# Extract, Transform And Load

**Technical Report**

**Airbnb And Hotels**

Extract, Transform And Load 1

1 INTRODUCTION 2

1.1 Summary 2

1.2 Scope 2

1.3 Technologies and Resource contributions 3

1.4 Definitions, Acronyms and Abbreviations 3

2 ETL DETAILS 3

2.1 Data Import/Extract Sources and Methods 3

2.2 Data Acquisition 4

2.2.1 Washington DC 4

2.2.2 New Orleans 4

2.3 Data Transformation 4

2.3.1 Washington DC 4

2.3.2 New Orleans. 5

2.4 Data Integrity 5

2.5 Data Refresh Frequency 5

2.6 Data Security 5

2.7 Data Loading and Availability 5

# INTRODUCTION

Data from multiple sites is extracted as a csv files for a given region to compare hotels and airbnb data. This data may be used to see if number of hotels in a city will have impact on airbnb in that city.

## Summary

The objective of this project is to extract data from multiple sources that consists of airbnb listing and hotel listing for a given city. The data is in csv and downloaded without knowing the quality. We want to analyze the data and clean the data and drop columns with not required and null values data. The scrubbed data is loaded into a database for future analysis.

## Scope

Data is extracted from multiple sources as below

1. AirBNB information: <http://insideairbnb.com/get-the-data.html>

2. Hotels information: [https://catalog.data.gov](https://catalog.data.gov/), <https://data.world/dcopendata/hotels>

From the data source webpage a query is written to extract the data for both New Orleans and Washington DC, in separate csv files and download and used for the project. The scope for this section of the project is to review the data, clean the data and make sure there are no nulls values in the zipcode. Zipcode is used to compare both airbnb and hotel data for that city. Only cleaning, formatting, elimination and loading data into mysql database is in scope. Producing graphs and dashboards on a web page, is not part of the scope of his project.

## Technologies and Resource contributions

The project team consists of two members: Srinivas Mudivarthy and Hasan Kaptan. Srini is working the Washington DC specific data and Hasan is working on New Orleans specific data.

Data is downloaded as a csv file after running queries for specific cities from the web site. Panda(python) is used to load the data into dataframe, and sqlalchemy is used to load data from dataframe into local MySql database on the laptop.

## Definitions, Acronyms and Abbreviations

ETL: Extract transform and load.

SQL: Structured query language

MySql: Database to hold data in tables

# ETL DETAILS

Subsections below will describe in detail how data for the above mentioned cities is extracted, processed, cleaned, selected and how the tables and keys are created in the database and loaded into the tables.

## Data Import/Extract Sources and Methods

The team made a decision to analyze the impact the presence hotels would have impact on airbnbs in that location. The impact of the type of location: business vs tourist, with regard to the number of airbnb or hotels in that location was also studied.

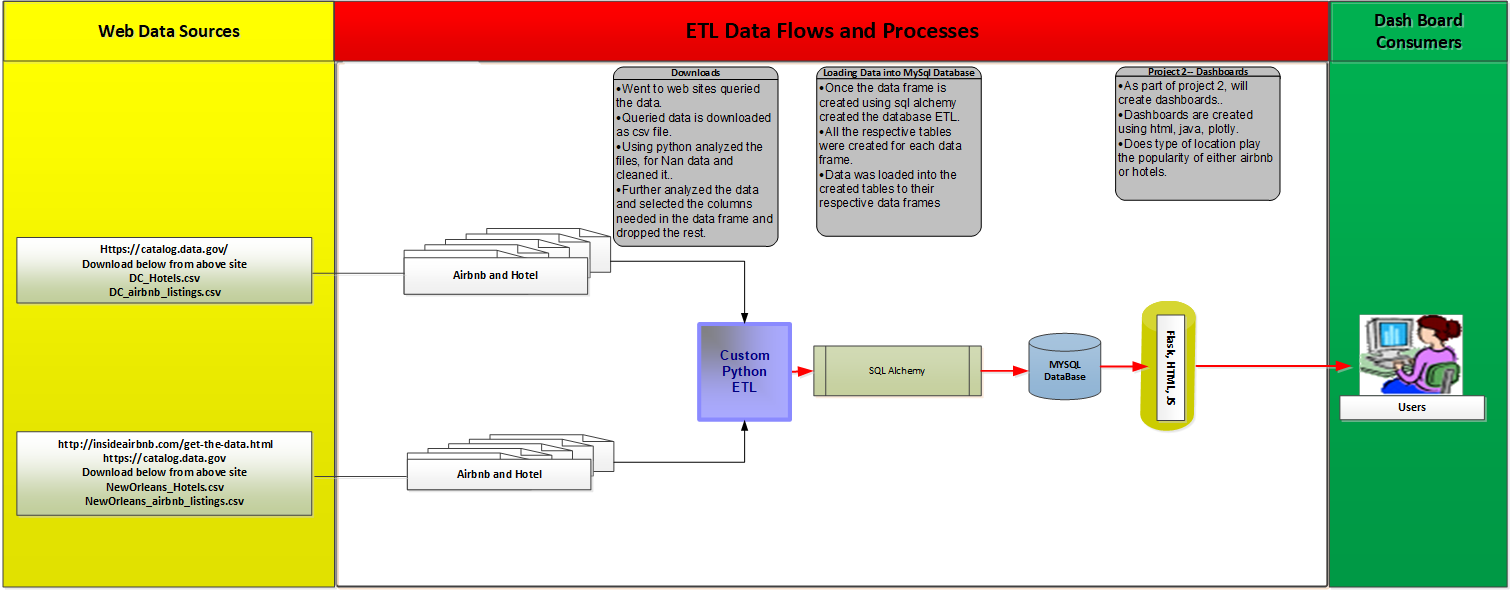
The sites below were identified and used in the query to get data for specific city. The results of the query were download as a csv file.

