AGILE ANALYTICS – COMBINING MULTIPLE DATASOURCES USING SQL AND NOSQL STRUCTURES

PETER ANNABLE, ROBERT BEUTNER, BALAJI DHAMODHARAN

Abstract— By identifying in three broad aggregations of data: real estate, quality of life and demographic; a detailed agile relational database has been constructed to enable detailed queries in order to answer a specific set of questions related to the identification of prime real-estate markets in an identified metropolitan market. A variety of commercial and public data sources provided the basis of the data warehouse. Data were then processed and made available for query using MySQL as the main relational database warehouse. From there, Tableau software for data visualization generated geographic data visualizations. Further development of the data includes the integration of No SQL database data managed using MongoDB for the development of a web-based Micro-service through and API.

Index Terms—…SQL, NoSQL, Relational Database, Query, Microservice, MongoDB, MySQL, Python-eve

I. INTRODUCTION

The availability of query enabled interactive datasets have become commonplace for searching for a variety of consumer goods, hotels and real estate. The main goal of each of these services is to provide the end user with the necessary datasets in order to answer a specific set of questions related to their purchase decision. Increasing availability of data from multiple data sources allows for the complex layering of information that adds increasing dimensions to query results and more importantly answer increasingly complex questions.

In this example, we have implemented a data warehouse for use by the end user to search for real estate information pertaining to the Houston, Texas metropolitan real estate residential housing market; with the added dimensions of data related to elements of geographic context such as quality of life and demographics. In addition to home sales data, we have structured the relational data warehouse to include quality of life data: crime statistics and school district information along with demographic data collected from the 2010 US Census and the 2015 ACS (American Community Survey).

As a proof of concept; using this data through query and information visualization as well as API-based micro services, we explore answers to key questions related to the consumer search for real estate.

- Which county has the highest average home price? Lowest home price?
- What is the nature of the relationship between SAT scores and home prices?
- Does per capita income have any correlation to the house prices?
- What is the impact of population density on home price? (PopulationDensity with HomePrice comparison)
- Is the home price on Steady Increase in Houston? Was there a period when the housing price got affected (recession)?

III. DATA PREPARATION AND LOAD

The agile analytics ecosystem includes data derived from several sources and fall into three broad categories: Real Estate Sales, Quality of Life (Crime and Schools) and Demographic data. Table (1) summarizes the source, and data schema for the raw data used to create this service.

Category of Data	**Data Source**	**Data Structure**
Real Estate *Balaji Dhamodharan*	Commercial real estate data: Zillow	Relational Data
Quality of Life *Peter Annable*	Texas school district performance data: Downloaded from http://tea.texas.gov/acctres/sat_act_index.html Simplified data to only show the all students groups, and removed some columns, and renamed others to harmonize with other sources Crime data for Houston 2016 data: http://www.houstontx.gov/police/cs/crime-stats-archives.htm	Criminal Statistics: No SQL Data School District Data: Relational Data
Demographic Data *Robert Beutner*	Public Data: US Census Data; ESRI Community Analyst https://communityanalyst.arcgis.com/	Relational Data

DATA PROCESSING:
Preparing the data for use in the data warehouse required different methods depending on the category and the source of the data.

Real Estate Data Processing:

MS PowerBI, Excel 2010 and TOAD and were used in the following steps:

1. Filter applied on "State" & "Metro" Level and chose "TX" & "Houston" respectively.
2. Removed, unwanted columns which contained Data from 01/1996 to 12/2006
3. Found the Average for the year and converted the Monthly Level data to Yearly Level Data
4. "PIVOT" Yearly Level Data & Price of the Home. The attached, "HomePrice.csv" contains all the required information
5. Used MySQL as backend DB. Used, TOAD for MySQL tool, and created table with the required columns called "HomePrice"
6. Used Import/Export Wizard on TOAD for MySQL and imported all the data to the HomePrice Table

Quality of Life Data Processing:

The Texas school performance extracts were simplified data to show SAT SCORES the "All Students" groups, editing the attributes, and renaming key attributes in order to harmonize with other sources using the county geography as the main field to join data sets together. These were loaded to the SchoolData table.

