**Gathern Documentation**

**Introduction**

It’s a platform licensed and approved by the Ministry of Tourism to book holiday houses around Saudi Arabia. Users can book villas, chalets, apartments, farms, camps, and more holiday houses around Saudi Arabia.

The platform operates in more than 42,000 properties in 207 cities and provinces around Saudi Arabia.

It was founded in 2016 with the aim to give users the ability to book from the platform private villas, chalets, rest houses, camps, resorts, farms and other places suitable for gatherings.

**Data Review**

The dataset has around 66k records, it has 35 columns and it has data for only 2022 Aug, also dataset is only for Riyadh.

We selected the popularity column, and it has three categories, which it High, Medium, Low.

**Business Problem**

We want to help Gathern to have more popular holiday houses.

More popular holiday houses will impact the platform repetition positively due to providing more popular and demanding holiday houses in the region.

**The MapReduce Program**

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

**Approach**

First, we toke the popularity column and put it in text file and then we import Joblib and load the column text file then we wrote the Map & Reduce in a Jupyter Notebook using Job Lib library then we saved as a script then we did upload it in Hadoop.

**Results**

The result we got it is that medium category has 33472 and high category has 16739 and low category has 16734. Therefore most of the holiday houses in Gathern datasetaremedium category.

**What can be improved**

Th improve we can do is to take more columns and specify what type of holiday houses are in medium or high or low category and specify what is the features that effect the popularity.