CIS 520-01 Term Paper: Chicago Food Inspection Analysis

Premdeep Parasa (305094210)

Adheesh Saxena (304746070)

Farhan Mahesania (305069081)

**Abstract**

Food Inspection is an essential aspect of food control system. Public health department ensures quality and safety of food by performing inspections on food facilities such as bakeries, restaurants, grocery stores, etc. Food Inspection is crucial to preserve consumers’ health by implementing requisite food controls in accordance with the requirements of national laws and regulations. It can be achieved by the delivery of rightful information on food facilitators to consumers. An effective analysis on food inspections can safeguard health and monitor wholesome food consumption.

**1. Introduction**

* 1. **Background**

The Food Protection Division of the Chicago Department of Public Health (CDPH) is committed to maintaining the safety of food bought, sold, or prepared for public consumption in Chicago by carrying out science-based inspections of all retail food establishments. Beginning January 4, 2010, the Division began entering inspections into the Digital Health Inspection System in order for the documents to be easily viewed online. Inspections focus on food handling practices, product temperatures, personal hygiene, facility maintenance, and pest control.

**1.2 Overview**

Our analysis on the project is primarily focused on the food violations which can be thwarted by increased inspections to ensure that food facilities comply with the standards set up by CDPH’s Food Protection Program. We used the data set consisting of such inspections from the year 2010 to present. We have analyzed the inspection patterns of all food facilities in Chicago using Hive and MapReduce. We queried the database to check the number of inspections carried by CDPH to determine which facilities which showed high risks in terms of food quality and to check the results of inspections that were carried out.

**2. Technical Specifications**

**System Requirements**

* **Storage:** Standard cloud-based object storage node with a business continuity system in place.
* **File System:** Hadoop Distributed File System(HDFS).
* **MapReduce Platform:** Apache Hive.
* **Analytics and Visualization:** Microsoft Excel enabled with ODBC driver and CloudBerry explorer for Azure Blob.

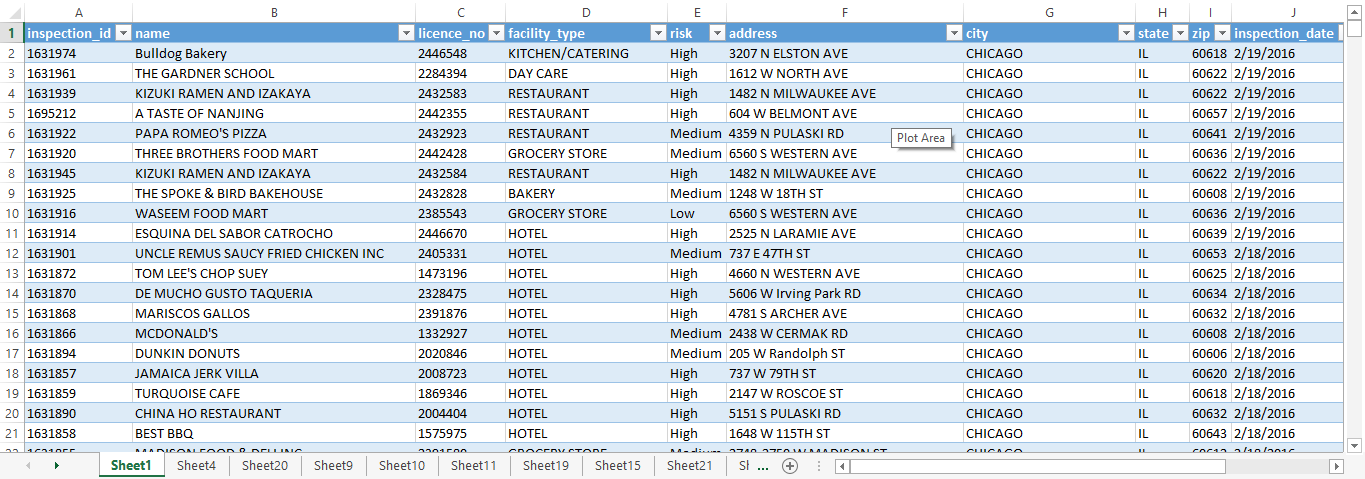
**3. Project Description**

* 1. **Project Scope**

Data analysis is done on the inspections done by Chicago Health Department. We have done analysis on the degree of risks involved among the different establishments. In accordance with that, we continued analyzing the results of those inspections. We will visualize the results of these analysis through Excel using pivot table, maps, pie charts and bar charts.

