OLAP Cube Visualization of Environmental Data Catalogs

Bora Beran[1], Catharine van Ingen[1], Ilya Zaslavsky[2], David Valentine[2]
[1]Microsoft Research, [2] San Diego Supercomputer Center

Microsoft Technical Report **MSR-TR-2008-70**
July 14, 2008

Abstract

Systems like SciScope and Data Access System for Hydrology (DASH) rely on data catalogs to facilitate data discovery. These catalogs describe several nation-wide data repositories that are important for scientists including US Geological Survey's National Water Information System (NWIS), Environmental Protection Agency's STOrage and RETrieval System (EPA STORET) and National Climatic Data Center (NCDC) data collections which contain a wealth of information reflecting the history and geography of environmental data collection efforts in the US.

OLAP (Online Analytical Processing) is an approach to organizing and querying large data collections. End user applications can request slices from OLAP servers and provide multi-dimensional visualizations that involve aggregations, ranking and algebraic/statistical calculations to gain greater insight into the data. We have applied the OLAP technology to environmental data catalogs using SQL Server 2008 Analysis Services and used Excel and Virtual Earth to visualize the query results.

1. Introduction

Dealing with distributed databases operated by different bodies involves maintaining metadata records (hereafter "data catalogs") for brokering communications between systems. These catalogs contain valuable information on data availability and geographical distribution in addition to the history of data collection efforts which is not often utilized. Catalogs are usually stored in relational database systems, which are typically optimized for online transactional processing (OLTP). While the transactional databases are good at executing a large number of small transactions concurrently, for scientists who need aggregates and summarized information, comparisons in space and time over a large number of occurrences for operations such as trends discovery, they are inefficient and difficult to use. [1] To address these issues we made use of OLAP (Online Analytical Processing) technology and examined catalogs from the three major data collection agencies for hydrologic, environmental and meteorological data in the US, namely US Geological Survey's National Water Information System (NWIS) [2], Environmental Protection Agency's STOrage and RETrieval System (EPA STORET) [3] and National Oceanographic and Atmospheric Administration (NOAA)'s National Climatic Data Center (NCDC) [4].

2. On-Line Analytical Processing

OLAP is "the name given to the dynamic enterprise analysis required to create, manipulate, animate and synthesize information from exegetical, contemplative and formulaic data analysis models. This includes the ability to discern new or unanticipated relationships between variables, the ability to identify the parameters necessary to handle large amounts of data, to create an unlimited number of dimensions, and to specify cross-dimensional conditions and expressions". [5]

OLAP technology relies on the multidimensional database approach. The dimensions represent the themes of interest and are organized hierarchically according to levels of granularity. For example "hydrologic unit" would be a dimension with a hierarchy starting with "hydrologic region" at the top and with "subbasin" at the bottom.

Dimensions contain members. "Upper Mississippi" would be a member of "hydrologic region" dimension while "Russian River" would be a member of "subbasin". The members of one level in the hierarchy can be aggregated to the members of higher levels. Time can also be a dimension in which case hierarchies would include items such as "day", "decade" and "year".

Other key concepts in OLAP are: measures, facts and data cubes. [6] The measures are the numerical attributes analyzed against the different dimensions which can built-in functions such as sum, count and average as well as complex, user-defined formulas. Dimensions provide the context for the measure and together they constitute facts. For example "count of dissolved oxygen measurements made by USGS in the Lower Washita watershed is 10" is a fact where "count" is the measure while "dissolved oxygen", "USGS" and "Lower Washita" are members of different dimensions.

A set of measures aggregated according to a set of dimensions comprise a data cube. [7] Several data cubes can be built from the same database for different analysis needs. Aggregations can be pre-computed to a certain level to increase the query performance.

Datacubes can be queried using the MDX (Multi-Dimensional eXpressions) query language [8]. MDX resembles the SQL (Structured Query Language) but has some differences to support multidimensionality.

Simplicity of using data cubes is because of the availability of GUI-based access tools. Excel [9], Tableau [10], Proclarity [11], and Cognos [12] are some examples of such software.

