Architect, Populate and Explore a Data Warehouse for Business Intelligence

National College of Ireland

Msc in Data Analytics

CA 2 Paper

21th April 2015

|  |  |
| --- | --- |
| Author(s) | Sapna Gupta |
| Degree | Msc in Data Analytics |
| Department | School of computing |
| Instructor | Dr Simon Caton |

# Introduction

ABC LTD is a famous UK wholesale company. This company sells goods to different Cash and Carry store all over UK. ABC LTD is interested in building new data warehouse for their company, to improve it’s corporate planning and report mechanism and address issues of competitive intelligence.

Two most commonly used approaches for building data warehouse introduced by [Bill Inmon](http://inmoncif.com/about/) and [Ralph Kimball](http://www.kimballgroup.com/html/about.html) are:

Bill Inmon’s enterprise data warehouse approach (the top-down design): A normalized data model is designed first. Then the dimensional data marts, which contain data required for specific business processes or specific departments are created from the data warehouse.

Ralph Kimball’s dimensional design approach (the bottom-up design): The data marts facilitating reports and analysis are created first; these are then combined together to create a broad data warehouse.

In my data warehouse implementation project I am following Kimball approach, as this is easy to implement , low cost , faster to build and easily scalable. This document summarises all the steps taken to implement data warehouse which gives 360 degree point of view of ABC business to it’s stakeholders, including the ETL process and Business intelligence reports drown from this data warehouse.

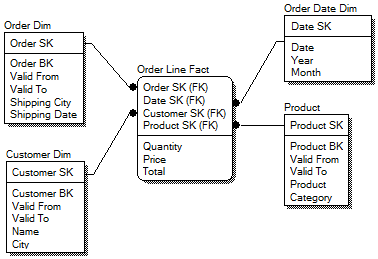
# Implementation and architecture

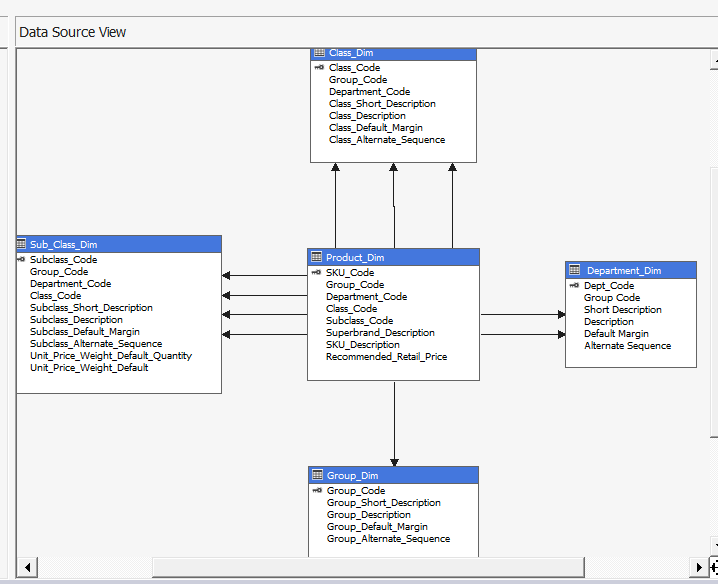
**Data Warehouse Architecture**

There are various data warehouse methodologies. The most widely utilised methodologies are modelling with confirmed dimensions and an enterprise data bus, or methodology with the database normalized. As the data is growing rapidly and organisations look for the business value that is concealed within non-structured data, they encounter the challenge of how to analyse complex data. Conventionally, data warehouses analyse structured, transactional data that is contained within relational databases. Data cleansing for confirmed dimensions will lead to loosing information.

Dimensional Model

In our dimensional model, we have a fact table in the middle and dimension tables around the fact table. We added order date as a separate dimension. Customer is directly joined to the fact table.





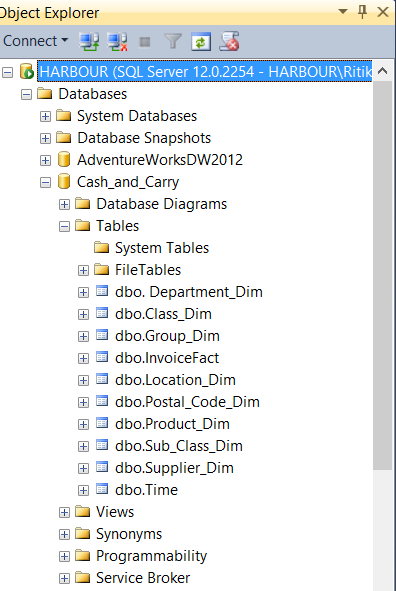
Data Base used:

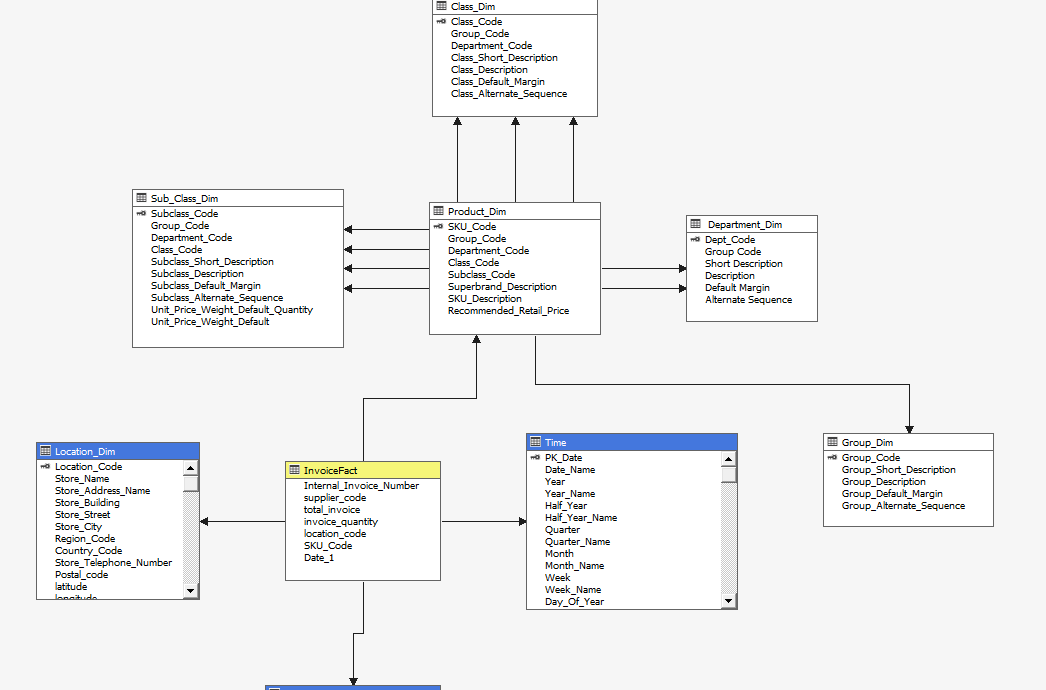
1. Mysql Server
2. SQL server
3. Hadoop HDFS ( NoSQL database )

Data Sources:

1. Structured Data – ERP system of ABC LTD ( Products , Invoice , Supplier , Store)
2. Unstructured Data – Company blogs stored in XML files with text column as unstructured data, Customer reviews in text document.
3. SemiStructured Data – Downloaded Longitude and Latitude from Web for Supplier and store locations using Power Query tool.

Looking at the complexity of the data set snowflex schema was created.





Data Cleaning

In order to answer Business queries correctly we need to have clean data and only required data to speed up business queries. Steps involved are.

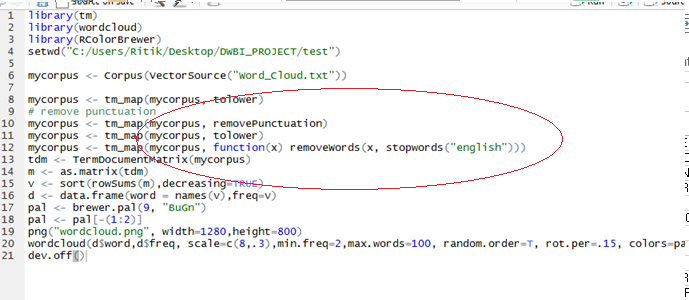
1. Business was only interested to analyse data of their 3 types of products. Bear , Wine and Tabcco.

So only information related to these 3 types of product was used to create Data Mart CashCarry.

1. Postal Code of Stores and Suppliers were missing. This information was populated manually using information provided in other address column.



Depending on the business requirement we need to use different techniques for data cleaning, like to analyse customer’s reviews I used text mining tools using R programming.



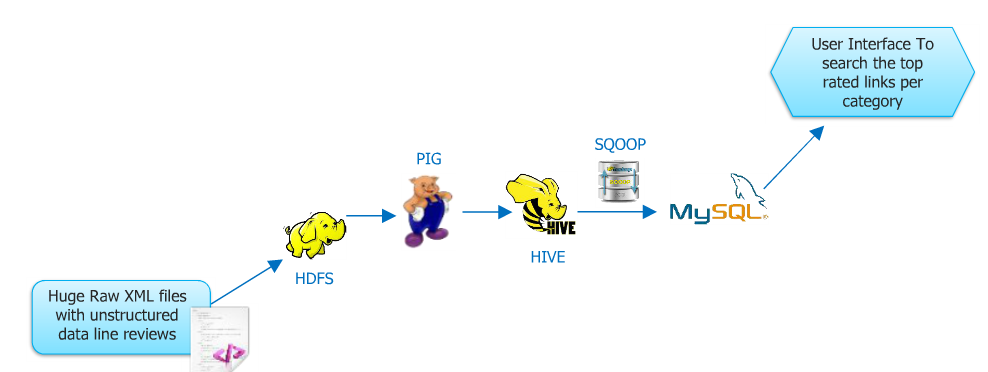
# ETL

ETL tools used are

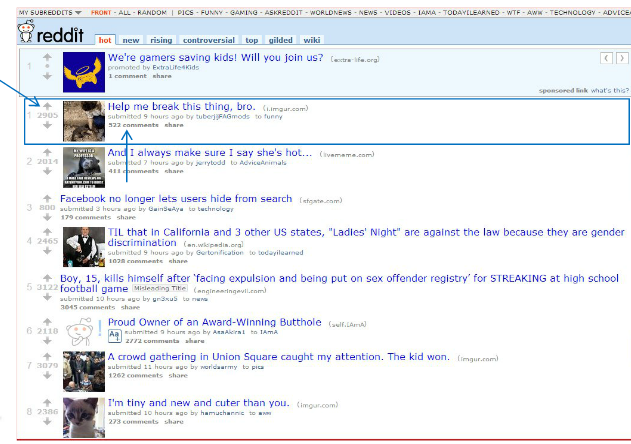
* Power Query
* SQL Server Integration Services (SSIS)
* SQL Server Analytics Services (SSAS)
* R programming
* Map Reduce
* Pig ,
* Scoop

Company’s customer reviews from company website were stored in XML file, which included unstructured reviews in one of the column. Modern ETL technologies were used for analysing unstructured data.