* 1. **Dataset**

We acquired the dataset from **Chicago Department of Public Health.** The data covers the inspections from January 2010 to February 2016. Format of the data is Comma Separated Value(.csv) and size of dataset is 135 MB. The data primarily consists of facility names, types of risks, inspection details and geographic locations.



* 1. **Storage Deployment**

The storage is deployed using Windows Azure Portal. The configurations are done using a classic Azure storage and storage location is Central US.

* 1. **Hadoop Cluster Deployment**

We have used Apache Azure HDInsight Cluster as Hadoop Distributed File System. It is deployed from the Azure portal, configured to have a Windows operating system and Hadoop version 2.6.0.

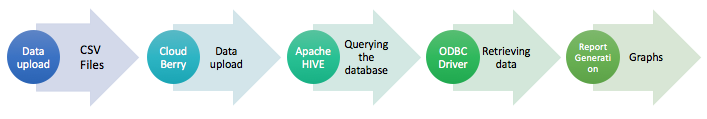
Its resources are comprised of four worker nodes with a total of 16 cores, 14GB RAM and 8 disks.

The head node is set-up with the same specifications.

* 1. **Data Upload**

The dataset is transferred using the CloudBerry Explorer client application. It is uploaded directly into the default container of the blob linked to the HDInsight cluster for easy access. Data uploaded in the system is in Comma Separated Value(CSV) format.

* 1. **Work Flow**

****

Data Uploaded from various data sources (usually CSV files). Cloud Berry is a tool for loading data into the cluster(projectg520). Apache HIVE is used to Querying the data base to get the desired output. ODBC Driver is used to retrieving data from cluster and exporting it to excel. Reports are generated and data can be opened using excel or data can be displayed with graphical representation

* 1. **Querying the database**

The MapReduce queries are executed using Azure’s Hive Editor

**Table Creation and Data Insertion:** *CREATE TABLE* and *LOAD DATA INPATH* scripts are executed to create tables and insert data into the table by uploading data from files with csv format.

**Select Queries:** *SELECT*queries are used to check the data from the table and to get desired output from the given dataset. *SELECT* queries are used to validate data from the dataset. Data is filtered using various aggregate functions like *COUNT, GROUP BY* etc. for getting result based on variety of conditions *WHERE* clause is used in queries.

*SELECT risk, count(Risk) AS Risk\_count FROM food\_inspections GROUP BY risk;*

The above Hive query determines the risk count based on three categories – High, Medium and Low.

*SELECT facility\_type, count(Risk) FROM food\_inspections WHERE risk='High' GROUP BY facility\_type, risk;*

Further we took high risk category to analyze which type of institute accounted for what percent of high risk. The following query illlustrates this scenario.

*SELECT name, year(inspection\_date), results, count(results) FROM food\_inspections WHERE name='Pizza Hut’ GROUP BY name, year;*

Going deep into our analysis, since the restaurant have highest risk among other food facilities, we check the inspection results of all restautants.

*SELECT facility\_type, results, count(results) AS results\_count FROM food\_inspections WHERE facility\_type='RESTAURANT' GROUP BY facility\_type, results;*

During our analysis we had considered a popular food chain of Pizza Hut to check the number of inspections carried out by CDPH from 2010 to 2016. The following query illustrates this scenario.

