FOSS based Interactive Spatial Temporal Analytical Tool: a GeoBI case study of retail data

Neha Pande and Dr. K S Rajan
Lab for Spatial Informatics, IIIT Hyderabad, India
(neha.pande@research.,rajan)@iiit.ac.in

KEYWORDS: GeoBI, Interactive data visualization, Open source tools, Spatial OLAP

ABSTRACT: Online WebGIS tools and services have become widely available in the recent past to visualize and query geospatial data. These are primarily spatial visualization of aggregated attribute data across various spatial scales or geo-hierarchy levels. Most of these tools lack the temporal profile of the data. Also, these are server centric. In the era of big data, the user data has various attributes collected over time across geographic detail. Also user may like to do visual analysis across these attributes at various levels of aggregation. Such server centric approaches give limited maneuverability to explore the data and discover hidden patterns and insights in data. This paper presents a Spatial-OLAP tool that has been developed over existing Open Source geospatial and database tools. It computes data aggregation across user-defined geo hierarchy with dynamic visualizations in a defined framework. Parameters in menu options are not pre-computed but added on the fly based on user-given data. The tool enriches user experience by combining additional data-aggregation functions over spatial toolkits like openlayers and allowing user to zoom to a set of temporal charts for a more extensive analysis of data across space and time.

To show the utility of this tool, a GeoBI case study was done on a large online retail data of three products sold in different countries during January 2009. The tool helped quickly identify the top five countries with maximum sales and the payment options in decreasing order of popularity. Also, relationships between quantity of product sold across the geographic regions was also extracted. As Online retailers need to draw insights from data to understand their customers across regions, we hope this tool can benefit them in finding such answers - what their popular products are, identifying regions where there is a demand for these, understanding regional preferences and many others.

1. INTRODUCTION

Today, a large number of online webgis tools and services are available for geospatial data visualization and exploration. However, very few of them allow users to perform a time-series or a trend analysis. Server-centric approaches are used in most of these tools. Therefore, data exploration using these tools can’t be done easily as they have limited mobility. Moreover, many of them have not incorporated OLAP techniques into the tools. With an OLAP tool, data can be analyzed from multiple perspectives and aggregation levels [7]. The developed tool addresses the challenges faced by these tools.

With advancements in technology, a large amount of data is being generated such as meteorological data, geographical data, medicine, and data from different kinds of sensors placed in buildings, roads, cars, cell phones, and so on. Location-aware data is also easily collected using smartphones and other handheld devices due to the presence of global positioning systems (GPS). Out of this, approximately 80% of digitally generated data contains a spatial element [1]. Many of them include a timestamp. With increasing amount of data, there is a need for a tool which helps the user to quickly analyze such vast amount of information and generate hypothesis.

This work demonstrates a web-based Spatial-OLAP tool which produces a map with dynamically generated charts from user-given data. It computes data aggregation across user-defined geo hierarchy with dynamic visualizations in a defined framework. Server-side approach is used by the tool for computing data aggregations and generating charts which are stored in chart cache. The tool provides the user with many chart configuration options such as selecting the one or many data attributes, aggregation function and the type for chart. It has been completely developed using open source geospatial and database tools.

The paper is structured as follows. Section 2 introduces related work, Section 3 describes the interactive spatio-temporal analytical tool, Section 4 demonstrates the case study undertaken and Section 5 concludes the paper.

2. RELATED WORK

There exists two types of geo-visualization frameworks: a thick client-server model based and thin client-server model. The first type requires the end-user to install the client program which limits its accessibility as you have to persuade the users to install the client program. Moreover, many versions of client programs have to be developed for different platforms. Google Earth and Microsoft Bing are examples of this.

Thin client-server based web visualization tools are accessible through a web browser to anyone anywhere who would have access to the internet. Mao and Ban (2011) employed X3DOM to visualize spatio-temporal data in 3D. Xia Li and Menno-Jan Kraak (2013) explored the multivariate spatio-temporal data with the time-tide. Natalia Andrienko, Gennady Andrienko (2013) analyzed movement data through visual analytics toolkit. Anil Ramakrishna, Yu-Han Chang, and Rajiv Maheswaran (2015) developed an Interactive Web Based Spatio-Temporal Visualization System using D3.js, QGIS, Django. Bo Mao, Zhiang Wu, Jie Cao (2012) proposed a framework for online spatio-temporal data visualization based on HTML5. Alfred David, Clarence J M Tauro (2015) worked on developing Web 3D Data Visualization of Spatio-Temporal Data using Data Driven Document (D3js). However, very few of these change the data content on map dynamically as the user zooms in, zooms out or changes geographic extent. Also, most of the frameworks use data previously stored in the database for visualization and don’t generate visualizations on the fly.

3. INTERACTIVE SPATIO-TEMPORAL ANALYTICAL TOOL

3.1 System Goals

The following system goals were established:

- Allow user to choose aggregation function such as min, max, sum.
- User can select one or more data attributes for aggregation
- User can choose chart type
- On-the-fly chart generation from user-given data
- Spatial aggregation of data
- Display different boundary shapefiles with change in zoom-level
- Aggregate and disaggregate data with change in zoom-level

3.2 Main Assumptions

Data input should contain files in csv (for data file) and zip (for boundary shapefiles) formats. For processing polygons of the boundary files, both .shp and .shx files would be required. Hence, the upload format for boundary files was decided as zip. Latitude and Longitude values are assumed to be present in the given csv data file. Hierarchy among the boundary shapefiles should be specified by the user while uploading them (from highest to lowest hierarchy).