OLAP has been used broadly in finance, sales and marketing but its applications in scientific studies are relatively new. [13] We have been working with OLAP using large environmental datasets and one particular application we found useful was examining the data inventory over space and time. We based the cube on the core components of a data model that we developed earlier. [14]

3. Cube Implementation

Figure 1 shows the structure of our cube. At the center is the catalog of time series data

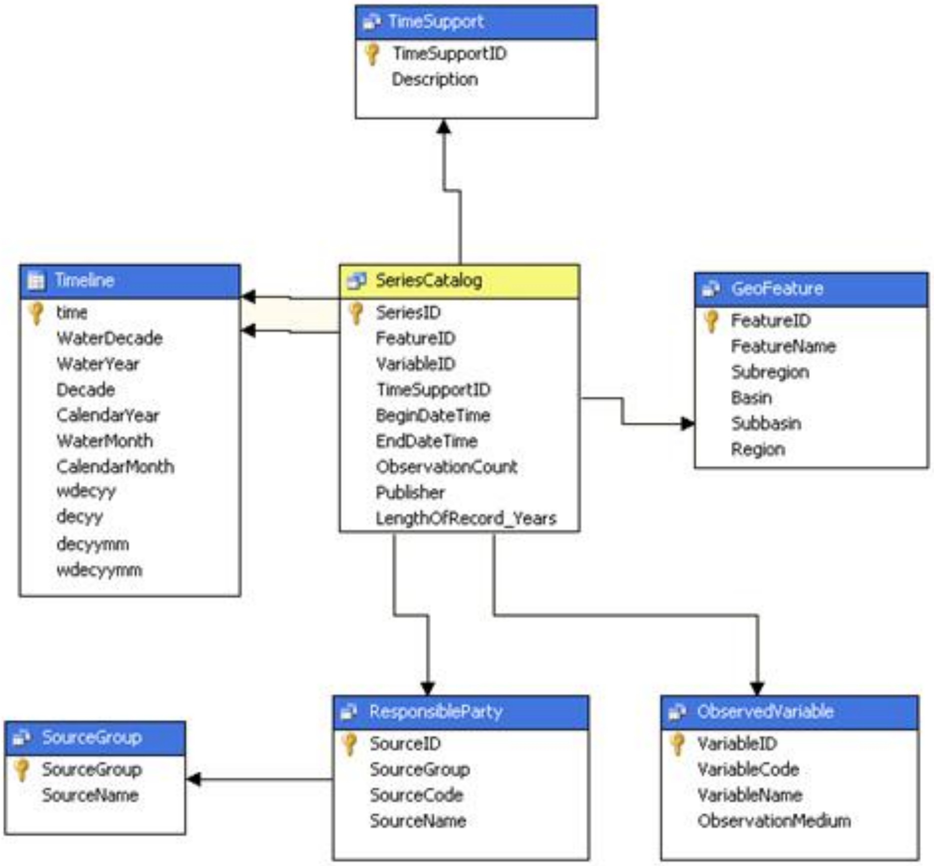

Figure 1 Cube Structure

with 5 dimensions providing spatial, temporal and thematic context of observations. Series Catalog contains 9.3 million entries representing a total of over 358 million observations.

GeoFeature dimension contains information about the location of sensor which collected the data. GeoFeature dimension has information on 1.7 million sensors in the US. It also contains information about which Hydrologic Units it belongs to. "A hydrologic unit is a drainage area delineated to nest in a multi-level, hierarchical drainage system. Its boundaries are defined by hydrographic and topographic criteria that delineate an area of land upstream from a specific point on a river, stream or similar surface waters." [15] USGS maintains a hierarchical hydrologic unit code (HUC) system for the United States which divides the country into 21 Regions, 222 Subregions, 352 Basins, and 2,149 Subbasins based on surface hydrologic features. Following the HUC system, in the cube hydrologic units are organized in a hierarchy. Hydrologic units and sensors are associated dynamically during the cube build using a view that involves a query like:

```
SELECT S.FeatureID as SensorID,
S.FeatureName as SensorName,
W.FeatureName as Region from
SciScope.dbo.ODCore_Feature S JOIN
SciScope.dbo.ODCore_Feature W on
W.GeoPosition.STIntersects(S.GeoPosit
ion)=1 and
W.FeatureType='HydroRegion' and
S.FeatureType='Sensor'
```

ObservedVariable dimension provides information about the measured variable and medium whereas **TimeSupport** clarifies how it is reported for example "Daily average", "Hourly incremental" or "15-minute instantaneous".

