Properties Rate Analysis

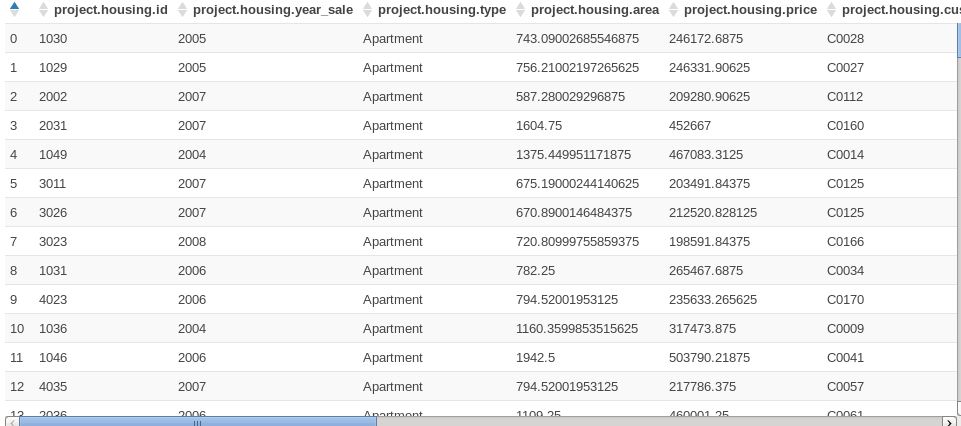
**Problem Statement**

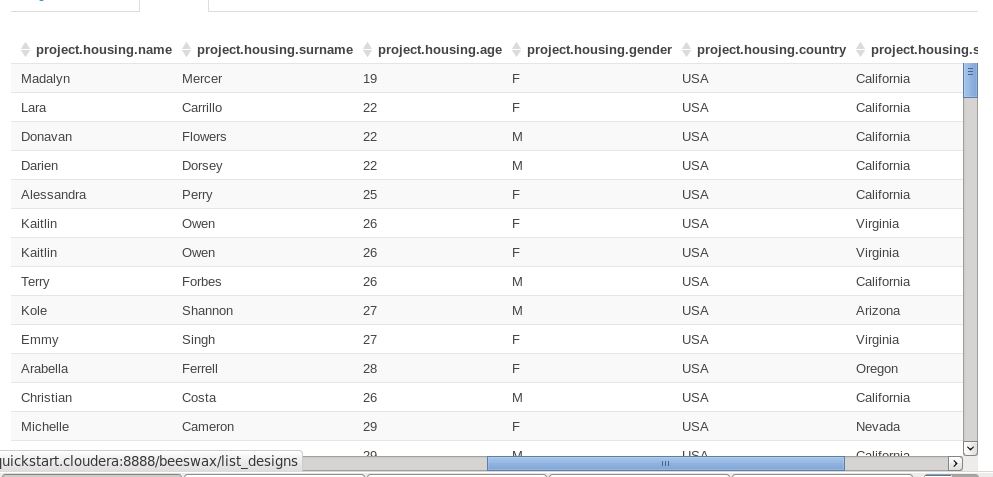
The various property and their prices attract a large amount of investor category audience. The engagement among the investors could be tracked using User-Generated Content (UGC). This report aims to provide a deeper analysis of the flow of worldwide interest in the price analysis, locality analysis, rating analysis, sentiment analysis, etc. This data is very big and unstructured form.

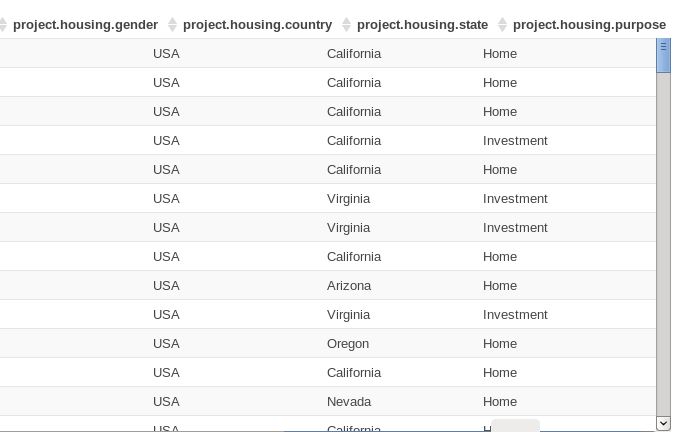
Apache facilitates with open-source software frameworks for distributed storage and processing big data information using the MapReduce environment. Tools like Apache Hadoop, Apache Pig, Apache Sqoop, Hive, and Apache HBase can enable the real-time analysis considering the aggregation of the results and user sentiments before, during, and after the episode has been broadcasted. This can help us to make strategies for making new TV shows depending on the viewer’s rating.