Data was stored in HDFS files, cleaned and categorised using Map Reduce. Positive and Negative Sentimental analysis was performed using Map Reduce. Data was restructured using PIG. Data was aggregated using HIVE queries. Data was moved from HDFS to Mysql server using SQOOP. This data is used in Business Intelligence report to search top rated links per category.



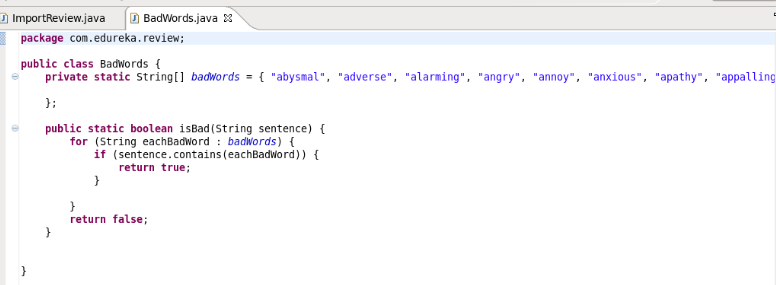
WebSite Reviews :

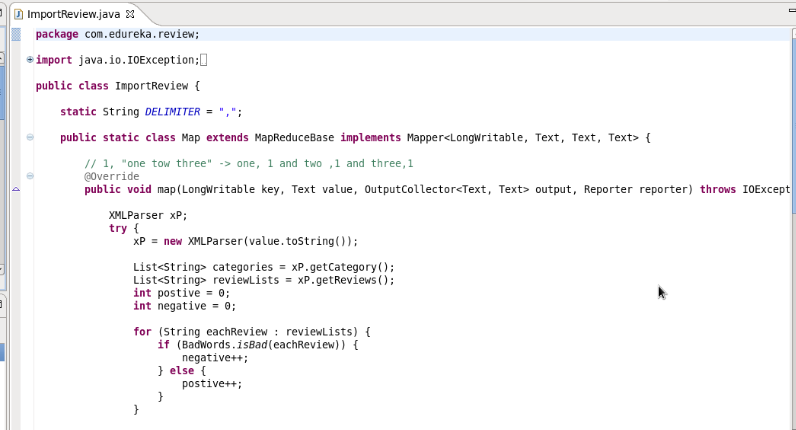


Sample of the reviews XML file.

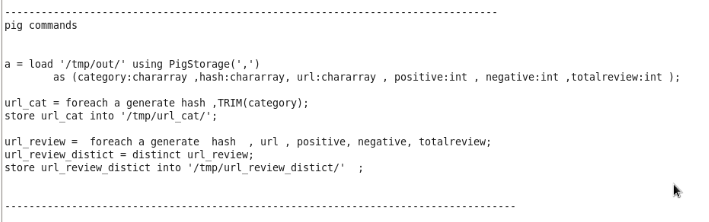
--

Following is the code used in Map Reduce to do sentiment analysis.

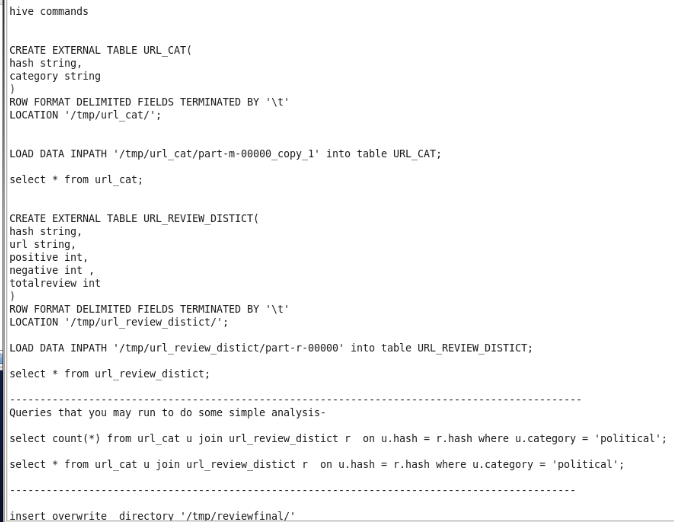




PIG Commads to load data into /tmp/out/ folder and giving schema to the unstructured data



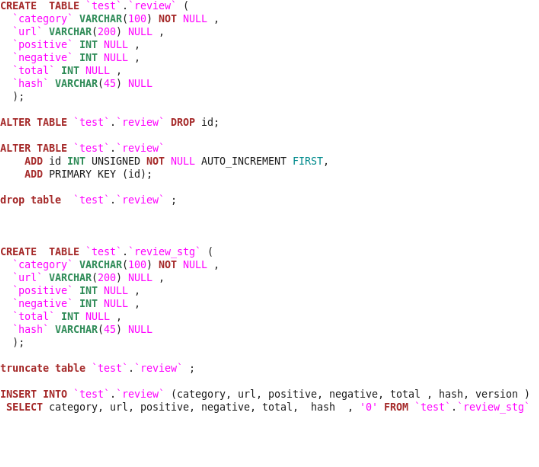
Following are HIVE commands to load the data into tables and analyse the data residing in the tables.



After analysing the results were moved into mysql server using following SQOOP command.