ResponsibleParty dimension contains information on the originators and publishers of the data. Originators are grouped under publishers in a hierarchy. For example EPA publishes data from 279 groups (states, Indian tribes etc.) thus the group named EPA STORET contains 279 children.

Timeline dimension contains the timeframe for time series and organized in two hierarchies: from decade to calendar month and water decade to water month. A water year starts in October of the prior calendar year and ends in September of the year it's named after. For example, water year 2008 covers the period October 1, 2007 through September 30, 2008. Since SQL Server Analysis Service (SSAS) [16] doesn't natively support water calendar, we created a custom timeline for our cube.

4. Browsing the cube

One piece of information that can be retrieved from a catalog cube is the continuity. Figure 2 shows the length of data collection activities for EPA STORET's water quality data by decade. It can be seen that starting in 1950 majority of campaigns lasted 5 or less years. In fact today %60 of the measurement series is 1 or fewer years of length.

If we examine the same question from a different point of view and using USGS data as in Figure 3 we can see how number of active stream gages and their length of record changes over time.

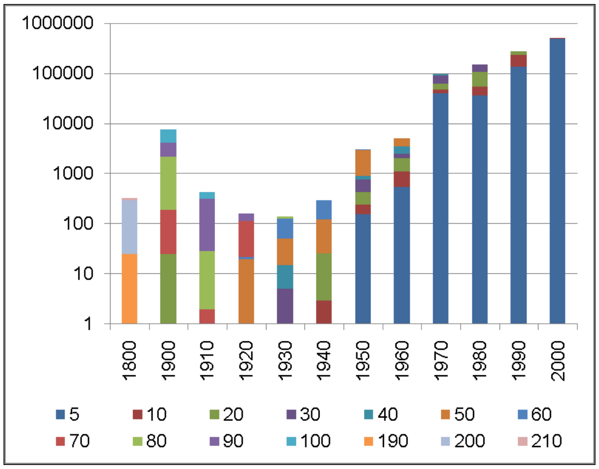

Figure 2 Site longevity in EPA STORET

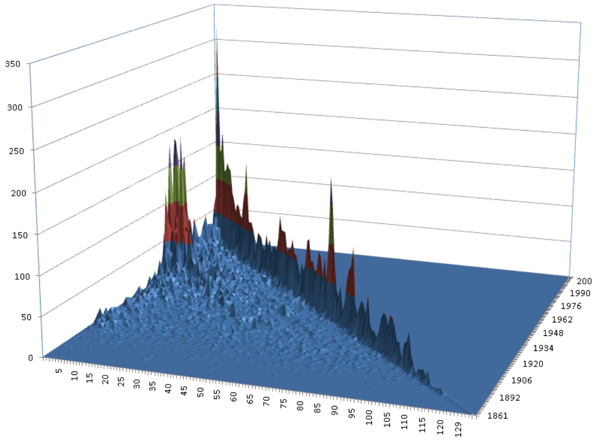

Figure 3 USGS sites and length of record over time

In Figure 3 from left to right extend the number of years of activity, z dimension contains the year of birth for sites while y dimension shows the number of sites. In this plot currently active sites fall on the diagonal. In 1960 Congress passed the first Clean Water Act, followed by Water Quality Act in 1965 and Wild and Scenic Rivers Act in 1967. It is possible to see this as a spike that appears around 1960-1973 that doesn't follow the diagonal. Around 1938-1940 can be seen another peak (this time on the diagonal) corresponding to Federal Stream

