Stephan, Eric and Gillen, D and Burke, J and Joslyn, CA and Olsen, Bryan and Critchlow, Terence and Peterson, Elena: (2009) "Geospatial Data Analysis for DHS Programs", in: Proc. 2009 IEEE Int. Conf. Technologies for Homeland Security (HST 09), pp. 553-560, IEEE, https://doi.org/10.1109/THS.2009.5168086

GeoSpatial Data Analysis for DHS Programs

Eric Stephan, John Burke, Carrie Carlson, Dave Gillen , Cliff Joslyn, Bryan Olsen, Terence Critchlow

Pacific Northwest National Laboratory
PO Box 999 Richland Wa 99354
eric.stephan@pnl.gov
john.burke@pnl.gov
carrie.carlson@pnl.gov
david.gillen@pnl.gov
cjoslyn@pnl.gov
bryan.olsen@pnl.gov
terence.critchlow@pnl.gov

Abstract— **The Department of Homeland Security law enforcement faces the continual challenge of analyzing their custom data sources in a geospatial context. From a strategic perspective law enforcement has certain requirements to first broadly characterize a given situation using their custom data sources and then once it is summarily understood, to geospatially analyze their data in detail.**

For the past three years the Generalized Data-Driven Analysis and Integration (GDDAI) project has operated through DHS Science and Technology on behalf of Immigration and Customs Enforcement (ICE) to develop and deploy hybrid data analysis methods that improve ICE analysts' ability to answer time-critical queries for their agents. GDDAI provides this hybrid environment through seamless customized interfaces built with commercial off the shelf (COTS) and government off the shelf (GOTS) tools.

At its core, to help characterize user data, GDDAI relies upon multidimensional relational data cubes which are based on OnLine Analytical Processing (OLAP) technologies that provide intuitive and graphical access to the massively complex set of summary views available in large relational (SQL) repositories. As an extension to OLAP we provide a proximity-based search engine, a rich map feature database which allows users to add their customized reports, and a geocoder that derives coordinates from addressable information in the user database.

While the design of data cubes can be technically challenging, our primary contributions are in the ability to effectively "drill-through" from a summary analysis to a detailed analysis when a population of interest is identified. In addition to this data transformation capability, our geospatial analysis infrastructure can perform proximity searches on the detailed "drill-through" records. Iteratively users can conduct increasingly more fine-tuned geospatial searches by combining or excluding previous geospatial results with their current drill-through results.

I. INTRODUCTION

Over the past three years the Generalized Data-Driven Analysis and Integration (GDDAI) Project has developed and deployed geospatial and mapping capabilities to our customers within the Department of Homeland Security (DHS) Immigration and Customs Enforcement (ICE) agency. ICE agents, like most law enforcement agencies, use information from a variety of data sources as part of their regular operations. This information can range from data structured in relational databases, to free text containing information about addresses and regions of interest, to other formats like images. Geographical information systems help users expand their knowledge about a particular problem by analyzing their data in the context of map features. Geospatial analysis integrates regional map data sets, user custom data, and geospatial algorithms. These algorithms analyze the custom data in the context of the regional maps.

A prominent geospatial analysis technique is known as proximity search. This spatial query allows users to specify source map features, target map features, and the qualifying distance between the source and target. If a given map feature source is within the specified distance of a feature target, then the source feature is selected. The search returns the set of identified source features selected by the proximity search.

The law enforcement community benefits from proximity searches by pinpointing spatial data of interest in the context of infrastructure, landmarks, specific points of interest, or established regions. Usually off-the-shelf solutions are well equipped to help users perform their analysis using either simple search criteria or small data sets. However from the novice user perspective proximity searches become far more complex when the user's dataset is large, multifaceted, and the search criteria consists of many iterative ad-hoc questions.

In this paper we describe our work developing an end-user analytical tool to quickly and efficiently perform complex proximity searches against large databases. This geospatial tool is a component within GDDAI which integrates multi-dimensional analysis, link analysis, statistics, and geospatial analysis. The core of GDDAI begins with providing a multi-dimensional view of the user's data. With visual multidimensional tools the user is able to navigate in an *ad-hoc* fashion to intuitively choose data characteristics and levels of summarization. Once a relevant summarization view