3.3 Technology

The developed tool is based on a suite of technologies, as listed here-

- Openlayers 2 (web mapping service)
- Apache Tomcat 7 (hosting server)
- Phantomjs export server (chart exporting server)
- Java/PHP Bridge
- HTML, CSS, AJAX, JQuery, javascript (client-side web technologies)
- Highcharts javascript library
- Php, ogr, shapely (for server-side scripting)

3.4 Development Environment

The development platform for the proposed framework is a PC with Intel Core 2.5GHz x4 CPU, 3.7GB RAM, and Ubuntu 14.04. The web page is deployed in Tomcat 7.0 server. The web browser is Mozilla Firefox. The Web Socket is deployed in port 8080. All tests and experiments are carried out in the local host.

3.5 Overall architecture and system components

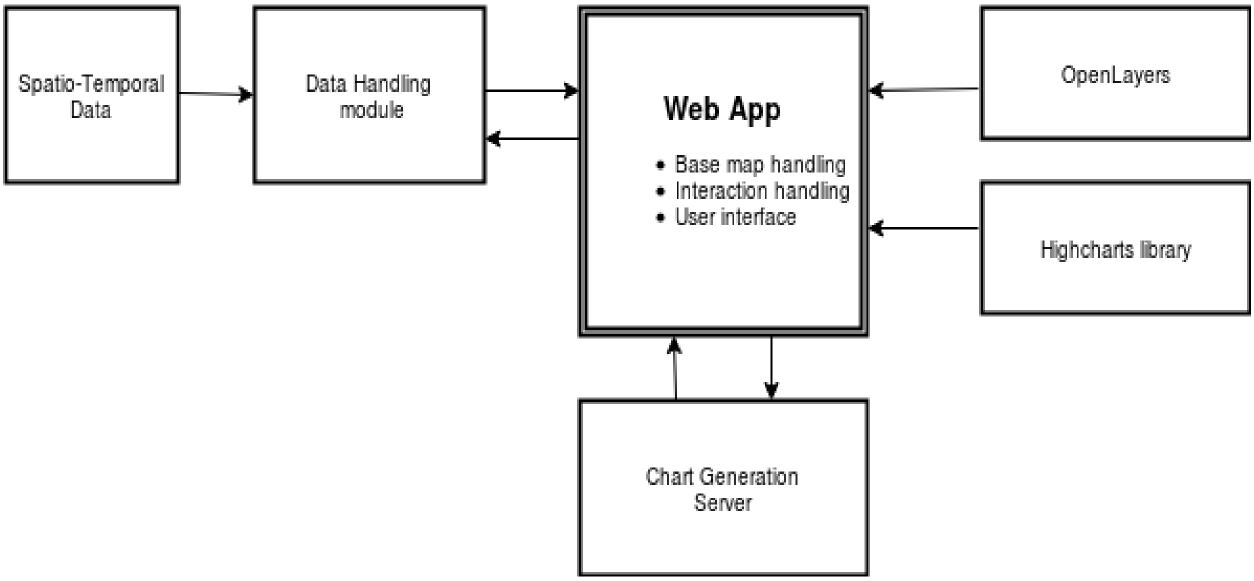


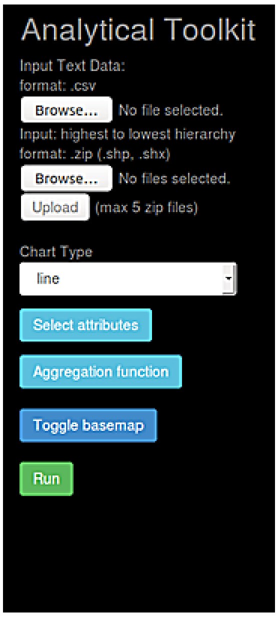
Figure 1. Overall architecture diagram

This tool uses a server-side approach for generating charts on the fly and displaying them on a map. In server-side approaches, the client asks the questions and the server responds to it in a defined framework. The responses are not auto-generated.

In current server based architectures, most of the data is preprocessed and it is displayed as per user request. In this tool, the parameters displayed in the menu are not preprocessed but computed on the fly. Moreover, the visualizations are generated by the tool dynamically based on user-given data.

Figure 1 shows the overall architecture of the system. Data handling module is responsible for storing and processing user uploaded data. The web application is hosted on apache tomcat 7. Openlayers library is used to display map on the webpage. Charts are configured using the highcharts javascript charting library and before adding to the web application. Phantomjs is the chart generation server. It exports charts on receiving chart generation requests from the web application.

* The chart labels can be seen on zooming in the map using *zoom to a set of charts* feature of the tool



Analytical Toolkit

Input Text Data:
format: .csv
Browse... No file selected.

Input: highest to lowest hierarchy
format: .zip (.shp, .shx)
Browse... No files selected.
Upload (max 5 zip files)