* 1. **Analytics Tools setup and Integration**

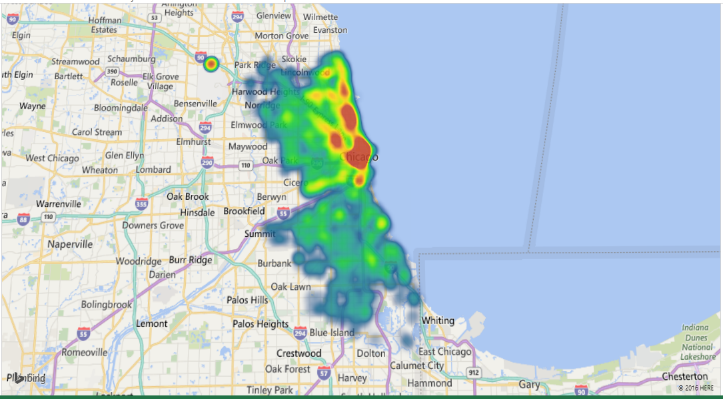
We used MS excel enabled with ODBC driver to transfer the output from apache hive and to create visualization graphs and charts.

* 1. **Testing**

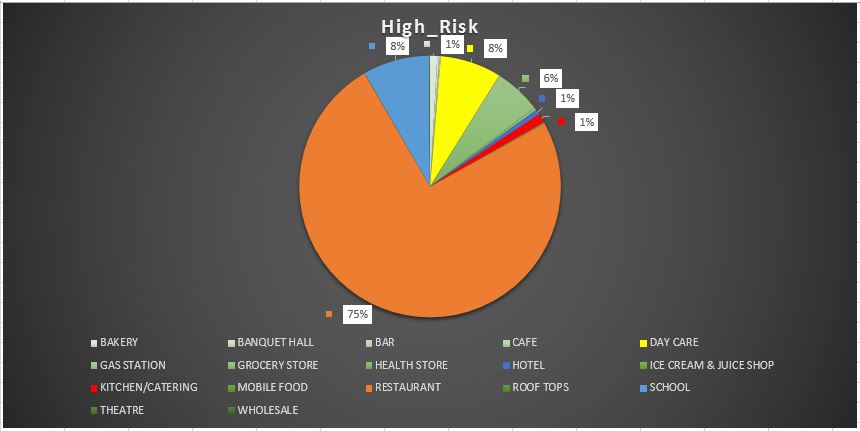
The testing of the end system should conforms to various conditions and criteria set during the business requirement analysis. The testing process is in accordance to the acceptance criteria wherein the data can be properly queried and displayed while maintaining accuracy. All queries should produce expected output based on which resultant graphs and reports are generated.

* 1. **Data Visualization and Interpretation**

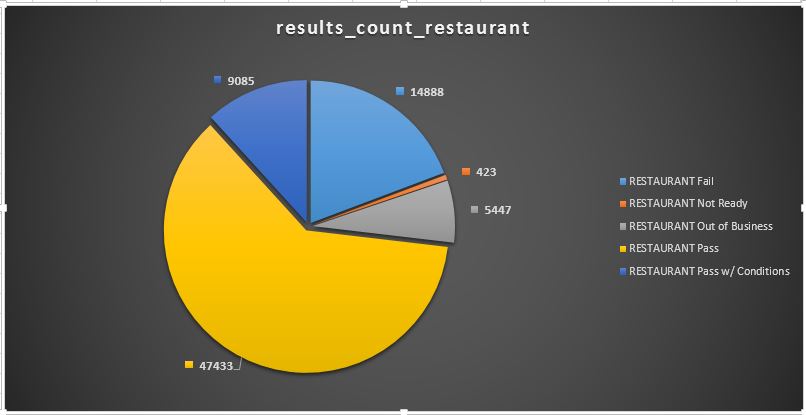
The following graphs were generated after the resultant data of the queries was exported in MS Excel.

The below heat maps was generated using latitude, longitude of each food facility and valued by risk. The red spots shows the areas with highest risk, followed by green ans medium risk and blue as low risk..