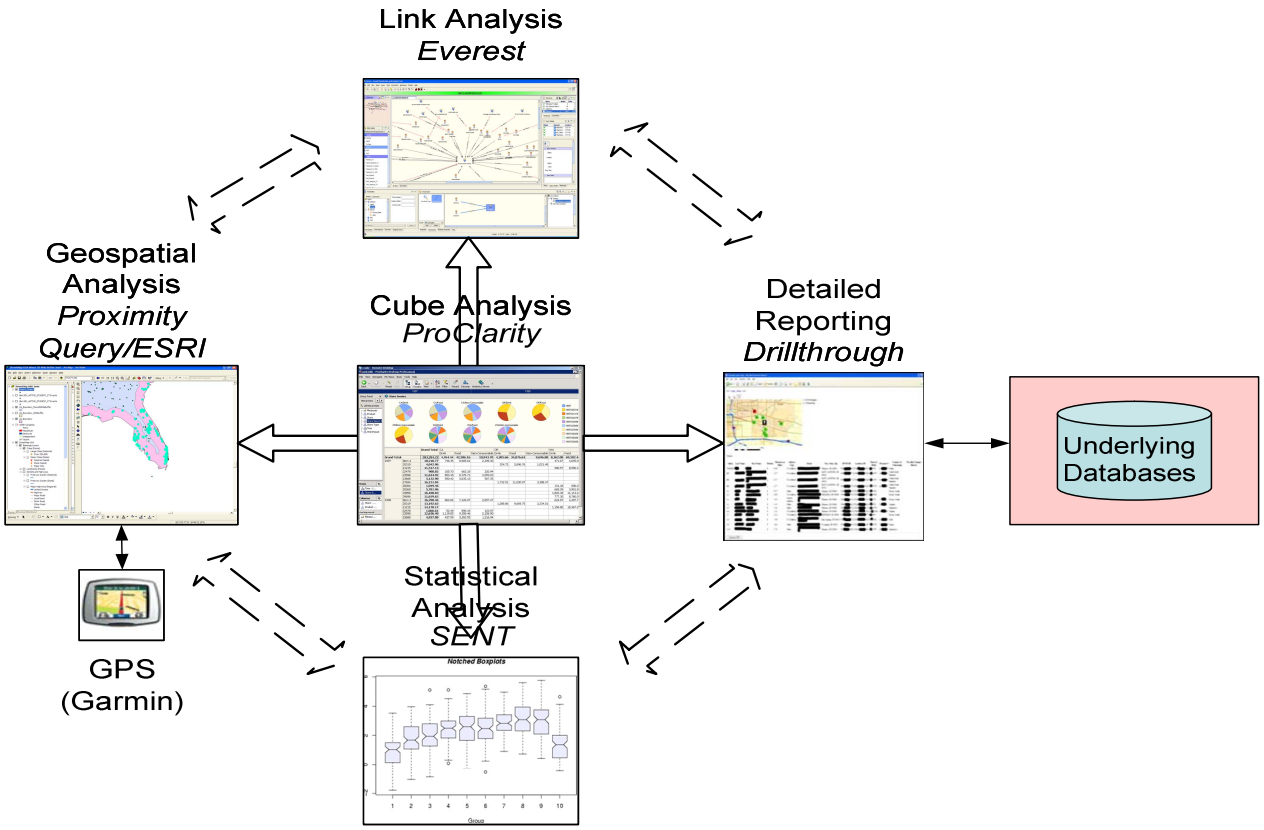


Fig. 1. GDDAI Conceptual Architecture

is chosen the user may choose to perform the link analysis, statistics or geospatial analysis on the subset of data.

GDDAI provides an environment to enable complex geospatial data analysis by:

- Helping users intuitively characterize analytical questions with OLAP tools and view answers at a summary level.
- Providing "drill-through" operations which retrieve detailed records based on the summary level answers.
- Iteratively performing complex proximity searches on the detailed records using a search wizard to perform spatial queries against a GIS.

This paper describes a use case that illustrates the problem space we are addressing, our technical solution, and provides a discussion section describing how proximity search problems can benefit from GDDAI.

