Open source geospatial Business Intelligence (BI) in action!

OGRS 2009 – Nantes, France

Dr. Thierry Badard
Etienne Dubé
Belko Diallo
Jean Mathieu
Mamadou Ouattara

GeoSOA research group
Centre for Research in Geomatics
Laval University, Quebec City, Canada

http://geosoa.scg.ulaval.ca
1. BI for dummies or a short BI 101 course
   - With BI, you enter a different world … Please, forget what you know about classical databases! ;-) 
   - Focus on some open source BI tools

2. Merging BI and GIS (Geospatial)?

3. An open source software stack for Geospatial BI
   - GeoKettle
   - GeoMondrian
   - Spatialytics

4. Outlooks
   - Integration projects
What is BI (Business Intelligence)?

- Wikipedia states “Business intelligence (BI) is a business management term, which refers to applications and technologies that are used to gather, provide access to, and analyze data and information about company operations.”

- Something your boss or client is possibly interested into, and asked you to investigate ...

- Rely on an architecture with complex components and applications:
  - Data warehousing
  - On-line Analytical Processing (OLAP) servers and clients
  - Reporting tools
  - Dashboards
  - Data mining

© 2005, United Feature Syndicate
Classical architecture of a BI infrastructure

* ETL stands for Extract, Transform (integration, data cleansing, data structure, “updating”, …) and Load

Data sources (OLTP systems)
- Transactional databases
- Web resources
- XML, flat files, proprietary file formats (Excel spreadsheets, …)
- LDAP
- …

* ETL systems

Data loading

Data Warehouse

Reporting tools

OLAP

Data mining
The Data Warehouse: the crucial/central part!

- Repository of an organization’s historical data, for **analysis purposes**.
- Primarily destined to analysts and decision makers.
- Separate from operational (OLTP) systems (source data)
  - But often stored in relational DBMS: Oracle, MSSQL, PostgreSQL, MySQL, Ingres, …
- Contents are often presented in a summarized form (e.g. key performance indicators, dashboards, OLAP client applications, reports).
  - Need to define some metrics/measures
The Data Warehouse: the crucial/central part!

- Optimized for:
  - Large volumes of data (up to terabytes);
  - Fast response (<10 s) to analytical queries (vs. update speed for transactional DB):
    - de-normalized data schemas (e.g. star or snowflake schemas),
      - Introduces some redundancy to avoid time consuming JOIN queries
    - all data are stored in the DW across time (no corrections),
    - summary (aggregate) data at different levels of details and/or time scales,
    - (multi)dimensional modeling (a dimension per analysis axis).
      - All data are interrelated according to the analysis axes (OLAP datacube paradigm)
  - Focus is thus more on the analysis / correlation of large amount of data than on retrieving/updating a precise set of data!
MDX query language

- MDX stands for MultiDimensional eXpressions
- Multidimensional query language
- *De facto* standard from Microsoft for SQL Server OLAP Services (now Analysis Services)
- Also implemented by other OLAP servers (Essbase, Mondrian) and clients (Proclarity, Excel PivotTables, Cognos, JPivot, …)
- MDX is for OLAP data cubes what SQL is for relational databases
- Looks like a SQL query but relies on a different model (close to the one used in spreadsheets)

```sql
SELECT
  { [Measures].[Store Sales] } ON COLUMNS,
  { [Date].[2002], [Date].[2003] } ON ROWS
FROM Sales
WHERE ( [Store].[USA].[CA] )
```
Query:

```sql
SELECT
    { [Product].[All Products].[Drink],
      [Product].[All Products].[Food] } ON COLUMNS,
    { [Store].[All Stores].[USA].[WA].[Yakima].[Store 23],
      [Store].[All Stores].[USA].[CA].[Beverly Hills].[Store 6],
      [Store].[All Stores].[USA].[OR].[Portland].[Store 11] } ON ROWS
FROM Warehouse
WHERE ([Time].[1997], [Measures].[Units Shipped])
```

Crosstab:

<table>
<thead>
<tr>
<th>Time</th>
<th>[Product].[All Products].[Drink]</th>
<th>[Product].[All Products].[Food]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Measures].[1997]</td>
<td>1002.0</td>
<td>7531.0</td>
</tr>
<tr>
<td>[Store].[All Stores].[USA].[WA].[Yakima].[Store 23]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Store].[All Stores].[USA].[CA].[Beverly Hills].[Store 6]</td>
<td>1420.0</td>
<td>7617.0</td>
</tr>
<tr>
<td>[Store].[All Stores].[USA].[OR].[Portland].[Store 11]</td>
<td>767.0</td>
<td>5723.0</td>
</tr>
</tbody>
</table>