Chart Type
line

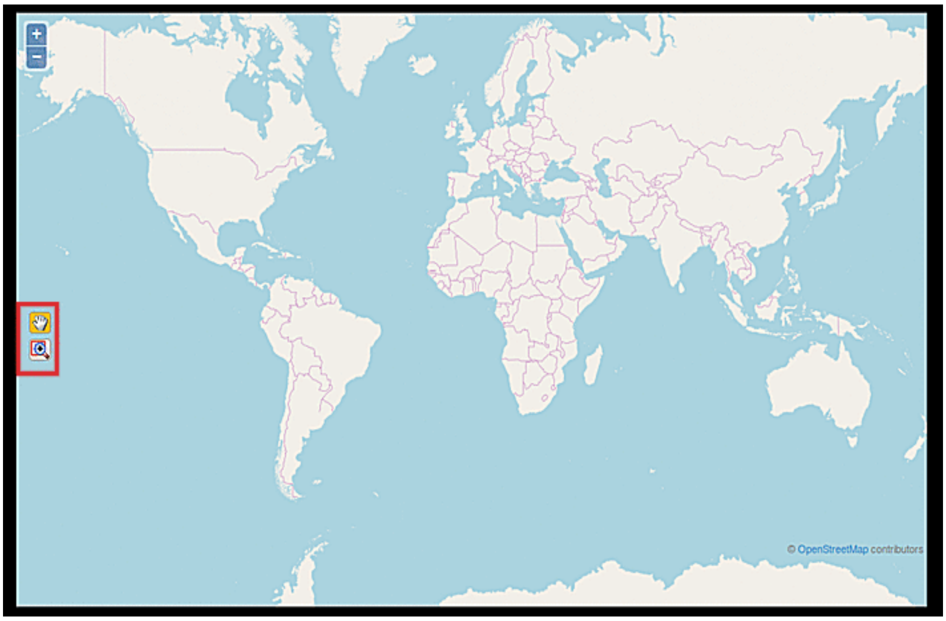


Figure 2a. Side Panel **Figure 2b. Map Panel**

The front-end of the web application comprises of a map panel for displaying the charts on the map and a side panel which allows user to upload files and choose different chart configuration options.

3.5.1 Backend Processing

It involves data extraction, preprocessing and generation of charts for displaying on map. Apache Tomcat 7 server has been used for hosting the application and phantomjs server for generating and exporting charts. Various components involved are listed as follows:

3.5.1.1 Spatial aggregation: This refers to the aggregation of data values which lie in the same geographical boundary. Following steps are taken for computing spatial aggregation. Firstly, the polygon boundaries are obtained from the user given boundary shapefiles. Next, point in polygon queries are run for all data points and a mapping is created between data points and polygons. This is done using shapely and ogr python libraries. This mapping is the then used to identify the data record ids which are to be aggregated. For chart generation, data from these record ids is aggregated and sent to phantomjs (chart export server).

3.5.1.2 Geographical hierarchy: This refers to the ordering between the different geographical boundary shapefiles. It enables the user to see data values at different geographic hierarchy levels. Example of geographical hierarchy is taluk level, followed by district, state and country levels. The ascending order of the user uploaded the boundary files is set as the hierarchy order from the highest to lowest hierarchy i.e. the first uploaded boundary file would be the highest in the hierarchy.

These boundary files are then mapped to specific zoom levels for display. At higher zoom levels, boundary files lower in the hierarchy will be loaded and vice-versa. As the user zooms in or zooms out of the map, new boundary files are loaded at specific zoom levels.

3.5.1.3 Chart Box Computation: Chart boxes are the bounds of the charts being displayed at different zoom levels. Their dimensions are decided based on bounds of the polygons in which they lie and the screen resolution. This is done primarily to fit the charts inside the polygons in which lie. This allows for much better ascetics and a clear visualization of data.

3.5.1.4 Zoom Level based Computation: The tool uses openlayers map as a background layer. Openlayers has zoom levels from 2 to 16. For the user to be able to explore the data interactively, data displayed on the map is changed based on zoom level. As the zoom level changes, the number of charts displayed, size of charts, data values and boundary files (geojson layer) also change. All these parameters are recomputed at the backend each time. In this way, more effective data visualizations are created which are easy to understand and interpret.

3.5.2 Frontend functionalities

Different functionalities present at front-end are listed below:

3.5.2.1 User Interface: It has two main components. A side panel with different menus is present on the left and a map panel on its right. The side panel provides various interactive controls for uploading files and choosing different chart configuration options such as the type of chart, attributes to be displayed and aggregation function to be selected (figure 3a). It also has an option for changing the map layer for visualization. The map panel displays the generated charts and the boundary layer on an openlayers map. It has controls which allow the user to zoom in, zoom out or shift the map towards left or right.

3.5.2.2 Attribute Selection: In order to discover patterns and relations in the data, a user should be able to explore one or more data attributes. A menu on the side panel of the tool allows the user to select the attributes to be displayed on the chart. Data values from the selected attributes are later sent to phantomjs server for chart generation.

3.5.2.3 Chart type Selection: Using this feature, the user can select the type of chart (bar, pie, line) in which he wants to represent data. Different chart types are suitable for different data. It depends on data characteristics as well as the type of data (geographical, sales, health data) being used . Hence, the tool lets the user to decide and select the best chart type for this data.

3.5.2.4 Aggregation function: Using this menu option, the user can select the aggregation function to be applied on the data values of the user-selected attributes. Different aggregation functions present are sum, max, min, mean and standard deviation.

3.5.2.5 Toggle basemap: By changing the basemap, user can visualize chart values on an openlayers or the geographical boundary layer.

3.5.2.6 Rectangle selection (zoom to a single or a set of charts): The rectangle selection feature has been highlighted in figure 2b (Red border). To zoom to a single or a set of charts, a rectangle can be selected around these charts on the map using this feature. This allows for a more extensive analysis and improves the user experience.

3.6 Methodology

On loading the webpage, the user-interface presents the user with a number of menus on the side panel and a map on the right (Figure 2). Using the first menu, user can browse and upload the data file. With the second menu, user can upload the geographical boundary files in zip format. The data uploaded by the user is processed and stored by the data handling module. Based on user uploaded file, parameters for made available to the user in the menu options. User can then select type of chart, aggregation function and attributes for aggregation. Once user has selected all menu options, he can click on 'Run' button. The webapp send chart generation request to the chart export server (phantomjs) with the chart configuration options set by the user. Layer consisting of the generated charts is overlayed on the user given geographical boundary and is rendered on the map.