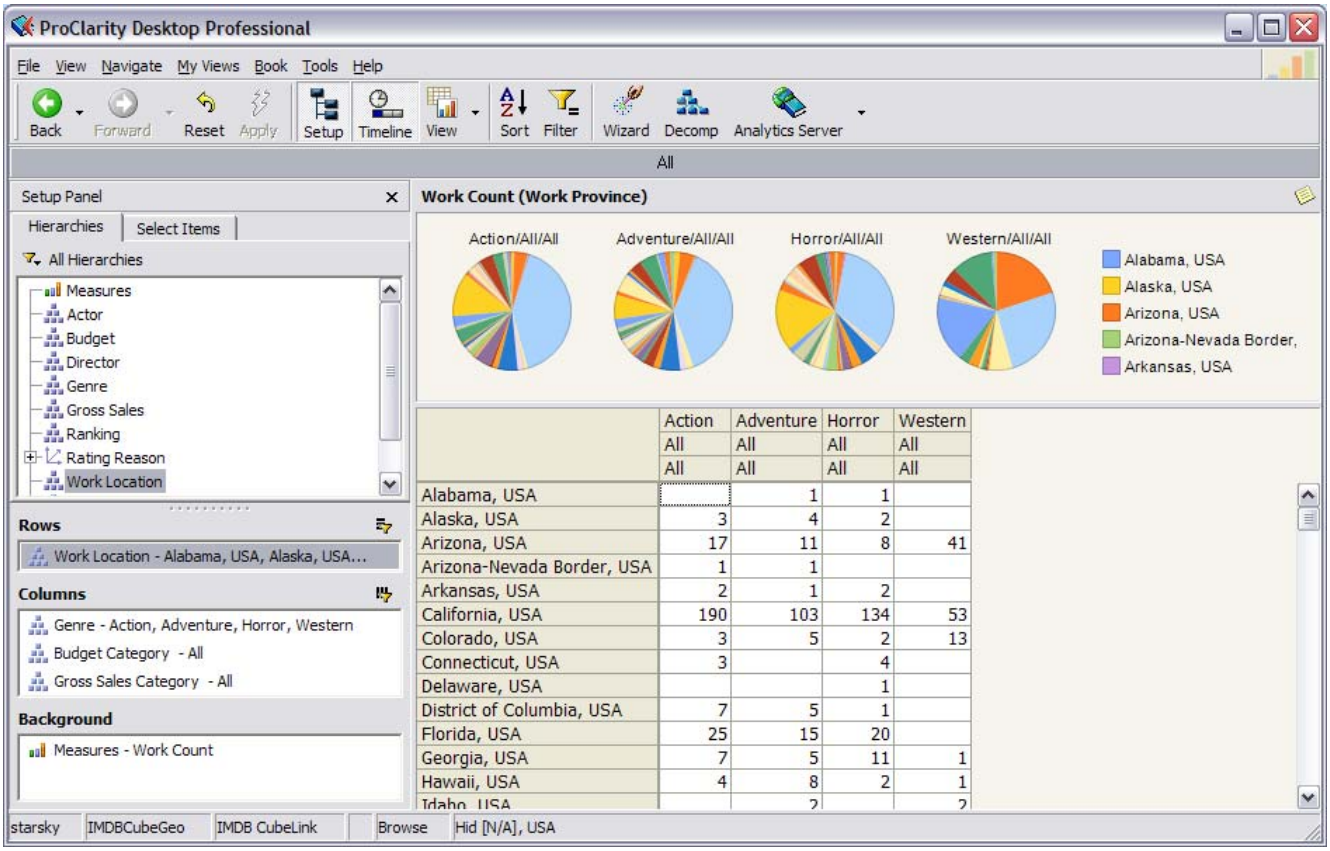
II. SIGNIFICANCE

Efforts to combine GIS with relational databases have long been established as a *de facto* standard. However from a user interface perspective very few ad-hoc interfaces offered to integrate the two technologies. ESRI provides ad-hoc query support in client tools but requires knowledge of query syntax, and are designed primarily to support simple queries. The IBM Spatial Extender system [3] integrated relational database queries and proximity related searches, but it is limited to DB2 and Data Warehouse Edition.

To provide users with rapid access to multiple summary, aggregated views of a large database, a relational technology OLAP [2] was developed. OLAP views databases through a collection of major foci called dimensions, organized as hierarchical collections of members. Constructing the Cartesian product of dimensions allows aggregation of quantities, called measures, over any collections of records projected through any subset of dimensions. These OLAP "data cubes" give flexible, visual reporting interfaces that are easily and rapidly driven by users.

The motivation for OLAP arose in the financial arena, where measures are primarily numerical quantities such as money. But in our law enforcement applications, measures are typically the numbers of entities (people, organizations, and events) having certain characteristics. Thus our approach rests on a specialized OLAP architecture we are calling count cubes built around count star schema [1].

The SOVAT system [4] focuses on a complete OLAP-GIS integration. This was later termed SOLAP (Spatial OLAP) used to hide the heterogeneity of each technology for an enhanced knowledge discovery process for the user.



	Action	Adventure	Horror	Western
	All	All	All	All
	All	All	All	All
Alabama, USA		1	1	
Alaska, USA	3	4	2	
Arizona, USA	17	11	8	41
Arizona-Nevada Border, USA	1	1		
Arkansas, USA	2	1	2	
California, USA	190	103	134	53
Colorado, USA	3	5	2	13
Connecticut, USA	3		4	
Delaware, USA			1	
District of Columbia, USA	7	5	1	
Florida, USA	25	15	20	
Georgia, USA	7	5	11	1
Hawaii, USA	4	8	2	1
Idaho, USA		2		2

Fig. 2. Example IMDB Work Location Analysis in Proclarity

GDDAI treats OLAP and geospatial analysis as components in an analytical workflow. GDDAI allows users to characterize their questions at a summary level using client-side business intelligence tools, such as ProClarity Desktop Professional, and then conduct their proximity searches on the detailed record(s) in the expanded GIS data source.

III. MOTIVATION

To illustrate knowledge discovery through the integration of OLAP and geospatial proximity searches we will describe a motivating use case and then later tie this use case into the Operations section (V).

