0205	3
254	18	20	0.125	0.0205	4
10	18	19	0.0367	0.0205	10
223	18	0	0.0079	0.0205	8
29	27	26	2.5	0.016	8
59	27	29	0.04	0.016	3

Data Visualization

Analysis for the purpose of this project can be divided into two criteria:

I. Trip Information
- Top hotels to stay and and top place to go to in the city

II. Tourist Information while on the trip
- Average number of smartphone functions used by a tourist
- Number of tourists who use social media during their travels
- Internet usage at places they travel to

1. Average Number of Smartphone functions used by the tourists:

```
SELECT AVG(no_of_smartphone_func) AS avg_phone_functions_used
FROM stat;
```

2. Total number of international tourists who use Facebook, Twitter, YouTube, and Instagram during their trip:

```
SELECT
   SUM(CASE WHEN fb_usage = true THEN 1 ELSE 0 END) AS fb_users,
   SUM(CASE WHEN twitter_usage = true THEN 1 ELSE 0 END) AS twitter_users,
   SUM(CASE WHEN youtube_usage = true THEN 1 ELSE 0 END) AS youtube_users,
   SUM(CASE WHEN ig_usage = true THEN 1 ELSE 0 END) AS ig_users
FROM social_media;
```

3. Average internet usage of tourists ordered by the type of place they visited:

```
SELECT place.type, ROUND(AVG(internet_daily_usage), 2) AS avg_usage
FROM behavior
INNER JOIN place ON behavior.place_id = place.place_id
GROUP BY place.type
ORDER BY avg_usage DESC;
```

4. Top 5 hotels to stay at based on number of ratings and of number of reviews

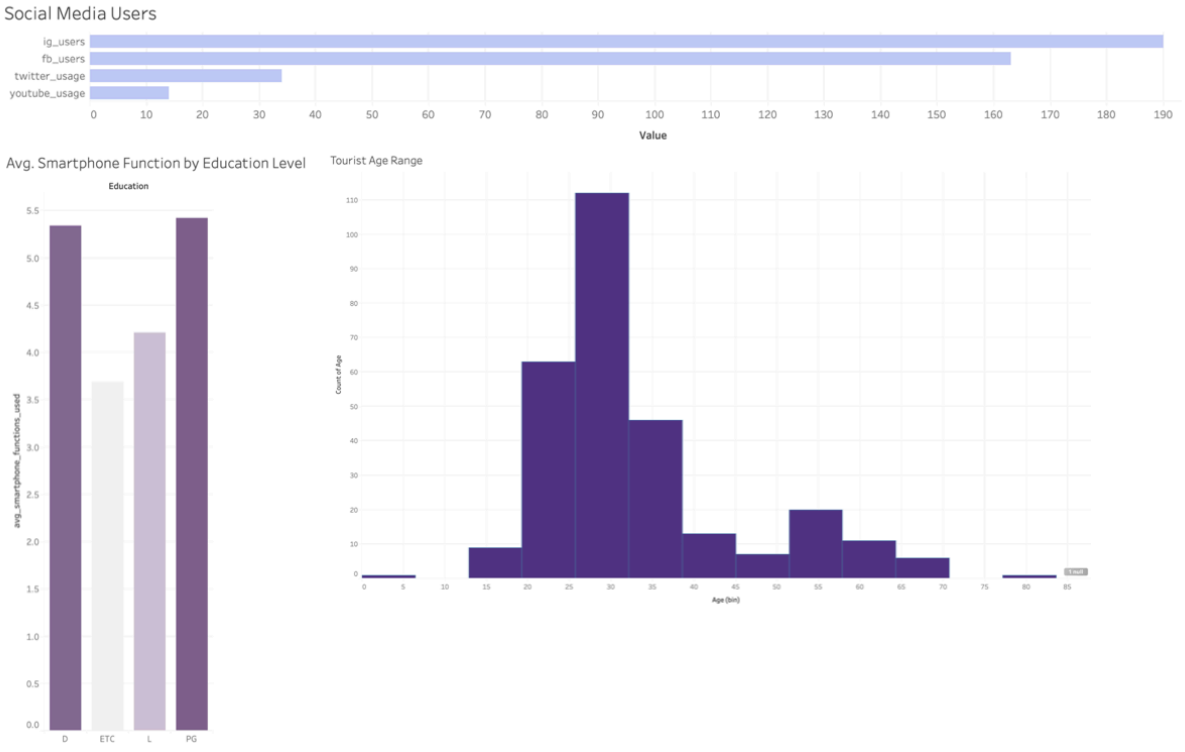
```
SELECT name, rating, no_of_ratings
FROM hotel
ORDER BY rating DESC, no_of_ratings DESC
LIMIT 5;
```