OLAP client software propose:

- Alternate representation modes (pie charts, diagrams, etc.)
- Different tools to refine queries/explore data
  - Drill down, roll up, pivot, …
  - Based on operators provided by MDX and on a complex logic implemented in the client part
Pentaho open source BI software stack

http://www.pentaho.org
Pentaho open source BI software stack

- Pentaho (http://www.pentaho.org)

- Kettle
- Mondrian
- Weka
- Pentaho Reporting

Data sources (OLTP systems) → Data extraction → Kettle (ETL systems) → Data loading → Data Warehouse

- CDF: Community Dashboard Framework
- Other projects: olap4j, JPivot, Halogen, …

Reporting tools

OLAP

Data mining
Pentaho open source BI software stack
Why merge BI and GIS software?

- Because …

“About eighty percent of all data stored in corporate databases has a spatial component” [Franklin 1992]

To support the geospatial component …

You can use:

- **GIS**
  - Implies the writing of very complex SQL queries
  - Sometimes, a long and hard job which requires dedicated human resources
  - Need to be done anew everytime data change or new analyses have to be achieved

- Classical BI tools (OLAP clients, reporting tools)
  - Unable to handle the spatial dimension of data (or only a very basic support)

- Merging GIS and BI tools (e.g. Spatial OLAP)
  - To fully exploit the spatial component
  - No need to write any SQL statements, just click away!
Components of a Spatial BI infrastructure?

- GIS file formats,
- Web Feature Services,
- Spatial DBMS
- Spatial ETL
- Spatial DBMS
- Spatial Business Intelligence
- Reporting tools
- OLAP
- Data mining
- SOLAP, Spatial data mining, Map-driven dashboards, …

Data sources (OLTP systems)

Data extraction ➔ Data loading

ETL systems

Data Warehouse

Require to consistently integrate the geospatial component in all parts of the architecture!
Open source geospatial BI software stack

- Pentaho ([http://www.pentaho.org](http://www.pentaho.org))

  + CDF: Community Dashboard Framework
  + Other projects: olap4j, JPivot, Halogen, …
Open source geospatial BI software stack

- **GeoSOA group** ([http://geosoa.scg.ulaval.ca](http://geosoa.scg.ulaval.ca))

- **GeoKettle**
- **GeoMondrian**

**Spatial ETL systems**

- **Weka**
- **Pentaho Reporting**

**Data sources** (OLTP systems)

**Data Warehouse**
- PostGIS
- Oracle Spatial

**Spatial Data mining**

+ **Spatialalytics**: new open source project and ongoing experiments with CDF & Jasper Server
GeoKettle is a "spatially-enabled" version of Pentaho Data Integration (Kettle)

True and consistent integration of the spatial component

- All steps provided by Kettle are able to deal with geospatial data types
- Some geospatial dedicated steps have been added

Current stable version: 3.2.0-20090609

Released under LGPL

http://www.geokettle.org
GeoKettle

Provides support for:

- Handling geometry data types (based on JTS)
- Accessing Geometry objects in JavaScript
- It allows the definition of custom transformation steps by the user (“Modified JavaScript Value” step)
- Topological predicates (Intersects, crosses, etc.)
- SRS definition and transformations
- Input / Output with some spatial DBMS
- Native support for Oracle, PostGIS and MySQL
- MS SQL Server 2008, Ingres and IBM DB2 can be used but it requires some tricks
- GIS file Input / Output : Shapefile (and GML soon)
GeoKettle

- Upcoming features:
  - Cartographic preview (*work in progress*)
  - Implementation of data matching and conflation steps in order to allow geometric data cleansing and comparison of geospatial datasets
  - Read/write support for other DBMS & GIS file formats
    - MapInfo (.tab or MIF/MID), KML, GeoJSON, GeoRSS, ESRI Geodatabase, ArcSDE
    - Native support for MS SQL Server 2008 and Ingres
    - WFS, Sensor Web (TML, SensorML, SOS, ...)
  - Implementation of a “Spatial analysis” step with a GUI
GeoMondrian