In this fictitious use case we're interested in planning a summer vacation where we visit action, comedy, and horror major motion movie locations filmed in the U.S. between 1970 and 1985. We are only interested in films that grossed $1M and that were rated PG and R. Since we're visiting the area we're curious about the other nearby attractions, and since it is a vacation, we need to know preferable lodging, restaurants, coffee shops, and local attractions such as ghost towns or covered bridges. For our custom database resource we will rely upon the Internet Movie Database (IMDB[1]) and for our proximity searches use publically available point of interest databases.

[1] http://imdb.com

IV. TECHNICAL APPROACH

The goal of GDDAI is to give users an easily navigated, multi-faceted analysis and reporting environment. Our vision for GDDAI is an integrated environment providing seamless interconnectivity among Data Cube Analysis, Link Analysis [1], GIS Analysis, Statistical Analysis [5], and Web Reporting. From a data cube it is possible to "drill through" into the GIS Analysis tool, the Link Analysis tool or a web report.

GDDAI Operational components fall into two broad categories:

- **Summary Views:** OLAP data cubes and statistical analyses provide summary, aggregated views over large collections of data. These tools enable trend analysis and quick identification of anomalous information across the entire data set.
- **Detailed Views:** Database drill-throughs, geospatial queries, and link analyses provide views of particular data elements of interest and their relations. These are generically referred to as "drill-through" capabilities, in that they are typically invoked from within a summary OLAP view.

Last year GDDAI presented a paper [1] on the integrated data cube analysis and link analysis component called CubeLink, so we will not discuss that further. The remainder

of this paper will focus on the linkages between data cube analysis, drill-through, and geospatial analysis.

A. Data Cube Analysis

Data cubes are an intuitive and sophisticated representation of multi-dimensional structured information within the OLAP database paradigm. This technology is complementary to both relational and graph-based representations, and provides a "third view" of the data supporting summary queries, including trend analysis and statistical analysis. GDDAI is working closely with DHS customers to develop and deploy data cubes customized to specific investigative areas using commercial tools like Microsoft SQL Server Analysis Services (SSAS), Oracle, and the ProClarity OLAP client tool. GDDAI is also developing new methods to facilitate navigation in OLAP spaces and connections to the link analysis environment. Summary data cube views are at the center of our operational deployments, and will serve users as an additional front-end for querying data.

SSAS is used for GDDAI's cube engine. Just like a relational database, SSAS has data structures defined by a developer, a language for querying and refreshing the data and facilities for administration of storage and memory settings. With SSAS Multidimensional Online Analytical Processing (MOLAP) storage setting, several summaries of the data are made in the internal data structures and all data is stored in a MOLAP optimized system, leaving the original data sources and staging area unaffected. SSAS utilizes a single server process to run and maintain the cube engine.

ProClarity Desktop Professional is a 32 bit "rich" client application that is available for Windows users to query the SSAS cubes. ProClarity also has a server component which can be configured as either a web thin client or a web fat client (ActiveX) for an additional cost. Within GDDAI, we leverage the "rich" client application because it provides an exposed API allowing the ProClarity to be extensible.

The design of data cubes closely mirrors the count star schemas constructed in the integration database with some minor variations. SSAS extracts data from source databases to construct cubes. The cubes can then be made available for ProClarity (cube) users.

B. Relational Database Drill-through

From a cube view (and, in the future, from other GDDAI operational components listed below), there is the ability to "drill-through" the current summary view to show the identified information within traditional detailed report listings and existing web interfaces. This provides users the ability to see detailed data records in their traditional views, and retains the connection to the underlying objects.

We extended the ProClarity Desktop Professional client application to support drill-through to the underlying data. This capability is the basic building block for the detailed view components. The ProClarity desktop client has been extended using Microsoft .NET libraries to provide additional options on the context-sensitive "right-click" menu. These options rely on a common library we have developed to extract the current data cube context. Cube context is simply defined as the ad-hoc selection criteria defined by the user. ProClarity provides the ability for users to visually construct their query and display the query results using charts and a table of the summary count based on selected dimensions, hierarchies, and background filters. In addition to this context, users are able to highlight specific table cells displaying the summary count to "drill-through" the summary into an XML detailed record listing. Depending on the specific operation performed, the XML can be stored as a file on the user's system, or sent to a web server through an HTTP command, or sent directly to another application via an API or web-service interface. For many of our drill-through functions, the XML is sent to a web server, parsed, and translated into a relational database query against the star schema underlying the cube. This query is used to retrieve the unique identifiers for the objects selected in the cells in the data cube. A second query is then issued against the data staging area to return all of the staged attributes relevant to those identifiers. In some cases, the traditional star schema may be extended into a "super" star schema to provide information that is not included in the cube dimensions (e.g. entity names or latitude and longitude coordinates). This information may be further augmented by performing queries directly against the original data source, although that is not typically required. The resulting records are returned to ProClarity, where they can be rendered as HTML for display within a web browser, or sent to another application, such as a geospatial analysis application like ESRI's ArcMap. Our specific implementation is described below in a subsequent section.