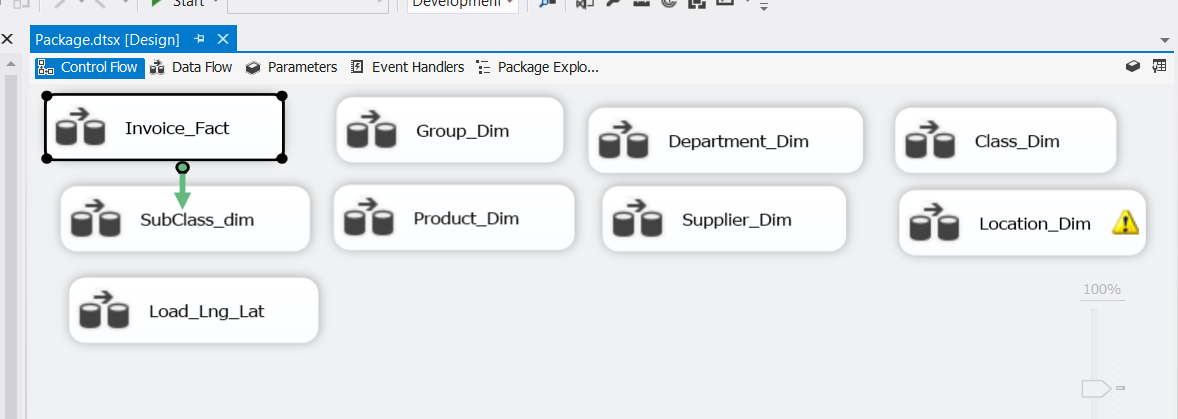
sqoop export --connect jdbc:mysql://localhost:3306/test --table review\_stg --username root --export-dir '/tmp/reviewfinal' --input-fields-terminated-by '\001'

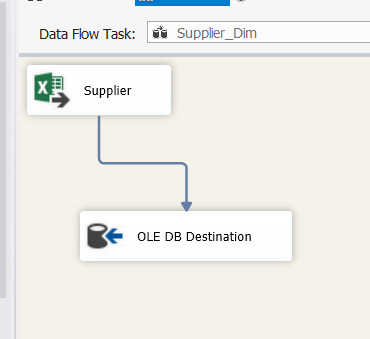
Following tables were created in Mysql before loading data using above SQOOP command.

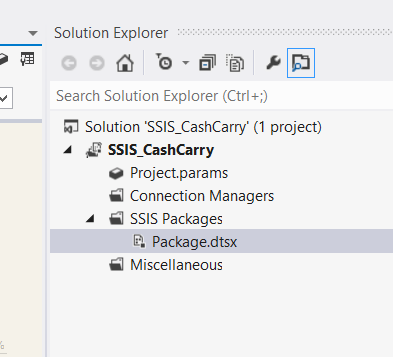


Data is moved from review\_stg table into review table using above table and each record is also versioned in Mysql server.

SQL Server Integration Services (SSIS) used to load data from spreadsheets to SQL server tables.

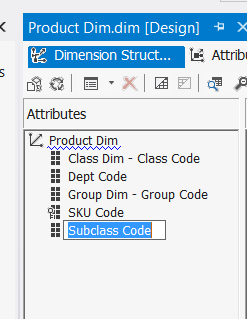


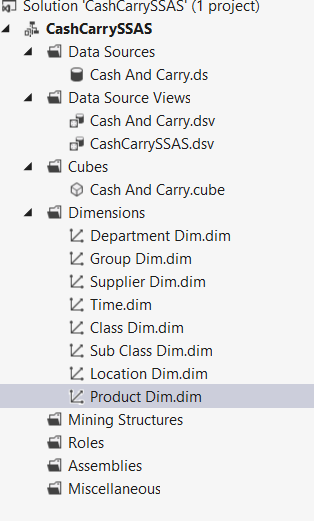




Data Analysis Process using SSAS

SSAS, is an online analytical processing (OLAP), data mining and reporting tool in Microsoft SQL Server. SSAS is used as a tool by organizations to analyse and make sense of information possibly spread out across multiple databases, or in disparate tables.





In this process we create Dimensions and Facts to create an OLAP cube for analytical purpose. Firstly the data source is added and created a data view. The Tables created and populated using SSIS has to be linked using a relational identifier or a Primary Key.

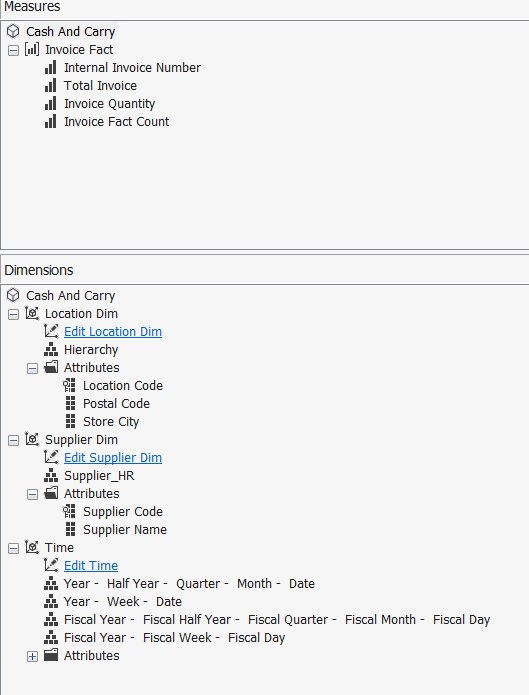
The concept of this Data warehouse is Snowflake schema. In this schema each point of the star explodes into more points. In a star schema, each dimension is represented by a single dimensional table, whereas in a snowflake schema, that dimensional table is normalized into multiple lookup tables, each representing a level in the dimensional hierarchy.