- GeoMondrian is a "spatially-enabled" version of Pentaho Analysis Services (Mondrian)
- GeoMondrian brings to the Mondrian OLAP server what PostGIS brings to the PostgreSQL DBMS
  - i.e. a consistent and powerful support for geospatial data.
- Released under the EPL
- [http://www.geo-mondrian.org](http://www.geo-mondrian.org)
GeoMondrian

- As far as we know, it is the first implementation of a true Spatial OLAP (SOLAP) Server
  - And it is an open source project! ;-)  
- Provides a consistent integration of spatial objects into the OLAP data cube structure
  - Instead of fetching them from an separate spatial DBMS, web service or a GIS file
- Implements a native Geometry data type
- Provides first spatial extensions to the MDX language
  - Add spatial analysis capabilities to the analytical queries
- At present, it only supports PostGIS datawarehouses
  - But other DBMS should be supported soon ...
Spatially enabled MDX

- Goal: bring to Mondrian and MDX what SQL spatial extensions do for relational DBMS (i.e. *Simple Features for SQL and implementations such as PostGIS*).

- Example query: filter spatial dimension members based on distance from a feature

  ```sql
  SELECT
      {[Measures].[Population]} on columns,
      Filter(
          {[Unite geographique].[Region economique].members},
          ST_Distance([Unitegeographique].CurrentMember.Properties("geom"),
                      [Unite geographique].[Province].[Ontario].Properties("geom")) < 2.0
      ) on rows
  FROM [Recensements]
  ```
Spatially enabled MDX

Many more possibilities:

- in-line geometry constructors (from WKT)
- member filters based on topological predicates (intersects, contains, within, …)
- spatial calculated members and measures (e.g. aggregates of spatial features, buffers)
- calculations based on scalar attributes derived from spatial features (area, length, distance, …)
Spatialytics

- Spatialytics is a lightweight cartographic component which enables navigation in geospatial (Spatial OLAP or SOLAP) data cubes, such as those handled by GeoMondrian.
- It aims to be integrated into existing dashboard frameworks in order to produce interactive geo-analytical dashboards.
- Such dashboards support the decision making process by including the geospatial dimension in the analysis of enterprise data.
- First version stems from a GSoC 2008 project performed under the umbrella of OSGeo.
- Released under BSD (client part) and EPL (server part).
- [http://www.spatialytics.org](http://www.spatialytics.org)
Spatialytics

- Is mainly based on OpenLayers and Dojo

- Allows:
  - the connection with a Spatial OLAP server such as GeoMondrian,
  - the navigation in geospatial data cubes,
  - and the cartographic representation of some measures as static or dynamic choropleth maps.

- More thematic capabilities will be added shortly!
  - A student currently funded by the Google Summer of Code 2009 program, under the umbrella of OSGeo, is currently working on this task.
GeoKettle, GeoMondrian and Spatialytics demos

- **GeoKettle**
  - A lab. has been held this week! ;-)
  - Contents will be available on the GeoSOA website shortly!

- **GeoMondrian/Spatialytics**
  - A basic online demo is available at:
    - [http://geosoa.scg.ulaval.ca/Spatialytics/](http://geosoa.scg.ulaval.ca/Spatialytics/)
  - It demonstrates the interaction with GeoMondrian and how the cartographic navigation in the geospatial datacube is performed.
  - Work in progress; more demos should be available soon …
Features in development:

- More map-driven OLAP navigation operators (drill by position, by member, roll-up to parent, etc.)
- Dimension member selection / navigation controls
- Legend display
- New thematic mapping styles:
  - Choropleth: quantiles, other statistical distributions
  - Graphics: proportional symbols, histograms, pie charts, ...
  - Styles for other geometry types (lines and points)
  - Some styles or combination of styles allowing representation of multiple members/measures on a single map feature
  - Multi maps: Maps for different periods of time
  - ...

Spatialytics – Integration projects

First experiments with JasperServer + iReport

- iReport is a graphical report designer for JasperReports
- Will provide a framework to produce highly customizable reports or static dashboards
- Displays the information in different ways: maps, charts and tables
- Allows synchronisation between the different representations when the user drills down or rolls up on the map or the charts or …

Integration into Pentaho CDF + GeoReport (Inova)

- Another student funded by the GSoC 2009 program under the umbrella of OSGeo

Other integration projects to come …
JasperServer + Spatialytics eye candy ;-)
Questions?

- Thanks for your attention!

Contact:

Dr. Thierry Badard
GeoSOA research group
Laval University, Quebec, Canada
Email: Thierry.Badard@scg.ulaval.ca
Web: http://geosoa.scg.ulaval.ca