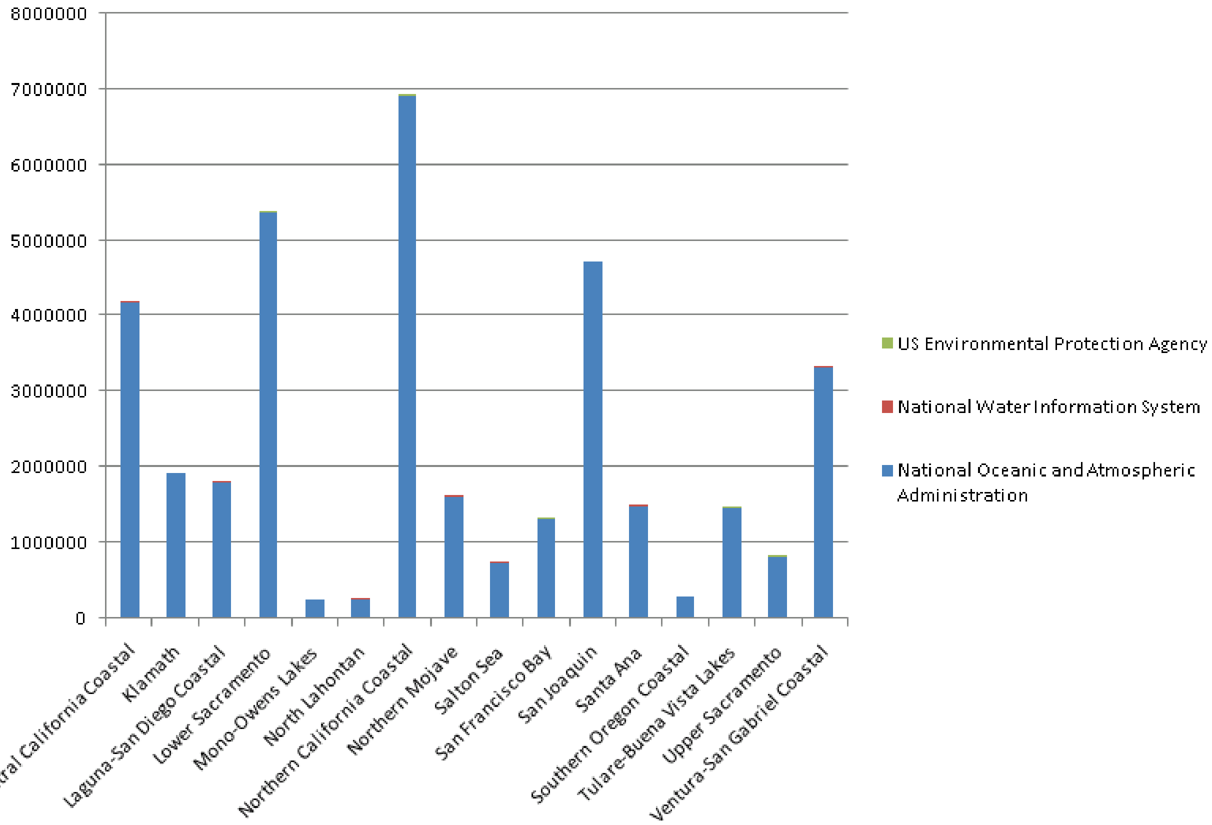

Figure 4 Precipitation data distribution across river basins in California

Pollution Bill that was approved by the Congress on June 13, 1938.

Figure 4 shows the availability of precipitation data in California and its distribution across river basins from NOAA's National Weather Service, EPA's STORET and USGS' NWIS. It can be seen that NOAA maintains the majority of rainfall data. A similar data availability analysis can be done to identify areas that have the necessary data for a particular type of study. The query that returned the results in Figure 4 takes less than 1 second using the OLAP data cube while it takes 326 minutes with a regular SQL query.

Since the data has a spatial component it is often useful to be able to visualize it on a map. Geographical relations may not only help spatial patterns to stand out but also makes the data more meaningful for people who're not familiar with the geography of a particular region.

5. Coupling OLAP with Virtual Earth

Spatial OLAP (SOLAP) brings Geographical Information Systems (GIS) and OLAP technologies together to provide an alternative way of examining the data. Benefits of map visualization, frequently exemplified by John Snow's 'Cholera map' from 1854 are well established. However common OLAP clients like Excel, Proclarity, Cognos and Novaview [17] do not support cartographic displays. While products and plug-ins with varying levels of GIS-OLAP integration are available, they require commercial mapping tools such as ArcGIS, ArcIMS [18] and MapInfo [19]. Some dashboard tools such as Dundas [20] and iDashboards [21] also provide basic map visualizations but lack the spatial depth of real mapping applications. Tools such as Virtual Earth [22] on the other hand provide detailed imagery and base maps but lack OLAP integration. Applications such as SciScope extend Virtual Earth with additional map layers making it an even more valuable tool for domain scientists. Hence to spatially enable our cube we decided to leverage Virtual Earth.