3.7 User Interaction (zoom-in/zoom-out)

Controls present on the map allow the user to zoom in or zoom out. With change in zoom level, there is change in the number of charts, the size of charts, the data values and the boundary file (geojson layer) to be displayed. The new set of charts to be displayed are obtained from chart cache by the tool and a vector layer is created.This layer is then rendered along with the new geojson layer in the map panel of the browser.

4. GeoBI Case Study - Online Retail Data

4.1 Aim

The main purpose of this geobi case study on a large online retail data is to understand the application of this tool and its utility to the online retailers. The dataset used is a publicly available spatio-temporal dataset with multiple attributes and multi-level geo-hierarchy.The results of geobi case study done can help online retailers to understand their customers across regions and find answers for questions such as - what their popular products are, regions where there is a demand for these, regional preferences and many others.

4.2 Data used

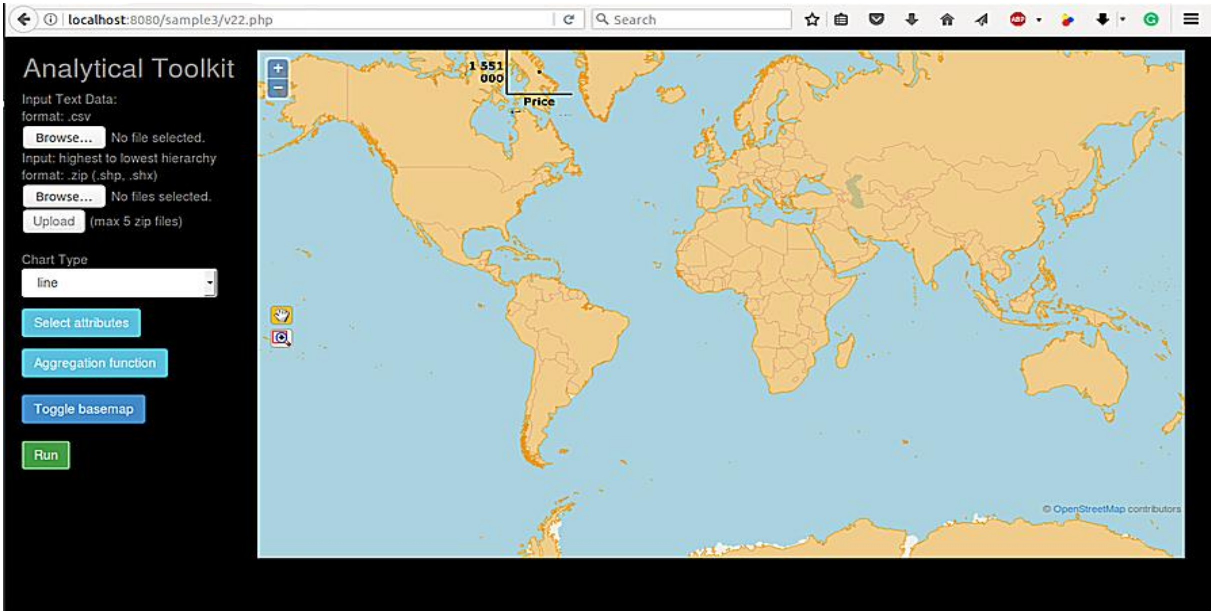
The input dataset used was online retail data of three products sold in different countries in January 2009 [6]. It consists of about 1000 records. Each record has 30 attributes. It is a spatio-temporal dataset with 4 levels of spatial hierarchy (city, state, country and continent) and day-wise sales for January 2009. Along with this, boundary shapefiles of World, Continent, UK and Canada were used [5, 9, 14]. This data was converted as per input format of the tool.

4.3 Data Analysis Results

A large number of combinations of possible questions are possible. Only the results of our analysis for different questions on the chosen dataset are presented in this section.

4.3.1 Total Sales of all products

The total sales for January 2009 all over the world was found to be 15,51,000. Figure 3 shows the chart representing this figure at the centroid of the polygon shapefile. A drill-down into map shows the sales of each continent for the January 2009 (Figure 4). It can be seen that the highest sales occur in North America (8,83,200) followed by Europe (6,06,300), Oceania (72,000), Asia (39,600), Africa (15,900) and South America (13,500). North America and Europe together constitute more than 95% of the total market share.