C. Geocoding

Geocoding services provide the connective glue between the user's database and geospatial analysis. Geocoding translates user address information (street, city, state or province, country, zip or postal code) into coordinate points, which are typically latitude and longitude. Each geocoding vendor has their own algorithms, and users typically have options to select the level of imprecision they are comfortable with. To geocode an address, the geocoder compares each user-supplied address with an address locator that designates coordinates for addresses. The geocoder provides the best guess for a coordinate point, as well as a score to indicate the confidence of the result. The confidence score may vary due to spelling inconsistencies, or to ambiguity in some of the address components.

There are many commercially available address locators that users can rely upon. We rely on the ESRI provided geocoder and address locators. We configure our geocoding services to run as a batch process periodically to refresh the list of coordinate points for a given set of addresses in our

database. These coordinate points are stored as part of the "super" star schema in our relational database tables.

D. Base Map

A base map in GIS is used to help orient users to a particular region. Base maps can contain vector data sets (points, lines, polygons, symbology, and labels) and imagery that best help give the user perspective. For the purposes of this demonstration our base map is supplied by *ESRI StreetMap USA 2007* which supplies a variety of map features. The base map is served through a web service on the ArcGIS Server to make it easily accessible to the GDDAI clients.

E. Connecting ProClarity with ArcMap

As described above, users can access a detailed view from the summary view in ProClarity. We added a "Send to GIS" option to ProClarity's right-click menu. Once a user selects this menu option, an extended ESRI ArcMap thick client application is launched which is configured to be "drill-through aware". ArcMap automatically detects the new drill-through data, causing the deposited drill-through results to be stored as a persistent map layer in the user's workspace. We implemented this through a combination of ArcMap's Microsoft .NET APIs and ESRI's python APIs.

F. Basic Proximity Searches

Once the data from ProClarity comes into ArcMap (figure 4), a wizard is automatically launched prompting the user for their search criteria. As part of the search criteria, users can select one or more map layers of particular interest that they want to include in their proximity searches. For demonstration purposes we chose a variety of Point of Interest databases publically available on the web, and data provided in ESRI's Street Map USA collection. The search criteria were designed to be flexible to provide many options to the user's investigative purposes. At the conclusion of the search wizard, we launch the actual query process by invoking specific spatial analysis tools available within ArcMap. The traditional interface for these tools is a bit confusing, so our wizard simplifies the process for constructing meaningful results.

G. Reporting

Proximity searches provide different types of reports to show which drill-through results match the criteria. The primary display of the results is through an interactive map, which displays the matches as selected points on the map. In addition, for detailed feedback on the search, html is streamed to a browser so users can view or save PDF reports at a very low maintenance cost. A second type of persistent report of the search results is also provided which caches the search results as a new map layer that can be used later as part of a more complex search.

H. Complex Search

This search facility extends the basic proximity search capability by allowing users to change their currently selected drill-through result set by including/excluding selected drill-through from previous proximity search reports. This capability adds the power of combining Boolean logic with basic set operations like union and intersection to fulfil an analyst's more advanced *ad-hoc* query needs.

V. OPERATIONS

To help provide an operational perspective in a test environment we used the IMDB data source to build a cube focused around questions about movies, their respective budgets, genres and work locations. Using ProClarity, users can dynamically combine the IMDB dimensions to ask a variety of questions. Knowing the types of questions users want to ask drives the overall design of the cube. A portion of the cube star schema for the IMDB is shown in figure 3. The dimensions are:

- Movie – The entity dimension
- Genre – The categorical dimension describing the movie genre.
- Budget – A scalar dimension holding the movie budget in dollars binned within $1M intervals.
- Location – A hierarchical dimension with levels of city, province (state), and country.

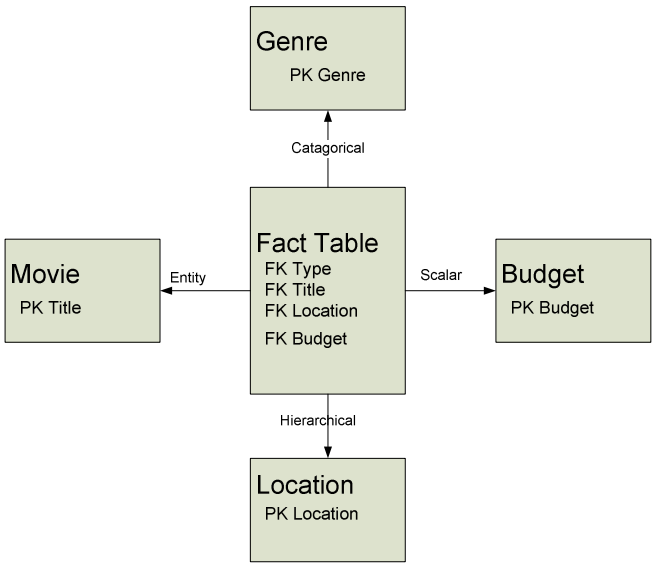


Fig. 3 A portion of the count star schema for the IMDB

By using ProClarity we can start to get the answers to the specific questions we posed such as: *What horror and action movies were filmed between 1970 and 1985 in the US?* ProClarity can return graphs or grids to represent the number of movies that fit certain criteria and categorical questions. While a relational database will provide thousands of results, ProClarity allows users to drill down into specific finetuned questions such as adding criteria like *film budgets* and *movie ratings*. To get the specific details we must use the Relational