Crime statistics for a portion of 2016 were downloaded from the City of Houston into a CSV file. From there, the CDXBingLocator excel add-on was used to determine zip codes for the data to provide better mapping in Tableau. The csv file was batch loaded to the table CrimeData.

To save time, both loads were done on the server using the mysql command prompt:

```
MYSQL --USER=PANNABLE -PASSWORD=XXXXX FINALPROJECT
>LOAD DATA LOCAL INFILE 'CRIMELOAD.CSV' INTO TABLE
CRIMEDATA COLUMNS TERMINATED BY ',' OPTIONALLY
ENCLOSED BY '"';
```

Demographic Data Processing:

The Demographic data extracts from the US Census Data collected during the 2010 decennial Census, as well as the 2015 ACS supplied the contextual demographic information for the Houston, Texas metropolitan area.

Using the ESRI Community Analyst data portal, the Houston Texas metropolitan area was the focused upon at the county scale. See Table 2 for a summary of the Demographic Data.

Table 2: Demographic Data Summary

Census Data Set	Attributes
Owner-Occupied Housing Units (2010)	Number of Housing Units per County (Integer)
Renter-Occupied Housing Units (2010)	Number of Units per County (Integer)
Total Housing Units (2010)	Number of Units per County (Integer)
Population Density (2010)	Population density at County level (decimal)
Per Capita Income (2016)	Per Capita Income by County (decimal)

The datasets downloaded were prepared for MySQL in Excel and following their proper formatting; inserted into their respective corresponding tables in MySQL using the CREATE Table and INESRT into Table commands.

Using MySQL a Data View aggregated the pertinent demographic relational data into one table aligned to their respective counties in the Houston.

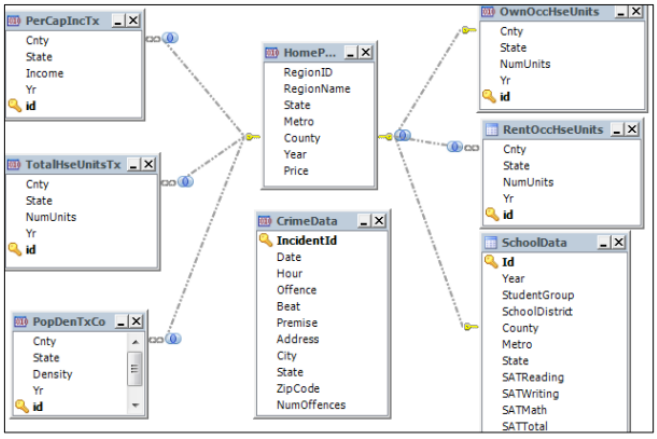
```
CREATE VIEW DEMODATAVIEW AS (
SELECT T.CNTY, T.STATE, O.NUMUNITS AS OWNEDUNITS,
R.NUMUNITS AS RENTEDUNITS, T.NUMUNITS AS TOTALUNITS,
I.INCOME, D.DENSITY

FROM TOTALHSEUNITSTX T, PERCAPINCTX I, POPDENTXCO
D, RENTOCCHSEUNITS R, OWNOCCHSEUNITS O

WHERE T.CNTY=I.CNTY AND D.CNTY = T.CNTY AND
R.CNTY=T.CNTY AND T.CNTY = O.CNTY);
```

SQL Entity-Relationship Design

To support simplified set up in Tableau, a de-normalized data structure was used, where most data tables had repeated meta data. The County name was harmonized and used across all data sets for joins. Crime Data was by address and zip code only, and was used in a stand alone fashion.