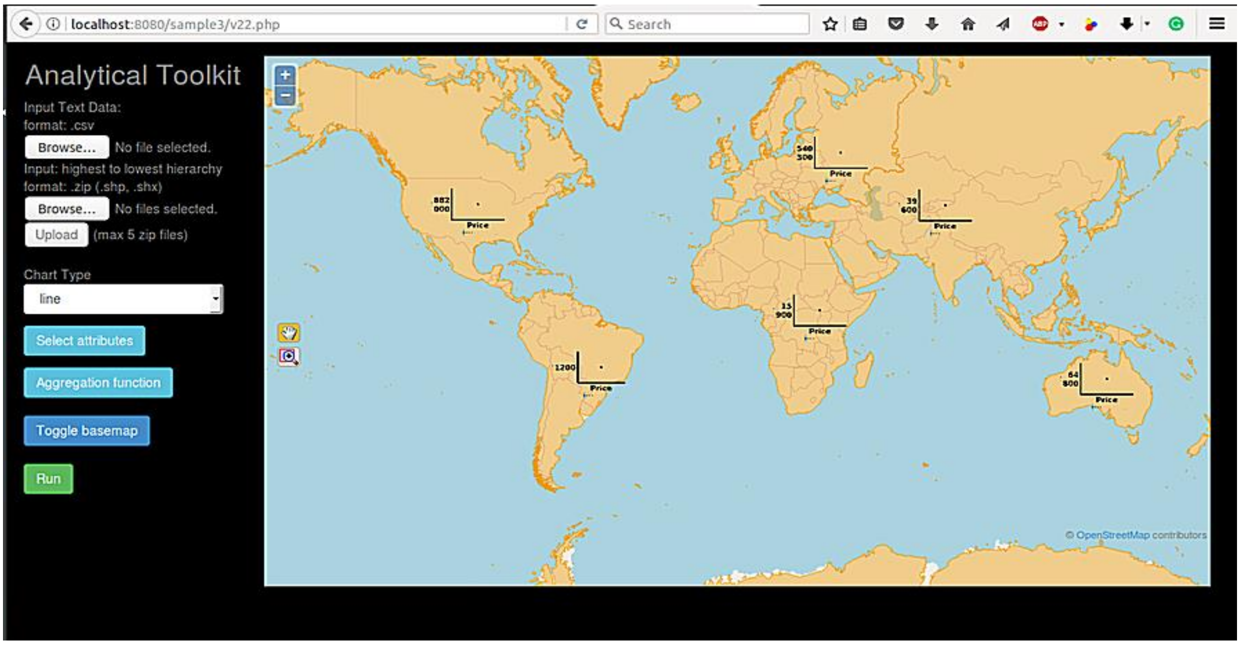


localhost:8080/sample3/v22.php

Chart Type: line

Figure 3*. Figure showing total sales all over the world

* The chart labels can be seen on zooming in the map using *zoom to a set of charts* feature of the tool



localhost:8080/sample3/v22.php

Analytical Toolkit

Input Text Data:
format: .csv
No file selected.
Input: highest to lowest hierarchy
format: .zip (.shp, .shx)
No files selected.
(max 5 zip files)

Chart Type: line

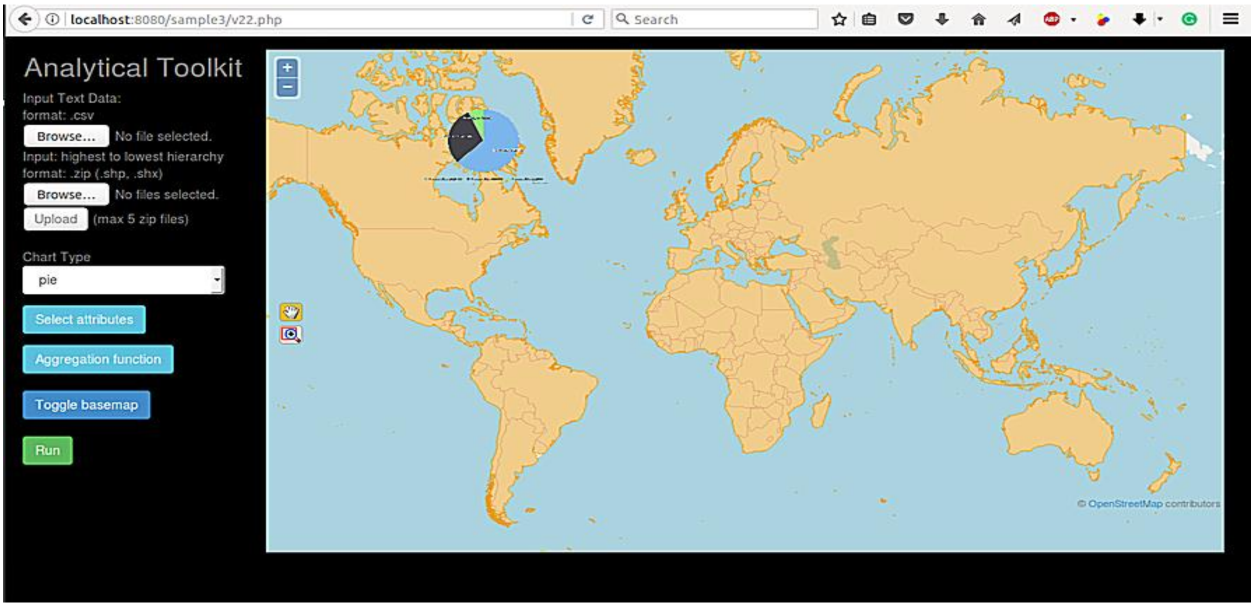
Figure 4*. Figure showing total sales in different continents

The top five countries with maximum sales were found to be United States of America, followed by United Kingdom, Canada, Switzerland and Israel on drilling down further into the map.

4.3.2 Product-wise Analysis

The dataset used consists of sales data of three products- Product1, Product2 and Product3. A comparison among the sales of these products was done (Figure 5). The blue color in the pie chart represents Product1, black color represents product2 and green represents product3. Product 1 (1028400) was found to be the most popular product, followed by Product2 (489600) and Product3 (112500).