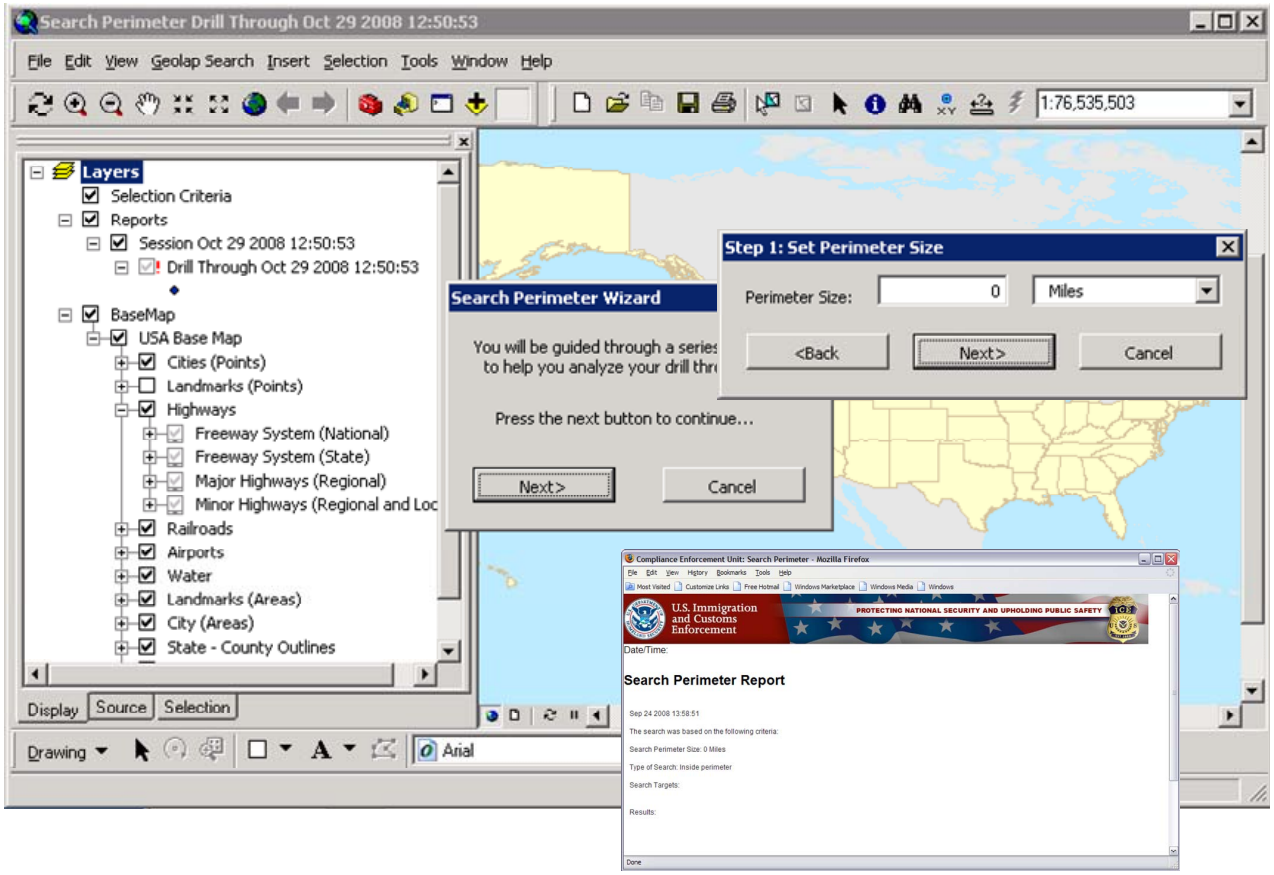


Fig. 4. Collage of Proximity Search Wizard and Reports

Drill-through capability to get the records about these movies from the source IMDB database, and to perform the geospatial searches on our points of interest. To perform the geospatial to latitude and longitude coordinates, which we can dynamically join to the Relational Drill-through results.

In ProClarity users can access these detailed results by right clicking on the cell of interest and selecting the "Send to GIS" menu option. When a user initiates the request, ProClarity generates its context XML, then sends this XML to a web server. The web server takes the XML and generates a relational database drill-through query, then returns the results of this query to ProClarity. The columns in this query include latitude and longitude coordinates, along with other database columns of interest. Since our geospatial analysis is focused around filming locations, each record returned will be focused around specific filming work locations. ProClarity takes these results and saves them to the user's hard disk in a configured location, where ArcMap can find them.