Our cube already contains nominal spatial references such as observation sites and hydrologic units. To be able to display the information on Virtual Earth these nominal references need to be associated with their geographical counterparts. In our implementation we used a SQL Server 2008 [23] database exposed through a SOAP web service for this purpose although it could have been a Web Feature Service (WFS) [24], a gazetteer or a flat file containing the coordinates. Figure 5 shows the data flow diagram for the system. ASP .NET application underlying the map accesses SQL Server Analysis Services (SSAS) using the ADOMD .NET Client [25] dynamic-link library and reads cube metadata such as dimensions, hierarchies and hierarchy levels then populates three TreeView [26] controls. Each node of the tree has certain actions associated with it implemented using JavaScript. For example a measure can be added to the values group while a regular dimension can be added only as a filter. On the other hand members of the spatial dimension can be added to both label axis and the list of filters. This is because the interface requires one axis to contain geographical references and the other to have values. The tool requires one dimension of the cube to be designated as “spatial dimension” any dimension that is not a measure or spatial are referred to as regular dimensions.

A spatial dimension should have field that uniquely identifies the geographic feature to be used in retrieving the coordinates from a spatial data source. These identifiers can be codes such as hydrologic unit code (HUC) or Federal Information Processing Standards (FIPS) codes for states, counties, International Organization for Standardization's ISO 3166 country codes or unique names. Filter, value and label settings are stored in a JavaScript array rendered in a way that resembles Excel’s pivot table interface. Clicking filters and labels pops up another window with a TreeView control that allows drilling down. Nodes are populated on the fly as user clicks since the total number of instances can be too high to load at once. For example our catalog contains about 1.7 million sensors while drilling down the hierarchy for a given subbasin the number is rarely more than a hundred. Drill down is processed using server side code which requires a postback and is handled using an ASP .NET update panel in order not to trigger a full page refresh. Once user sets up the query parameters through the interface (filters, measures and a spatial dimension) and submits the form, an Asynchronous JavaScript and XML (AJAX) call is made to a SOAP web service. Web service creates an MDX query based on user’s selections and interrogates the SSAS database. MDX query returns measures coupled with nominal spatial references which are converted to geographical coordinates using a spatial data source which is SQL Server 2008 in our case. Interface displays the geographical coordinates on Virtual Earth using colors generated by normalizing measures and converting them to RGB values. WKT GeometryTypes are converted to Virtual Earth (VE) ShapeType counterparts. For example POINT, LINESTRING, and POLYGON are converted to VE pushpin, polyline and polygon respectively while MULTIPOLYGON is made into multiple VE polygons. Figure 5 shows the data flow diagram for the system. Depending on the capabilities of the spatial data source, it is fairly easy to extend the spatial OLAP capabilities with density visualizations, area-weighted aggregations and geographical area identification using spatial relationships (e.g. ST_Touches, ST_Crosses, ST_Overlaps). Figure 6 shows some examples of OLAP query results visualized using Virtual Earth. In this figure, to the left is a visualization of sensor count in Upper Gila basin in Arizona. Next image shows sensor density (#/km2) for all subbasins in Upper Gila basin. Third image shows data availability for individual sensors, while the rightmost image shows how polygons in 2nd image appear over a digital elevation