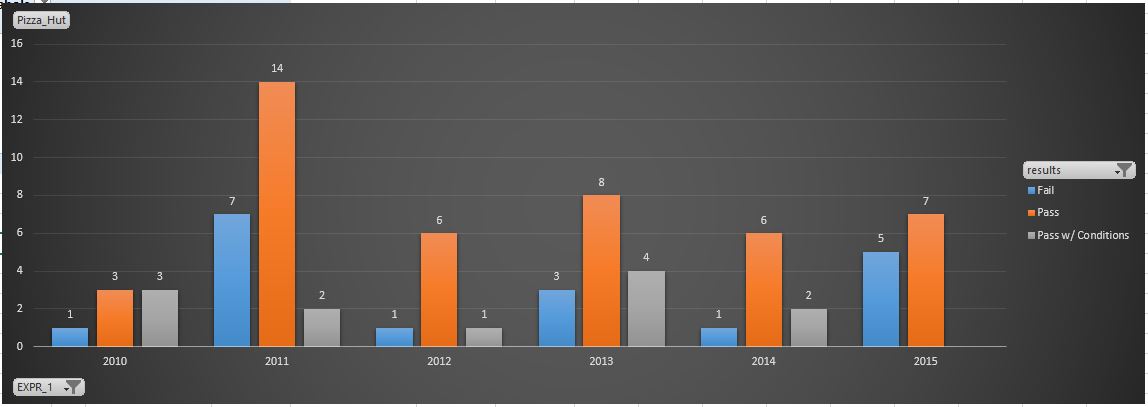
The pie charts below shows the by categorizing the count of facility types which has high risk. We can clearly see that the restaurants have the highest high risk of 75%r followed by 8% high risk for schools and 8% high risk for day care centres.



Since the risk factor for restaurants being high, we have analyzed the inspection results of all the restaurants whether they have passed or failed when the inspection is performed irrespective of the inspecton type. However, from the below pie chart it is understood that though the risk is more in restautants 80.6% of the restaurants have passed during the inspection.



The last query is an example of a popular food chain called Pizza Hut which had to undergo many inspections and the resultant graph is based on various inspection type carried by CDPH at all Pizza Huts in Chicago region from 2010 to 2016. The result of inspection is also categorized in following categries – pass, fail and pass w/conditions during each year.



**4. Conclusion**

From the resultant data we could determine the risk factor of each institution type. The restaurants lead in risk factor followed by schools and day care, whereas hotels, kitchen/catering and bakeries show low risk.

We also got the results of the inspections carried out at restaurants in Chicago region where majority of the restaurants passed inspection carried out by CDPH with passing count of 47433 , failing count of 9084, passed with condition count of 14888 out of total count of 77276 with 80.6% of pass percentage. Surprisingly, the pass percent of inspection after a complaint is 81.2% which proves that food facilities are following high standards and adhering to the regulations set by CDPH. From our analysis on inspection count, it is understood that with the food facilities increasing each year, there is significant growth in inspections each year. We can also conclude that Pizza Hut franchise has failed to fully comply with the safety standards and quality of the food by checking the results of inspections carried out by CDPH from year 2010 to 2016. During 2015, their pass:fail ratio is 60:40 which was 85:15 during the year 2010.

**5. References**

*[1]* ***DATASET URL*** *:* [*https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5/alt*](https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5/alt)

*[2]* ***GITHUB CODE*** *:* [*https://github.com/farhanmahesania/ChicagoFoodInspection.git*](https://github.com/farhanmahesania/ChicagoFoodInspection.git)

*[3]*[*https://azure.microsoft.com/en-us/documentation/articles/hdinsight-connect-excel-hive-odbc-driver*](https://azure.microsoft.com/en-us/documentation/articles/hdinsight-connect-excel-hive-odbc-driver)

*[4]*[*http://doh.dc.gov/service/understanding-food-establishment-inspections*](http://doh.dc.gov/service/understanding-food-establishment-inspections)

*[5]*[*http://www.cityofchicago.org/city/en/depts/cdph/provdrs/environ\_health/svcs/restaurant\_food\_inspection.html*](http://www.cityofchicago.org/city/en/depts/cdph/provdrs/environ_health/svcs/restaurant_food_inspection.html)

*[6]* [*http://www.w3schools.com/sql/default.asp*](http://www.w3schools.com/sql/default.asp)

*[7]* [*https://azure.microsoft.com/en-us/documentation/articles/hdinsight-connect-excel-hive-odbc-driver/*](https://azure.microsoft.com/en-us/documentation/articles/hdinsight-connect-excel-hive-odbc-driver/)