5. Top 5 places to visit based on number of ratings and of number of reviews

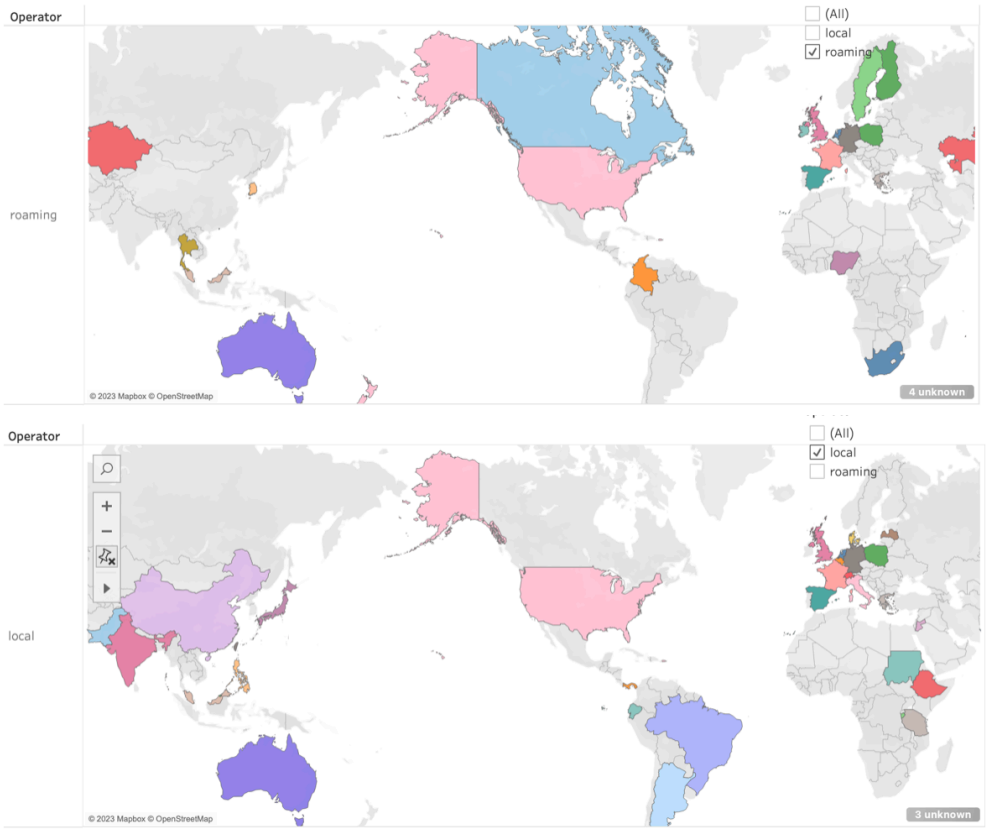
```
SELECT tourist_attraction, type, rating, no_of_ratings
FROM place
ORDER BY rating DESC, no_of_ratings DESC
LIMIT 5;
```

6. Number of tourists who visited a particular type of tourist attraction

```
SELECT place.type, COUNT(DISTINCT behavior.person_id) AS count
FROM place
INNER JOIN behavior ON place.place_id = behavior.place_id
GROUP BY place.type
ORDER BY count DESC;
```