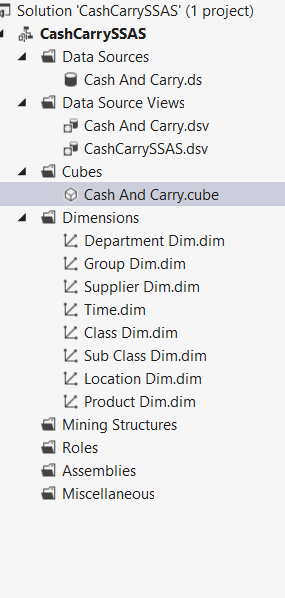
Creating Dimensions and Hierarchy

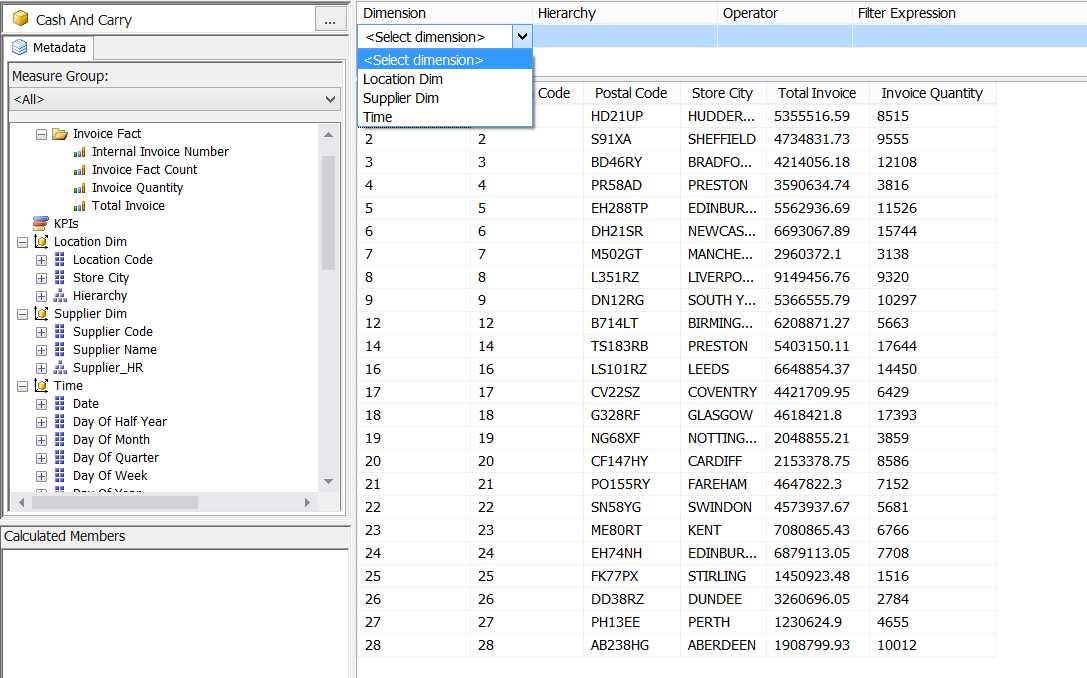
Dimensions are a fundamental component of cubes. Dimensions organize data with relation to an area of interest, such as customers, stores, or to users. Dimensions in Analysis Services contain attributes that correspond to columns in dimension tables. These attributes appear as attribute hierarchies and can be organized into user-defined hierarchies, or can be defined as parent-child hierarchies based on columns in the underlying dimension table.

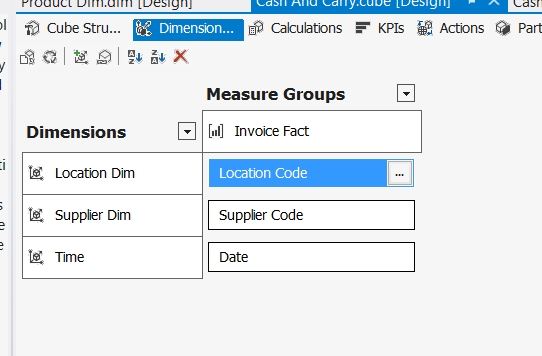
OLAP Cube Structure

OLAP Cube is created and Time dimension is to be added and configure with the Sales Date in Transaction table before processing the Cube. Once the cube is successfully processed the data can be sliced and diced according to Measures and Dimensions in the browser pane.









# Case Studies

1. Sales for each quarter of the year for each group of items.
2. Categorising or clustering suppliers based on their income.
3. Location based report on the UK Map.
4. Sentiment analysis using Word Cloud.
5. Finding Customers review on the brands of the 3 product ( Wine , Beer , Tobacco)
6. Relation between unit price and sales using Correlation.

To complete these case studies tools used are:

1. SSRS

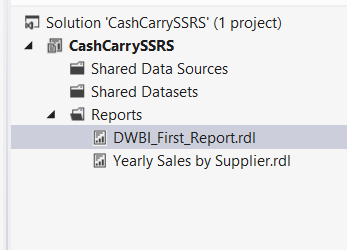
2. Power BI Suit ( Power Query , Power Pivot , Power Map )