A further drill down into the map shows the sales distribution for all products for each continent (Figure 6). We can see that in continent South America and Oceania, only Product1 was sold. In Asia and Australia, only Product1 and Product2 were sold. All the three products were sold in the rest of the continents. Differences in product preferences for each continent can be seen clearly in the figure. Moreover, in all continents except Africa, total sales were highest for Product1 followed by Product2 and Product3.



localhost:8080/sample3/v22.php

Analytical Toolkit

Input Text Data:
format: .csv
No file selected.
Input: highest to lowest hierarchy
format: .zip (.shp, .shx)
No files selected.
(max 5 zip files)

Chart Type: pie

Figure 5*. Pie chart showing total sales for different products all over the world

* The chart labels can be seen on zooming in the map using *zoom to a set of charts* feature of the tool

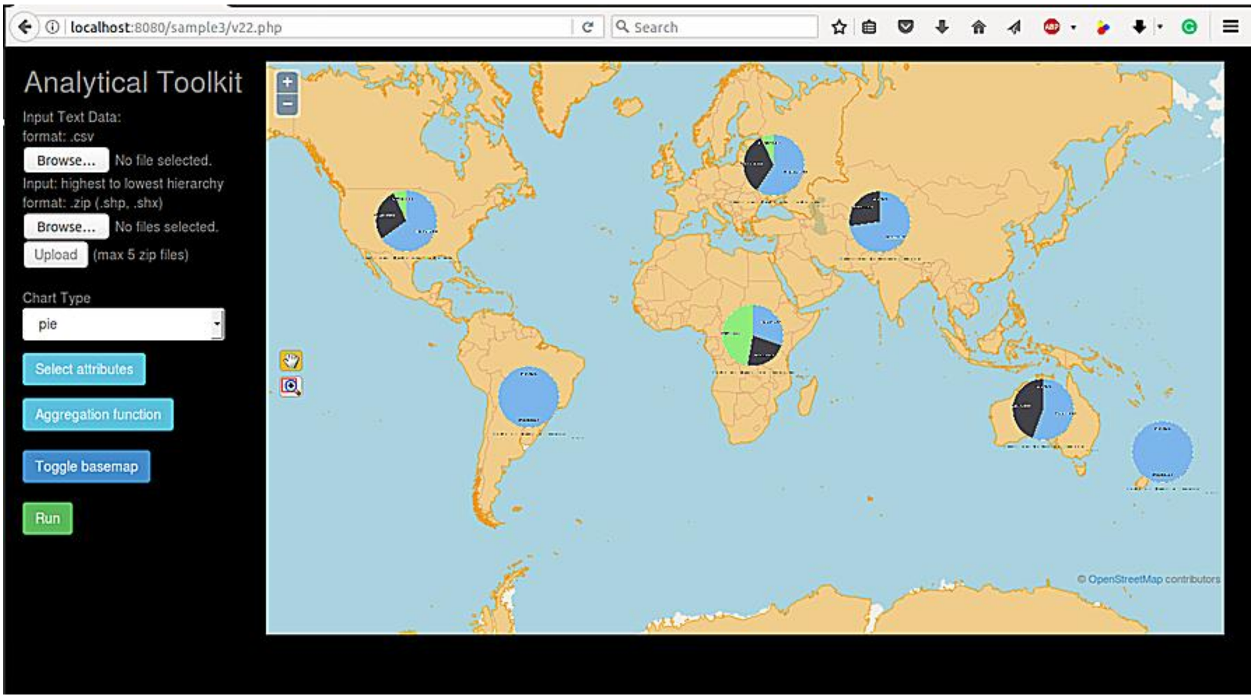


Figure 6* . Pie chart showing total sales for different products for different continents all over the world

Analysis of sales of all the three products in Canada was done. Only Product1 and Product2 were found to be sold. Product preferences of different provinces in Canada can be seen from figure 7 below.

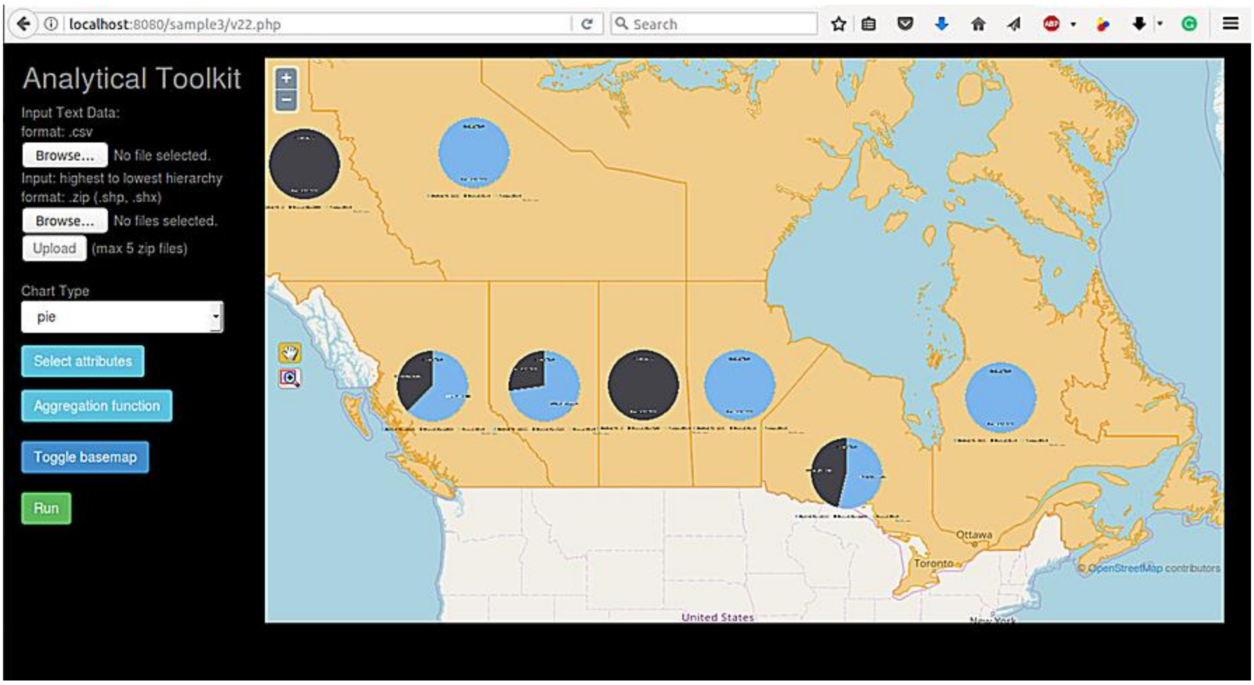


Figure 7*. Pie chart showing total sales for different products for different provinces in Canada

4.3.3 Time-series analysis

To understand the peak hour for online sales, sales data for each hour-wise was analyzed. It was found that maximum sales occurred at 7am in the morning (18000) (Figure 8). A huge number of products were sold between 9am to 11am. 12 pm onwards it declined and continuously declines. It came to a halt at 4pm. Markers in figure 8 show the geographic location of products sold in United Kingdom (UK).