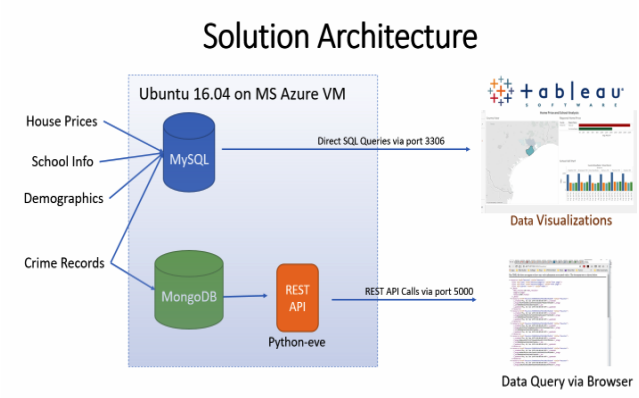
IV. Architecture and Implementation

For our project, we used MySQL running on a Microsoft Azure Virtual Machine running Ubuntu 16.04. A standard installation was done using the apt-get process from https://dev.mysql.com/doc/mysql-apt-repo-quick-guide/en/#apt-repo-fresh-install.

Accounts for each of the team members were setup to allow all to work on various parts of the data solution. Additionally, MongoDB community edition was installed via apt-get, using the procedures at https://docs.mongodb.com/master/tutorial/install-mongodb-on-ubuntu

To support a REST API interface to MongoDB, the Python-eve package was installed using instructions at http://python-eve.org/install.html#install. This package implements REST services on top of MongoDB. More about this part of the setup is later in this paper.

Figure 1 – Solution Architecture



V. Evaluation

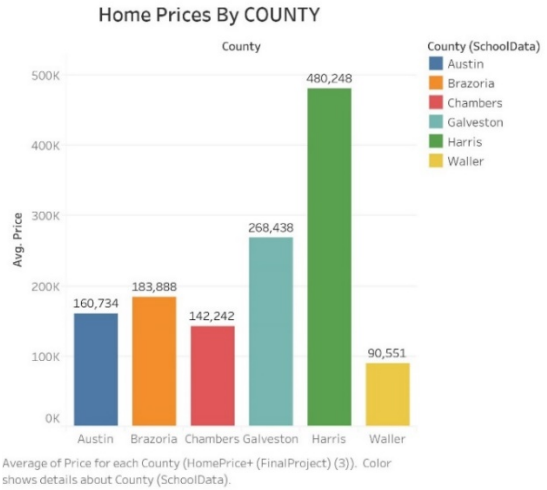
Data Manipulation and visualization using Tableau and Micro-services:

For the purposes of working through the needed queries and related transactions, the following questions were to approximate a user request of the information through query:

1. Which county has the highest average home price? Lowest home price?
2. What is the nature of the relationship between SAT scores and home prices?
3. Does per capita income have any correlation to the house prices?
4. What is the impact of population density on home price? (PopulationDensity with HomePrice comparison)
5. Is the home price on Steady Increase in Houston? Was there a period when the housing price got affected (recession)?

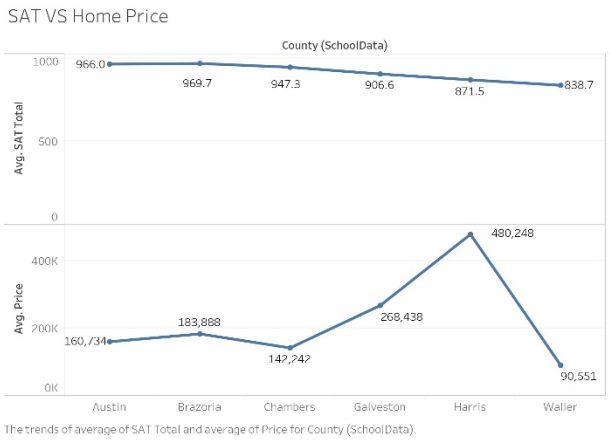
The following are some example data visualizations were the result of posing the above questions to the data warehouse architecture:

Figure 2 – Question 1 - Average Home Price by County and School District



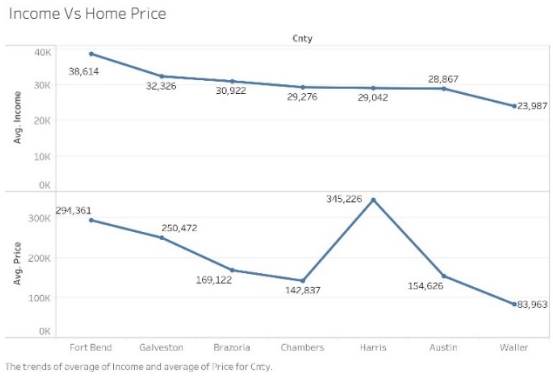
Houses in "Harris" County have the highest average sales price of 480,248. Similarly, the lowest housing prices in "Waller" County at average sales price of 90,551

Figure 3 – Question 2 – Average School District SAT Score and County Average Home Price



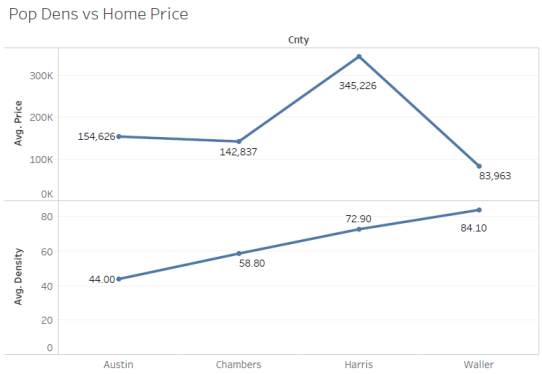
Preliminary data analysis and visualizations provide some evidence that SAT scores and House Prices do not appear to have any significant relationship.

Figure 4 – Question 3 – Per Capita Income and Average Home Price



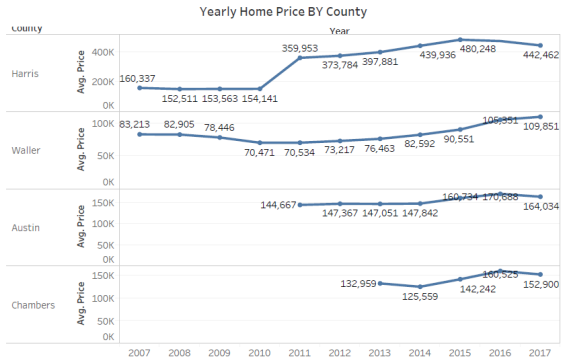
Preliminary data analysis and visualizations provide evidence of a need to work with less disaggregated spatial data, as county level data may be exhibiting a limitation of the data analysis due to spatial auto-correlation or other statistical factors.

Figure 5 – Question 4 – Per Capita Income and Average Home Price



Preliminary data visualizations suggest that there are possible trends in the Houston residential housing market that run counter to the expectation of housing price and population density.

Figure 6 – Question 5 – Home price over time in Houston by county



Preliminary data visualizations suggest that there are definitive trends in housing prices in the residential housing market with an impressive 276% increase in housing price over the decade for Harris County.

VI. Further Software Development Potential: Micro-service Development and Demonstration

Our original goal was to demonstrate the connection of data from MongoDB to Tableau. Because Tableau does not have a native connector to MongoDB, we built an API interface to return JSON datasets using the Python-eve framework. Unfortunately, Tableau could not consume this data directly without an additional Web Connector add-on and we ran out of time to implement this additional layer. However, we were successful in querying data directly from the browser via REST calls.