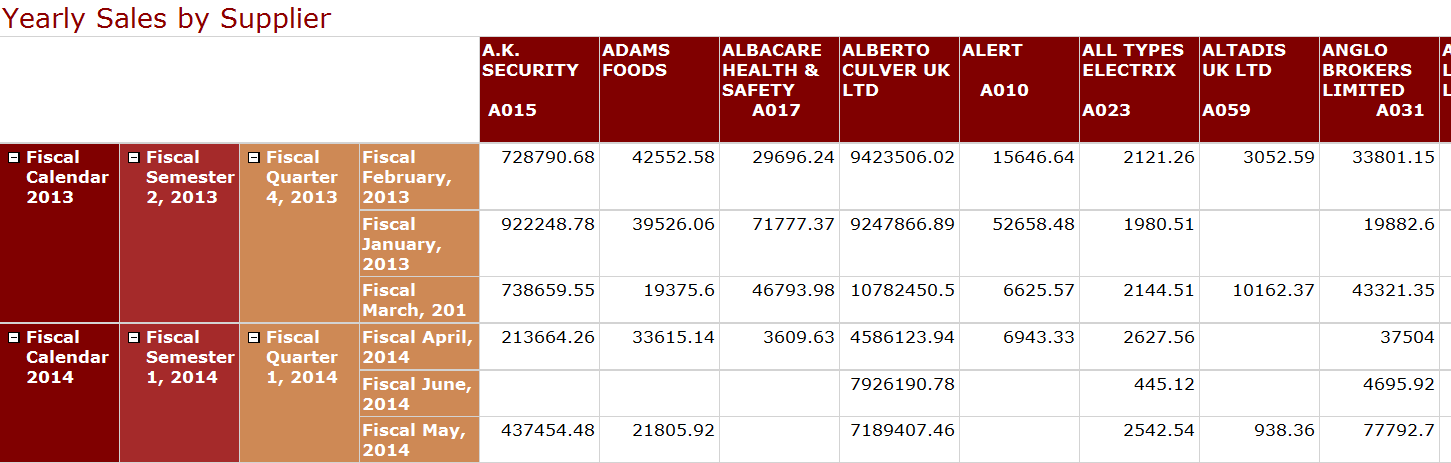
4. R Programming

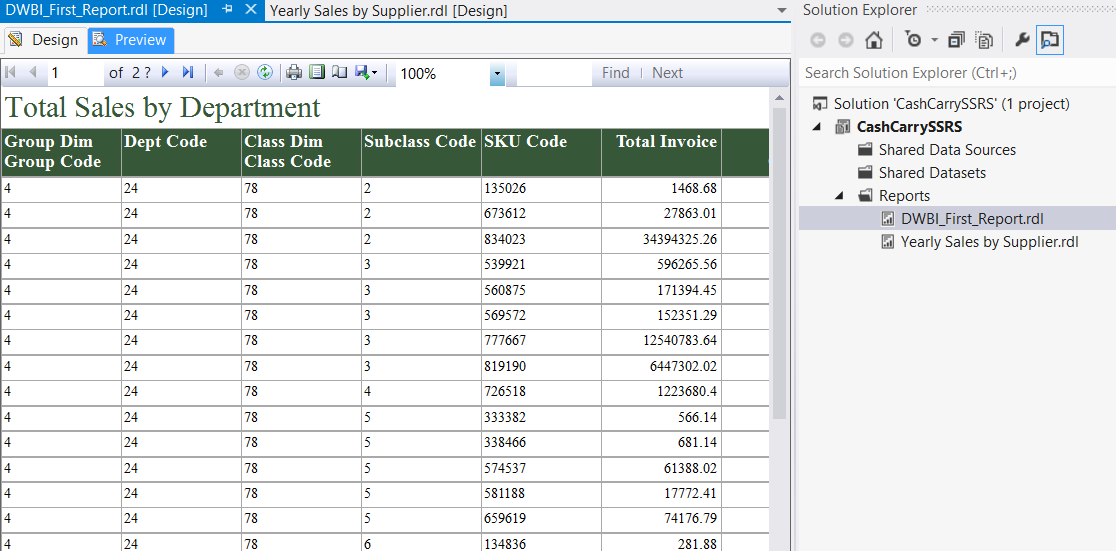
5. Hive QL

Creating Reports using SSRS

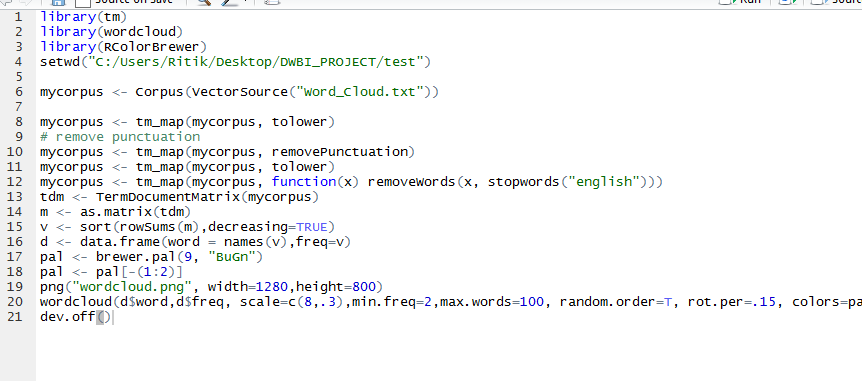
The prime requirement of creating Data Warehouse is to analyse and present in data in good shape and structure. SSRS is a reporting tool gathering data from the Analysis Services and generate reports using complex SQL Queries. The query builder enables less technical IT workers to create Reports according to Business needs.

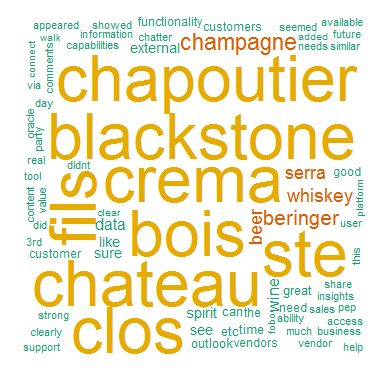
Firstly the Report Wizard enables to connect to Analytical Services and Data warehouse database system. Then using the Query builder the Report query can be created with measures and dimension fields available in the GUI. The sequence of process is shown below.