Some interesting points can be observed here. There is not a one-to-one relationship between movies and their filming locations. Some movies were filmed in more than one location, and *vice versa* locations can appear in more than one film. In ProClarity, this movie exists as one entity being counted, but when this movie is included in a result set sent to ArcMap, there will be two records sent. This causes the number of records sent to ArcMap to be higher than the sum of the selected cells in ProClarity. But some movie locations cannot be geocoded successfully, maybe due to a data entry error, or a mismatch between the filming location and our geocoder. These records are omitted from the result set sent to ArcMap (we give the user the option to view these records in a web browser report, but since there is no longitude or latitude coordinates, they cannot be rendered within ArcMap). These kinds of locations can cause the number of records sent to ArcMap to be reduced. It is rare that the number of records sent to ArcMap equals the sum of the selected cells in ProClarity.

Once the drill-through has created a new map layer the user can begin to ask the second part of their question: *Which movies were filmed within 20 miles of a ghost town or a covered bridge?* To answer this easily, a search wizard is initiated to lead the user through a series of questions to perform a proximity search. The proximity search assumes that you want to ask questions in the context of things being in the proximity of your drill-through (e.g. film work locations) being your search source. Based on the search results those drill-through map features matching your search criteria will be selected and visibly highlighted.

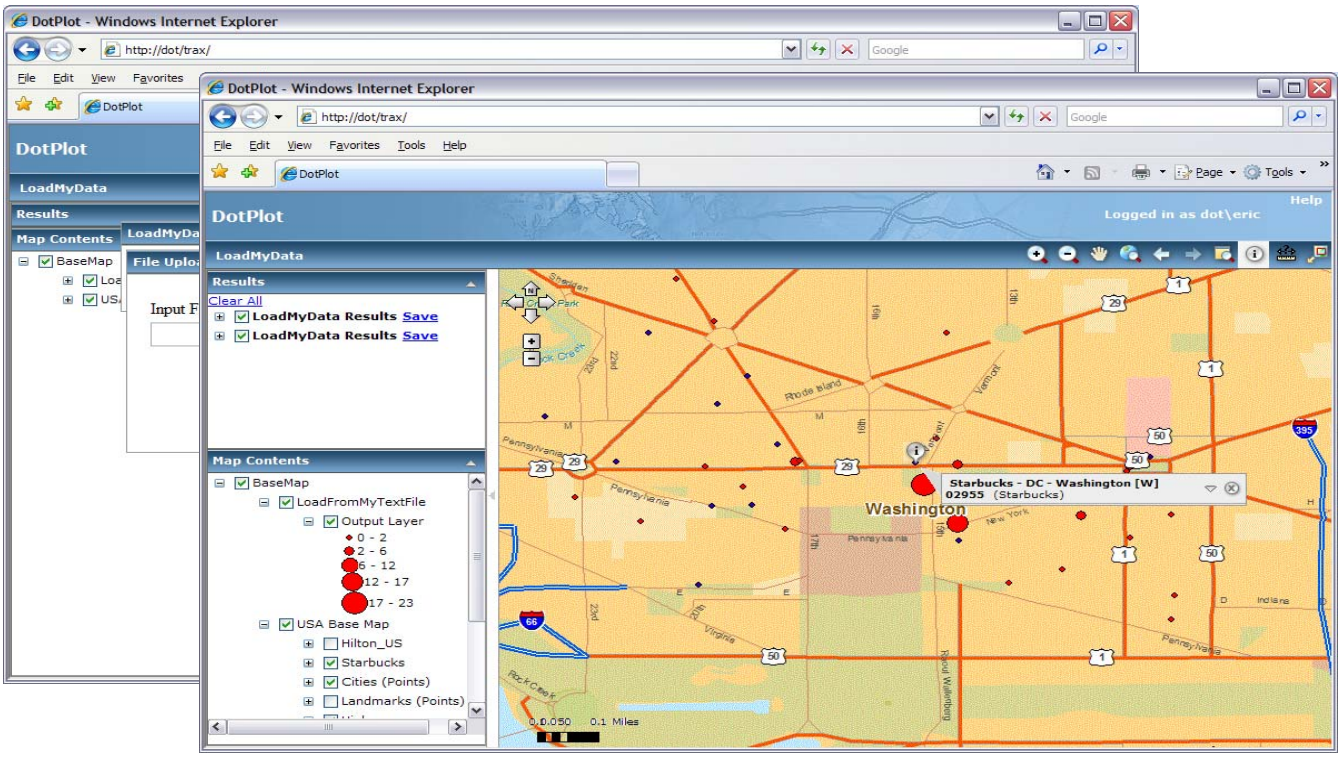


Fig. 5. Drill-through import script reused to become an ArcGIS Server web application

The search criteria are related to:

Proximity distance – users are able to select the circular buffer surrounding the target and specify the unit of measurement.