* The chart labels can be seen on zooming in the map using *zoom to a set of charts* feature of the tool

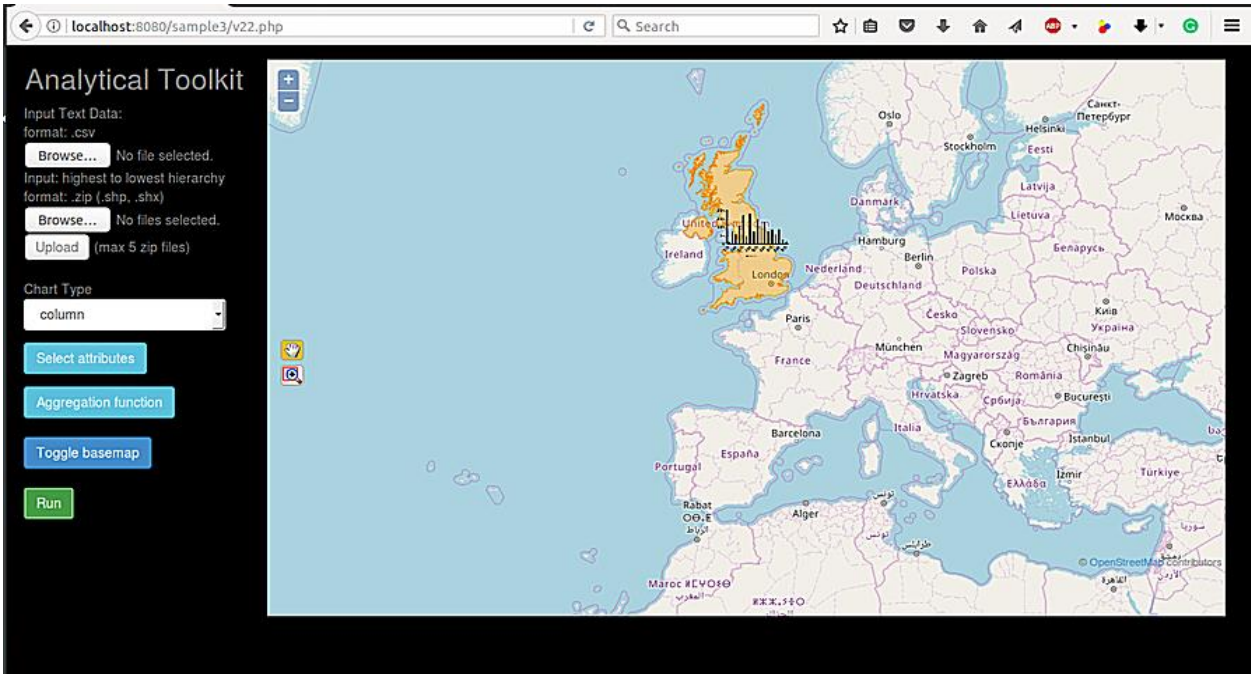


Chart Type: column

Figure 8*. Column chart showing hourly sales of products in United Kingdom (UK)

4.3.4 Payment Option Analysis

Figure 9 shows us the total sales for each payment option- Amex, Visa, MasterCard and Diners.Visa is the most popularly used option followed by MasterCard, Amex and Diners.