**Methodology**

**Dataset:**

****





The dataset consists of the following attributes:

**housing\_id:** House id.

**year\_sale**: year of selling.

**housing\_type:** type os house.

**housing\_area:** area code in which the house is located.

**housing\_price:** Price of the house.

**housing\_name**: Name of the house owner.

**housing\_surename**: Surename of the house owner.

**housing\_age**: age of the house owner.

**housing\_gender**: gender of the house owner.

**housing\_country**: country in which the house is located.

**housing\_state**: State in which the house is located.

**housing\_purpose**: the purpose of the home.

**Technology Used:**

We are using Hive to create databases and HIVEQL to run the complex queries. In this project, we are using big data. The term ‘Big Data’ is used for collections of large datasets that include huge volume, high velocity, and a variety of data that is increasing day by day. Using traditional data management systems, it is difficult to process Big Data. Therefore, we use a framework called Hadoop to solve Big Data management and processing challenges. Hadoop is an open-source framework to store and process Big Data in a distributed environment. Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy. Initially, Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open-source under the name Apache Hive. It is used by different companies. For example, Amazon uses it in Amazon Elastic MapReduce. It stores schema in a database and processes data into HDFS.It provides SQL type language for querying called HiveQL or HQL. It is familiar, fast, scalable, and extensible.

We are using Cloudera to implement Hive. Cloudera is the Open Source Distribution of Apache Hadoop and related projects. It delivers the core elements of Hadoop scalable storage and distributed computing – along with additional components such as a user interface, plus necessary enterprise capabilities such as security, and integration with a broad range of hardware and software solutions. They develop a Hadoop platform that integrates the most popular Apache Hadoop open-source software within one place. Hadoop is an ecosystem and setting a cluster manually is a pain. Basically going through each node, deploying the configuration through the cluster, deploying your services, and restarting them on a wide cluster is a major drawback of distributed systems and requires a lot of automation for administration. Cloudera developed a big data Hadoop distribution that handles installation, updates on a cluster in a few clicks. They also integrate regularly Apache open source or custom made software to ease the life of Hadoop sysadmin and dev. To sum up, Cloudera is a major actor in the big data industry that was able to provide a product that eases Hadoop integration and that provides professional services to integrate the solution.

Hive provides a CLI to write Hive queries using Hive Query Language (HiveQL). Generally, HQL syntax is similar to the SQL syntax that most data analysts are familiar with. Hive's SQL-inspired language separates the user from the complexity of Map Reduce programming. It reuses familiar concepts from the relational database world, such as tables, rows, columns, and schema, to ease learning. Most interactions tend to take place over a command-line interface (CLI). Hive provides a CLI to write Hive queries using Hive Query Language (Hive-QL). Generally, HiveQL syntax is similar to the SQL syntax that most data analysts are familiar with. Hive supports four file formats that are TEXTFILE, SEQUENCE FILE, ORC, and RCFILE (Record Columnar File). Hive provides Built-in operators for Data operations to be implemented on the tables present inside the Hive warehouse. These operators are used for mathematical operations on operands, and it will return specific values as per the logic applied.

**Hardware Configuration:**

**● Local Machine Configuration: 8 GB ram, 4 Core(s).**

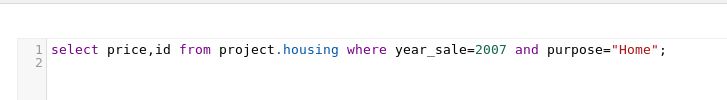
**● Virtual Machine Configuration: 4 GB ram, 2 Core(s).**

**Implementation**

1. **Loading Data**

****

1. **Query analytics and Visualization**

****