<http://insideairbnb.com/get-the-data.html>

[https://catalog.data.gov](https://catalog.data.gov/)

<https://data.world/dcopendata/hotels>

The diagram below shows the overall architecture and data flows for this project.



## Data Acquisition

Data acquisition from web sites are discussed below for cities of New Orleans and Washington DC.

### Washington DC

After exploring various options on Google for airbnb and hotel data for Washington DC. The site below was selected to pull the from:

[https://catalog.data.gov](https://catalog.data.gov/)

The query is built to choose the data and download the result set into .csv file. The downloaded data had 106 columns and close to 10000 records for airbnb and 11 columns and 150 records for hotels. This data may gets updated once a year from other sources by the user community. This source didn’t have the capability nor the frequency to automate the process of extracting the data. The query cannot be saved unless the user has an account/profile.

### New Orleans LA

After exploring various options on Google for airbnb and hotel data for New Orleans LA. The site below was selected to pull the from:

<https://catalog.data.gov>

The query is built to choose the data and download the result set into .csv file. The downloaded data had 80 columns and close to 70000 records for airbnb and 11 columns and 344 records for hotels. This data may gets updated once a year from other sources by the user community. This source didn’t have the capability nor the frequency to automate the process of extracting the data. The query cannot be saved unless the user has an account/profile.

## Data Transformation

As described above the data set for both cities are collected from multiple websites.

### Washington DC

Data is published to the [https://catalog.data.gov](https://catalog.data.gov/) site from the airbnb web site; hence the data is reliable and complete. Similarly data for hotels in the DC area are collected from the same site.

* Airbnb:
  + Airbnb data has 106 columns, from those 20 columns were chosen which were more than required. Important columns were zipcode and location information.
  + Bad data was found in 6 columns, after further analysis these columns were dropped from the Dataframe.
  + The Dataframe was analyzed to check if there were any null values and few null values for found. But emphasis was made to eliminate data which has null zipcodes and those records were rejected and a final clean dataframe was created to be loaded into database table.
* Hotels:
  + Hotels had lesser data and columns compared to airbnb data. Hotels had 11 columns that included zipcode and geo-location data. Unfortunately the hotels data didn’t have city, state, country code and country information.
  + Missing city, state, country code and country columns were created in the table in the database and default values of city, state, country code and country were put as part of the table creation, so once the data got loaded into the table those default values got populated instead of null, since the source did not have values. This was done to make data consistent with airbnb data. This is important for future analysis.

### New Orleans.

Data is published to the [https://catalog.data.gov](https://catalog.data.gov/) site from the airbnb web site; hence the data is reliable and complete. Similarly data for hotels in the New Orleans area are collected from the same site.

* Airbnb:
  + Airbnb data has 80 columns, from those 9 columns were chosen which were more than required. Important columns were zipcode and location information.
  + Bad data was found in 1 column, after further analysis these columns were dropped from the Dataframe.
  + The Dataframe was analyzed to check if there were any null values and few null values for found. But emphasis was made to eliminate data which has null zipcodes and those records were rejected and a final clean dataframe was created to be loaded into database table.
* Hotels:
  + Hotels had lesser data and columns compared to airbnb data. Hotels had 11 columns that included zipcode and geo-location data. Unfortunately the hotels data didn’t have city, state, country code and country information.
  + Missing city, state, country code and country columns were created in the table in the database and default values of city, state, country code and country were put as part of the table creation, so once the data got loaded into the table those default values got populated instead of null, since the source did not have values. This was done to make data consistent with airbnb data. This is important for future analysis.

## Data Integrity

The data sources are reliable since they are from airbnb and government sites. There were some special characters in non important columns and hence they were dropped. The data sources are updated on a yearly basis as this data has listing information. Currently a scheduled data pull is not established.

In the future once the system is established on a server as the data is loaded into the table email triggers will be sent to the data consumers of the refreshed data.

## Data Refresh Frequency

As of now no refresh frequency is established.

## Data Security

As of now no data security is established.

## Data Loading and Availability

In future a dimensional data model will be created and loaded as per the user requirements. The data will be accessed by the users through a web interface which will have dashboards.