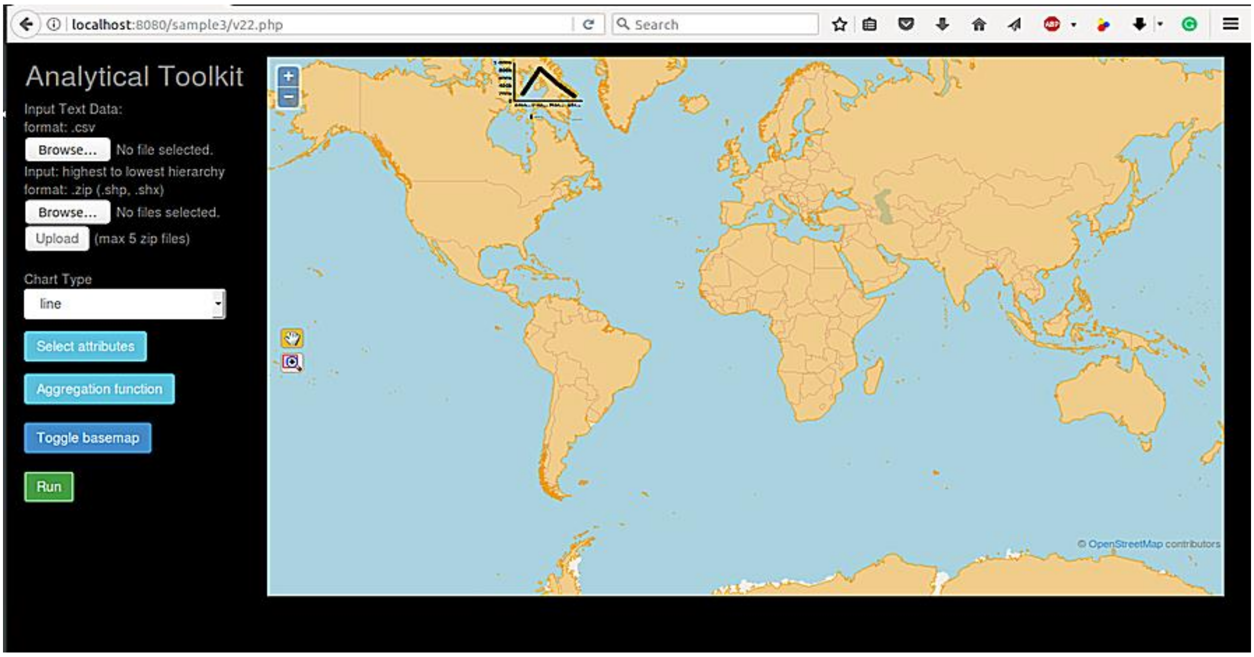

localhost:8080/sample3/v22.php
Search
Analytical Toolkit
Input Text Data:
format: .csv
Browse... No file selected.
Input: highest to lowest hierarchy
format: .zip (.shp, .shx)
Browse... No files selected.
Upload (max 5 zip files)
Select attributes
Aggregation function
Toggle basemap
Run
© OpenStreetMap contributors

Chart Type: line

Figure 9*. Line graph showing total sales using different payment types all over the world

5. CONCLUSION

An interactive, web-based, Spatial-OLAP tool to find hidden patterns using both user-driven and data-driven analysis has been presented. The tool produces an interactive map with online analytical capabilities (OLAP) and computes data aggregation based on user-defined geo-hierarchy with dynamic visualizations in a defined framework. It has been entirely developed using open source suite of technologies. With this tool, any user can successfully visualize data with the help of a web browser without requiring any additional software, training or

* The chart labels can be seen on zooming in the map using *zoom to a set of charts* feature of the tool

knowledge of GIS. A server-side based approach has been used for generating charts and displaying them on a map. The parameters displayed in the menu are added on the fly and not precomputed.

A GeoBI case study on a large online retail data of three products sold in different countries during January 2009 has been presented to understand the application of this tool and its utility to the online retailers. Total sales all over the world were found to be 1551000. It was found that the top five countries with maximum sales were United States of America, followed by United Kingdom, Canada, Switzerland and Israel. Maximum quantity of product1 was sold (1028400), followed by product2 (489600) and product3 (112500). Product preferences for all continents were found. Similar analysis for done for different provinces in Canada. The payment options in decreasing order of popularity were found to be Visa, MasterCard, Amex and Diners. To understand the peak hour for online sales in United Kingdom (UK), time-series analysis was done. It was found that maximum sales occurred at 7am in the morning (18000). Almost no sales occurred after 4pm. A working model of this tool has been developed. It will be made available on the server or as a service at a later stage.

REFERENCES

[1] Alan M. MacEachren and Menno-Jan Kraak. Research challenges in geovisualization. Cartography and Geographic Information Science, 28(1), 2001.
[2] Andrienko, G., Andrienko, N., 1999. Interactive maps for visual data exploration, International Journal of Geographical Information Science, 13(4), pp. 355-374.
[3] Andrienko, N., Andrienko, G., Gatalsky, P., 2000, Towards Exploratory Visualization of Spatio-Temporal Data, In: 3 rd AGILE Conference on Geographic Information Science- Helsinki/Espoo, Finland, May 25 th -27 th, 2000.
[4] Apache Tomcat (2017) Apache Tomcat [Online] http://tomcat.apache.org/
[5] ArcGIS Continent Shapefile http://www.arcgis.com/home/item.html?id=3c4741e22e2e4af2bd4050511b9fc6ad
[6] BigML Data Repository https://bigml.com/user/czuriaga/gallery/dataset/555c67beaf447f4a73001457#info
[7] Chaudhuri, S., Dayal, U.: An overview of data warehousing and OLAP technology. ACM
SIGM OD Rec. 26(1), 65–74 (1997).
[8] Diansheng Guo, Jin Chen, Alan M. MacEachren, and Ke Liao. A visualization system for space-time and multivariate patterns (vis-stamp). IEEE Transactions on Visualization and Computer Graphics, 12(6):1461–1474, November 2006.
[9] Global Administrative Areas http://www.gadm.org/download
[10] Highcharts (2017) Highcharts [Online] http://www.highcharts.com/
[11] Jianghui Ying, Denis Gracanin, and Chang-Tien Lu, "Web Visualization of Geo-Spatial Data using SVG and VRML/X3D," Proceedings of the Third International Conference on Image and Graphics (ICIG'04).
[12] JQuery. (2017) JQuery. [Online]. http://jquery.com/
[13] Natalia Andrienko, Gennady Andrienko, and Peter Gatalsky. Exploratory spatio-temporal visualization: an analytical review. Journal of Visual Languages & Computing, 14(6):503 – 541, 2003.
[14] National Weather Service OST/SEC GIS Map Group http://www.nws.noaa.gov/geodata/catalog/national/html/province.htm
[15] Openlayers https://openlayers.org/
[16] PHP Tutorial https://www.w3schools.com/php/
[17] Q. Ho, Q., P. Lundblad, P., T. Åström, T., and M. Jern, M., 2011, A Web-Enabled Visualization Toolkit for Geovisual Analytics Visualisation and Data Analysis, In: Proceedings of SPIE: Electronic Imaging Science and Technology, Visualization and Data Analysis, San Francisco, USA Jan 2011.
[18] Thematic Mapping- World Borders Dataset http://thematicmapping.org/downloads/world_borders.php