**Sentiment analysis using Word Cloud using R programming to find most popular brands of wine , whiskey among customers.**





**hive>** select distinct category, positive ,negative , total ,url from review\_stg where upper(category) in ('WINE','WHISKEY','CHAMPAGNE','BEER','SPIRIT');

+-----------+----------+----------+-------+----------------------------------+

| category | positive | negative | total | url |

+-----------+----------+----------+-------+----------------------------------+

| wine | 771 | 75 | 846 | http://www.codeweavers.com/ |

| Beer | 6106 | 76 | 6182 | http://www.1001freefonts.com/ |

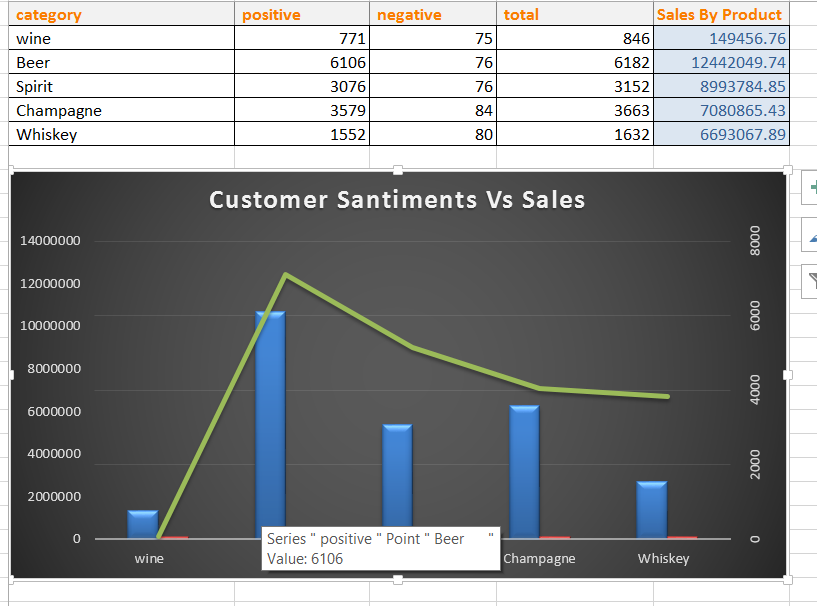
| Spirit | 3076 | 76 | 3152 | http://www.charitynavigator.org/ |

| Champagne | 3579 | 84 | 3663 | http://newounds.com/ |

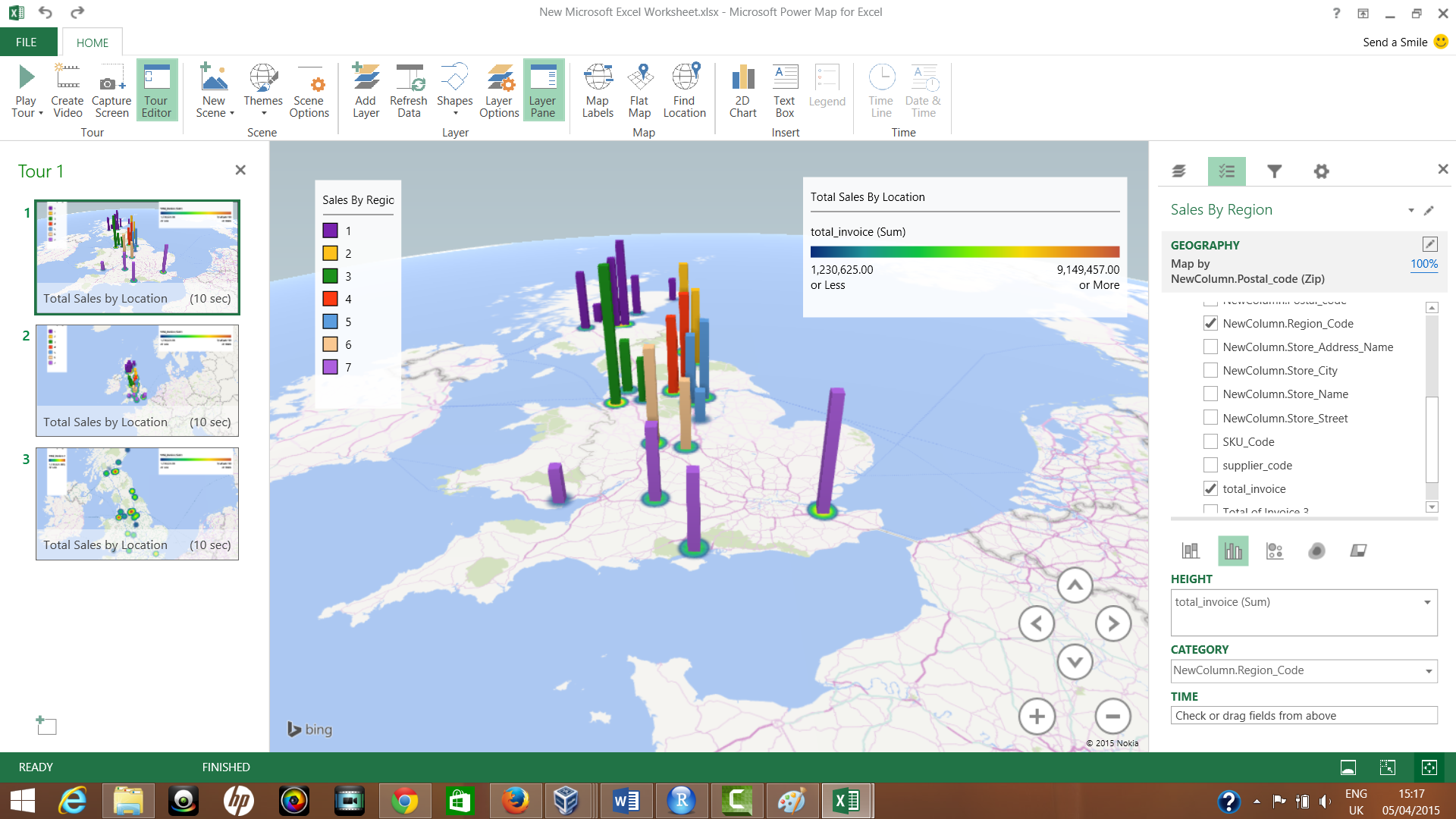
| Whiskey | 1552 | 80 | 1632 | http://www.google.com/chrome |