MongoDb workflow:

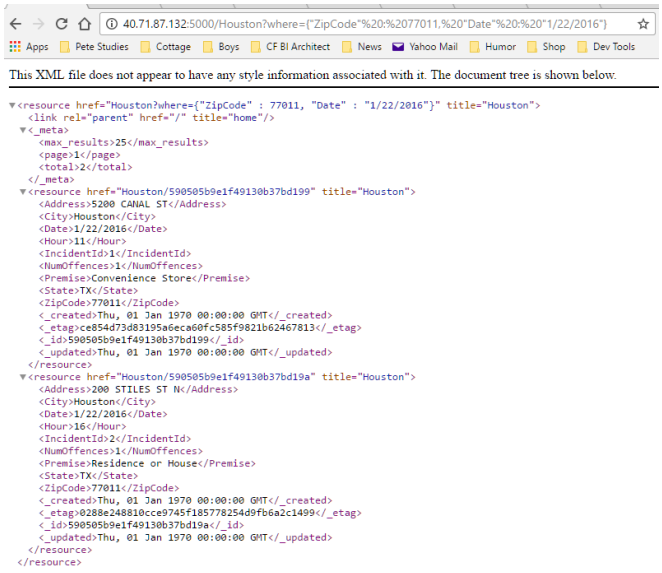
```
> USE CRIME
SWITCHED TO DB CRIME
> DB.CREATECOLLECTION("HOUSTON")
{ "OK" : 1 }
> SHOW COLLECTIONS
HOUSTON
> ANNABLEPJ@DSKB-VM:~/CRIMEDATA/HOUSTON$ MONGOIMPORT --DB CRIME --COLLECTION HOUSTON --TYPE CSV --HEADERLINE --IGNOREBLANKS --FILE CRIMELOAD.CSV
```

To configure Python-Eve, two files are configured, the run.py and settings.py files. These define the application and the schema of the data in MongoDB.

The final result looks like this:
Find all incidents on Jan 22, 2016for ZipCode 77011 :

http://40.71.87.132:5000/Houston?where={%22ZipCode%22%20:%2077011,%20%22Date%22%20:%20%221/22/2016%22}

Figure 7 – Microservice Demonstration



VII. Conclusion

Data acquisition and processing took the most time by far for this exercise. However, once the data were made available to the
MySQL workspace, it proved to be a consistently simple relational data warehouse to set up, maintain and interface with for query development and data visualizations using Tableau. All datasets employed simple data structures, mainly 2nd normal form.

Database transactional performance varied considerably resulting in variable query times when issued from Tableau software. This may have been due to using a very small VM with limited memory.

As a NoSQL technology, MongoDB made it much simpler to load data sets as table structures did not need to be set up in advance; exhibiting strong horizontal scaling. In addition, NoSQL data handling appears to reduce the processing steps between data manipulation and delivery.

As a proof of concept; using various data sets through query and information visualization as well as API-based micro services, we have explored SQL and NO SQL technology workflows to provide end-user focused tools to help answers key questions related to real-estate purchase questions. A tool that can scale as well as provide depth of information for decision-making through the increasing diversity of data sets, types, platforms and structures; has been developed and could provide the basis for further development of decision-making tools.

Acknowledgments

The authors wish to thank Professor Ding, and Yi Bu for their support during this course.

Peter Annable: Cloud resource management, ecosystem design and quality of life extraction, processing and warehousing. MongoDb connection development and data processing.

Rob Beutner: Demographic information extraction, processing and SQL warehousing. Information production and presentation (project narrative production).

Balaji Dhamodharan: Real estate information extraction, processing and SQL warehousing. Tableau dashboard development and visualizations.

References

[1] Installation of MySQL on Ubuntu 16.04 via apt-get : https://dev.mysql.com/doc/mysql-apt-repo-quick-guide/en/#apt-repo-fresh-install

[2] Installation on MongoDB https://docs.mongodb.com/master/tutorial/install-mongodb-on-ubuntu

[3] Installation of Tableau http://www.tableau.com/products

[4] Installation on Python-eve http://python-eve.org/quickstart.html