Target map feature(s) – (e.g. ghost towns, covered bridges, Starbucks or Peets coffee shops) select one or more map features that may be near or far to the drill-through source.

Near/Far context – Does the user want to determine if the source is within or outside the proximity buffer. For instance the user may want to create a personal black list of all film locations that aren't in the vicinity of any of the points of interest.

Once all the criteria have been given to the search wizard the application finds all the matching drill through and highlights them for you. At this point users are able to leverage any available geospatial analysis tool through ArcMap to perform a more detailed analysis such as what distance is the ghost town from my hotel or nearest roadway or we can ask the more complex questions such as *where are all the other action and horror films?*

VI. Discussion

There were many lessons learned providing geospatial analysis in GDAAI. The first crucial lesson was the steep learning curve one must overcome to provide a comprehensive geospatial analysis solution to clients. While we are greatly encouraged by the painstaking strides taken by geospatial providers such as ESRI to advance technical approaches and capabilities, there will always be an initial high price of admission for providing complete and integrated solutions. We feel quite strongly that end to end solutions such as GDDAI are crucial to enabling deeply integrated analytical solutions within Homeland Security.

We chose ArcMap initially for building and deploying maps, and testing embedded analysis tools. When considering thick "rich" client versus "thin" web client based solutions it is important to realize that for geospatial analysis it is not entirely seamless and users who rely on file-based geodatabases, ArcSDE, and web based services need to take great care studying what combinations can be used for analysis. For instance while we did rely upon web services to provide our base map, all geospatial analysis was performed using file-based geodatabases on network file shares or local disk space We found it was important to find a solution where all of the data used in the proximity searches were compatible.

Geocoding also represented an extremely high learning curve for our initial implementation. It relies on the address completeness, quality of the user data, and knowledge of what geolocator databases will provide the best results. In the IMDB work location dimension we found even though country, province, and city were our addressable information that many inconsistencies existed in the data source. For instance the movie "Meatballs" was erroneously reported that it was filmed in the city "Ontario" and country Canada with no province. Record mismatches such as these require hand curation or sophisticated algorithms to help correct data entry mistakes.

We heavily leveraged geoprocessing scripts as a facility to perform our geospatial analysis, this is allowing us to leverage the ArcGIS server's versatility to transform scripts into client-side tools, and client-side tools into web-based services. To prove the benefits of this approach we migrated our drill-through script with slight modifications to create an application (figure 5) that imports a user spreadsheet of point data and depicts concentrations of target map features in the vicinity of the imported spreadsheet points. This was achieved by merely using the ArcGIS Server application wizard. In ArcGIS server 9.3 we hope to also leverage Web 2.0 technology such as JSON and REST APIs to more effectively build web applications.

VII. CONCLUSION

Within the Department of Homeland Security (DHS) law enforcement there is a fundamental need for geospatial analysis of their custom data sources. The focus of GDDAI has been to provide an approach that ties geospatial analysis with other analytical solutions such as data cubes. Using this integrated approach allows users to intuitively ask complex ad-hoc questions, and perform proximity searches based on a greatly reduced set of identified locations saving greatly on time and effort. In the near future we plan on expanding the GDDAI geospatial analytical solution to develop ties between our geospatial and link analysis tools. This will provide users with another seamless analytical workflow within the GDDAI environment.

We also envision that GDDAI's end-to-end solution will one day provide a general platform to practically integrate new types of geospatial analysis coming from industry and the research community where these capabilities can be infused into supporting law enforcement units providing new innovations to support today's challenges.

VIII. ACKNOWLEDGMENT

Our thanks to all of our colleagues within DHS Immigration and Customs Enforcement (ICE). We would also like to acknowledge funding for this project from the DHS Science and Technology Directorate.

REFERENCES

[1] Joslyn, C., D. Gillen, et al. (2008). "Hybrid Multidimensional Relational and Link Analytical Knowledge Discovery for Law Enforcement." 161-166. Technologies for Homeland Security, 2008 IEEE Conference on.

[2] Chaudhuri, Surajit and Umeshwar Dayal: (1997) "An overview of data warehousing and OLAP technology", *ACM SIGMOD Record*, v. pp. 65 - 74

[3] Davis, J. (1998). "IBM's DB2 Spatial Extender: Managing Geo-Spatial Information Within The DBMS." IBM corporation May.

[4] Scotch, M. and B. Parmanto (2005). SOVAT: Spatial OLAP Visualization and Analysis Tool.

[5] CA Joslyn, JS Burke, T Critchlow, N Hengartner, E Hogan: (2009) ``View Discovery in OLAP Databases through Statistical Combinatorial Optimization", submitted to the 2009 Conf. on Statistical and Scientific Database Management (SSDBM 09).