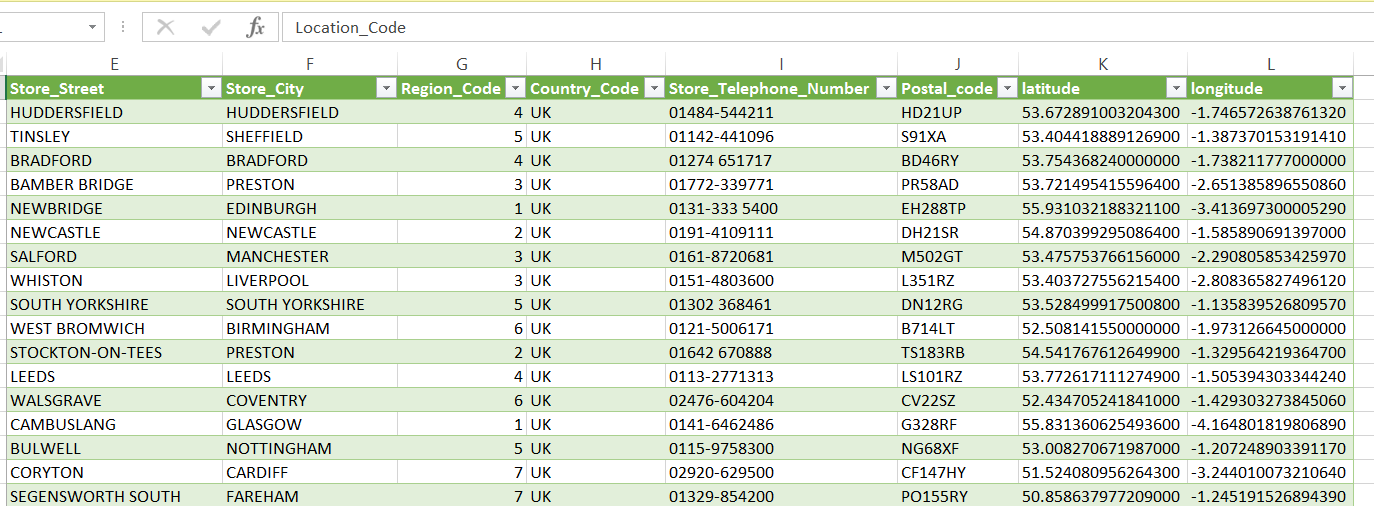
+-----------+----------+----------+-------+----------------------------------+

# rows in set (0.00 sec)

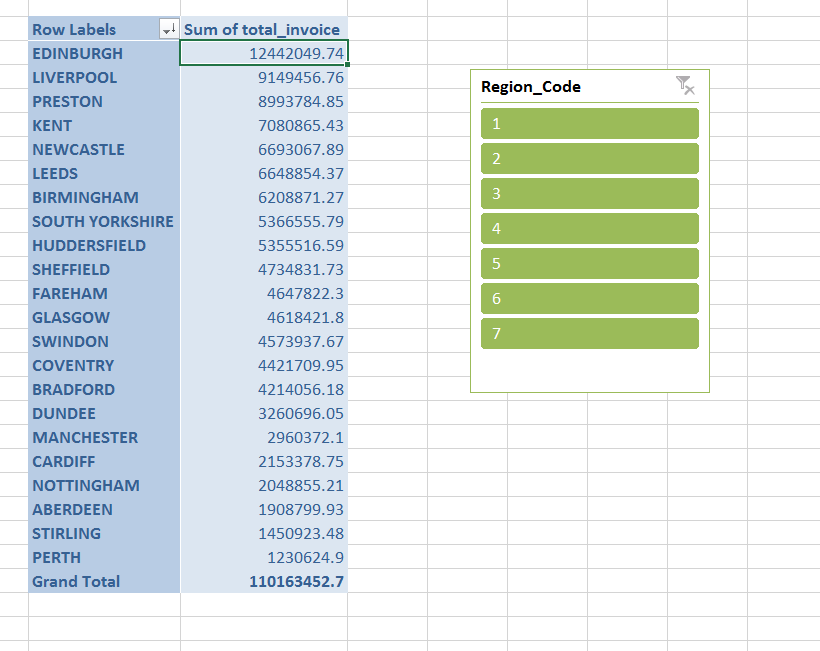


**Location based report on the UK Map using power BI tool of Microsoft.**





**Invoice Total by Location**



## References

1. Data Warehouse Architecture [Online].

URL:http://www.1keydata.com/datawarehousing/data-warehouse-architecture.html

1. Inmon, William H. and Hackathorn, Richard D. Using the Data Warehouse. NewYork, NY: John Wiley & Sons,
2. <http://en.wikipedia.org/wiki/Extract,_transform,_load>
3. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data
4. <http://hortonworks.com/hadoop-tutorial/hello-world-an-introduction-to-hadoop-hcatalog-hive-and-pig/>
5. <http://www.edureka.co/?utm_source=Google&utm_campaign=Brand-Search-Other&utm_medium=cpc&utm_term=Edureka&network=g&utm_content=40946180054&siteTarget=&campaign_id=00&adgroup_id=03&af_id=75fe73cf&gclid=CIGy1rbl38QCFSGL2wodAHUATQ>
6. <https://georeferenced.wordpress.com/2013/01/15/rwordcloud/>
7. <http://searchbusinessintelligence.techtarget.in/tip/Inmon-vs-Kimball-Which-approach-is-suitable-for-your-data-warehouse>