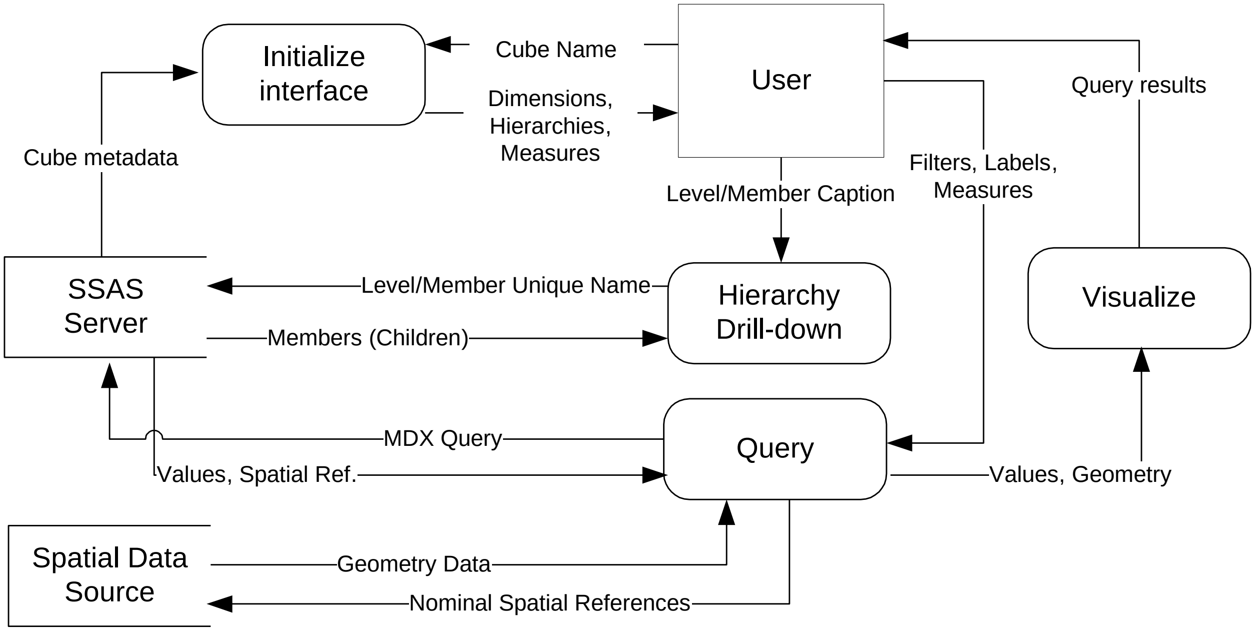

Figure 5 Data flow diagram for the system

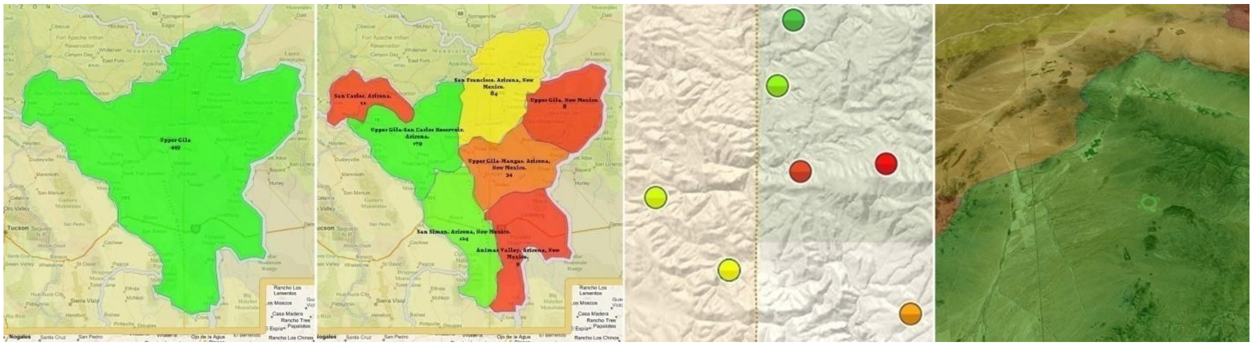

Figure 6 SOLAP with Virtual Earth (Polygons, Points on 2D and 3D maps, left to right)

model (DEM) in Virtual Earth 3D.

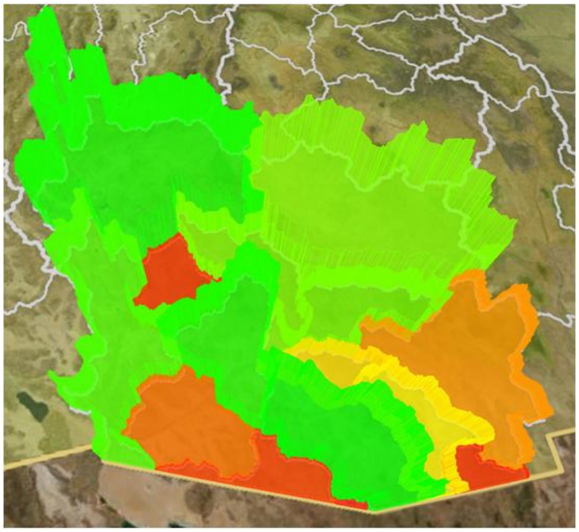

Figure 7 Using elevation for thematic mapping

Elevation can also be used as another axis for thematic mapping. Figure 7 shows a 3D map that uses colors to indicate sensor counts and elevations to represent observation counts.