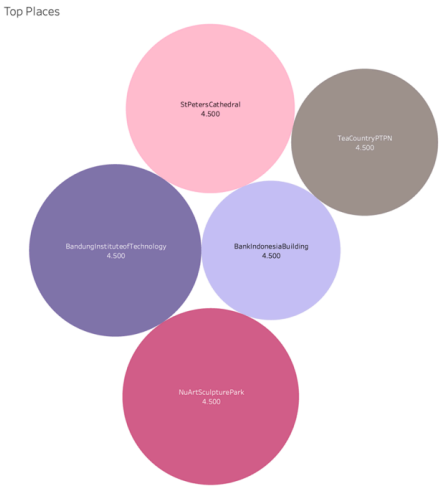


Dashboards


Operator
(All)
local
roaming
roaming
© 2023 Mapbox © OpenStreetMap
4 unknown
Operator
(All)
local
roaming
local
© 2023 Mapbox © OpenStreetMap
3 unknown

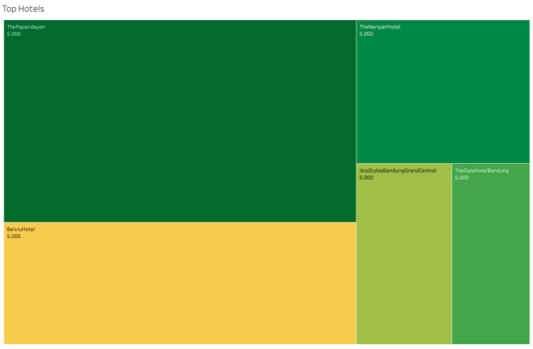
Project Results

1. Tourist



Internet Usage Hours by Place

Type (Avginternerus..	
DepartmentStoresShoppi..	11.330
HotSpringsampGeysers	8.800
Waterfalls	7.800
ArchitecturalBuildings	7.750
Neighborhoods	7.480
PointsofInterestampLand..	7.250
FactoryOutlets	7.000
ScienceMuseums	6.530
SpecialityMuseums	6.470
ScenicWalkingAreas	6.430
Parks	6.000
PointsofInterestampLand..	5.830
PointsofInterestampLand..	5.820
NatureampWildlifeAreas	5.600
PointsofInterestampLand..	5.500
ShoppingMalls	5.100
AmusementampThemePa..	5.080
ArtGalleriesParks	3.500
ReligiousSites	2.500
NeighborhoodsHistoricW..	2.500



Tourist Attraction (Bestplaces)
- BandungInstituteofTechnology
- BankIndonesiaBuilding
- NuArtSculpturePark
- StPetersCathedral
- TeaCountryPTPN

Name (Besthotels)
- BelviuHotel
- ibisStylesBandungGrandCentral
- TheGaiaHotelBandung
- TheNaripanHotel
- ThePapandayan

demographic: The majority of tourists are in their 30s and have a diploma, and they use social media platforms like Facebook and Twitter to share details of their trip. This information can be helpful for businesses that cater to tourists, such as hotels, restaurants, and tour operators, to target their marketing and advertising efforts towards this demographic.

2. Local operators vs. roaming: Tourists from India and the US are more likely to use local operators instead of roaming, while tourists from Australia and Canada are more likely to use roaming. This suggests that tourists from different regions may have different preferences and needs when it comes to mobile network services. Businesses in the telecommunications industry could use this information to tailor their services to better suit the needs of different tourist groups.

3. Top places and hotels: The top places and hotels have been calculated based on both the number of ratings and the highest reviews. This information can be useful for tourists who are looking for popular and highly rated places to visit or stay, as well as for businesses that want to improve their ratings and reviews by providing quality services and experiences to their customers.

4. Internet usage: The average hours of internet usage by places visited shows that internet usage is lowest in places like religious sites and highest in places like shopping malls. This information can be helpful for businesses that offer free Wi-Fi or other internet-related services, as they can target their efforts towards places where internet usage is higher. Additionally, it suggests that tourists may have different needs and preferences when it comes to internet access depending on the places they visit.