6. Conclusions and Future Directions

Leveraging OLAP technology for visualizing environmental data catalogs can be quite useful in understanding the data availability and distribution over space and time. We demonstrated some use cases based on observations metadata from 3 major environmental data repositories in the US also outlined the implementation the cube and an ASP .NET SOLAP client built on Microsoft Virtual Earth.

While this report deals with data catalogs, a similar system could be used for visualizing observation results. However different observations require different ways of spatial aggregation. In some cases e.g. census data the way the aggregations work is no different from the data catalogs. On the other hand data from rain gages or stream gages require special aggregation methods also depending on the type of the analysis the scientist is after. For example use of Thiessen Polygons is very common for rain gages while in order to calculate total inflow/outflow to/from a watershed, one needs to know about parameters like flow direction and stream convergence. Possibility of creating such custom applications using OLAP technology indicates that OLAP has a lot of potential for use in handling scientific data.

References

[1] Date, C.J., 2003. An Introduction to Database Systems, 8th edition. Addison-Wesley.

[2] United States Geological Survey, http://www.usgs.gov and http://waterdata.usgs.gov/nwis .

[3] Environmental Protection Agency, http://www.epa.gov .

[4] National Climatic Data Center, http://www.ncdc.noaa.gov/.

[5] Codd, E.F., Codd, S.B., Salley, C.T., 1993. Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate. E.F. Codd Associates. http://www.cs.bgu.ac.il/~dbm031/dw042/Papers/olap_to_useranalysts_wp.pdf

[6] Berson, A., Smith, S.J., 1997. Data Warehousing, Data Mining and OLAP. McGraw-Hill.

[7] Thomsen, E., Spofford, G., Chase, D., 1999. Microsoft OLAP Solutions. John Wiley and Sons.

[8] Multi Dimensional eXpressions (MDX), a query language to query the SQL Server Analysis Services, http://msdn2.microsoft.com/en-us/library/ms345116.aspx

[9] Excel Pivot tables, http://www.microsoft.com/dynamics/using/excel_pivot_tables_collins.mspx

[10] Tableau, A tool for querying and analyzing OLAP databases without any knowledge of MDX, http://www.tableausoftware.com/info/OLAP_Front_End/OLAP_Front_End_fw.php

[11] Proclarity, http://www.proclarity.com

[12] Cognos, http://www.cognos.com/solutions/index.html

[13] Ozer, S., Szalay, A., Szlavecz, K., Terzis, A., Musaloiu-E., R., Cogan, J., Using Data-Cubes in Science: an Example from Environmental Monitoring of the Soil Ecosystem, MSR-TR-2006-134, 2006.

[14] Beran, B., Valentine, D., van Ingen, C., Zaslavsky, I., Whitenack, T., A Data Model for Environmental Observations, MSR-TR-2008-92 , 2008.

[15] Federal Standards for Delineation of Hydrologic Unit Boundaries. FGDC, 2004, ftp://ftp-fc.sc.egov.usda.gov/NCGC/products/watershed/hu-standards.pdf

[16] SQL Server Analysis Server, An integrated view of business data for reporting, OLAP analysis, Key Performance Indicator (KPI) scorecards, and data mining, http://www.microsoft.com/sql/technologies/analysis/default.mspx

[17] Panorama Novaview, http://www.panorama.com

[18] ESRI ArcGIS and ArcIMS, http://www.esri.com

[19] MapInfo, http://www.mapinfo.com

[20] Dundas, http://www.dundas.com

[21] Idashboards, http://www.idashboards.com/

[22] Virtual Earth, http://www.microsoft.com/virtualearth/

[23] SQL Server 2008, http://www.microsoft.com/sqlserver/2008/en/us/default.aspx

[24] Vretanos P. A., (Ed), 2005. Web Feature Service Implementation Specification, Open Geospatial Consortium Inc., 117 pp.

[25] ADOMD.NET Client Concepts and Object Model, http://msdn.microsoft.com/en-us/library/bb522641.aspx

[26] ASP .NET CSS Friendly Control Adapters, http://www.asp.net/cssadapters/