GeoKettle: A powerful open source spatial ETL tool

FOSS4G 2010

Dr. Thierry Badard, CTO

Spatialytics inc.

Quebec, Canada

tbadard@spatialytics.com

Spatialytics

Genuine Geospatial Business Intelligence

Barcelona, Spain – Sept 9th, 2010

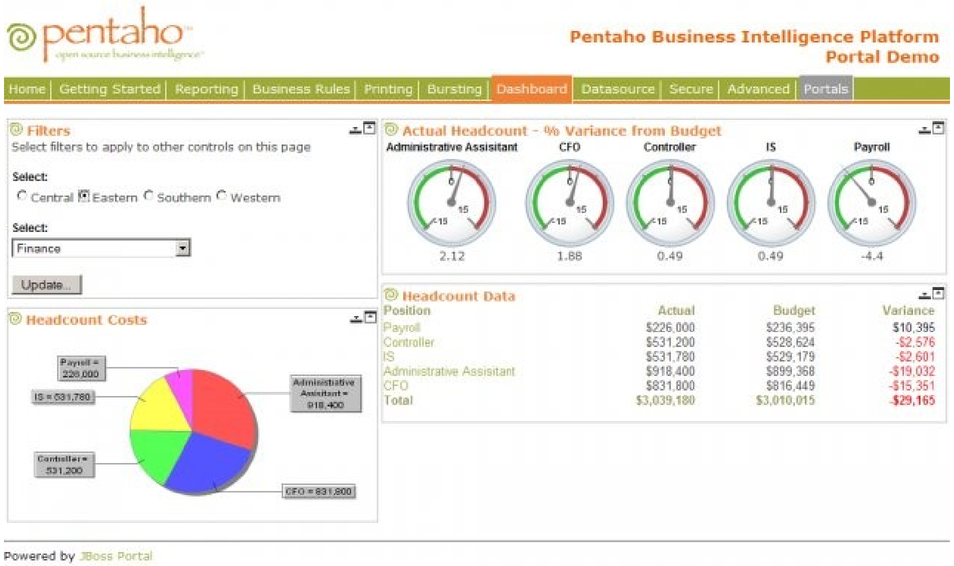
What is GeoKettle?

- It is part of the geospatial BI software stack developed initially by the GeoSOA research group at Laval University in Quebec …
  - GeoKettle — GeoKettle Spatialytics.org ETL Tool
  - GeoMondrian — GeoMondrian Spatialytics.org SOLAP Server
  - SOLAPLayers — SOLAPLayers Spatialytics.org Map Component
- But are now developed and supported by Spatialytics
  - http://www.spatialytics.org (open source community)
  - http://www.spatialytics.com (professional support, training)
- OK but … what is geospatial BI? ;-)

As you probably know …

- Business Intelligence applications are usually used to better understand historical, current and future aspects of business operations in a company.
- The applications typically offer ways to mine database- and spreadsheet-centric data, and produce graphical, table-based and other types of analytics regarding business operations.
- They support the decision process and allow to take more informed decision!

Data visualization to support decision …



pentaho
open source business intelligence
Pentaho Business Intelligence Platform
Portal Demo
Home | Getting Started | Reporting | Business Rules | Printing | Bursting | Dashboard | Datasource | Secure | Advanced | Portals

Filters

Select filters to apply to other controls on this page

Select:
○ Central ☒ Eastern ○ Southern ○ Western

Select:
Finance

Update...

Actual Headcount - % Variance from Budget

Administrative Assisitant	CFO	Controller	IS	Payroll
2.12	1.88	0.49	0.49	-4.4

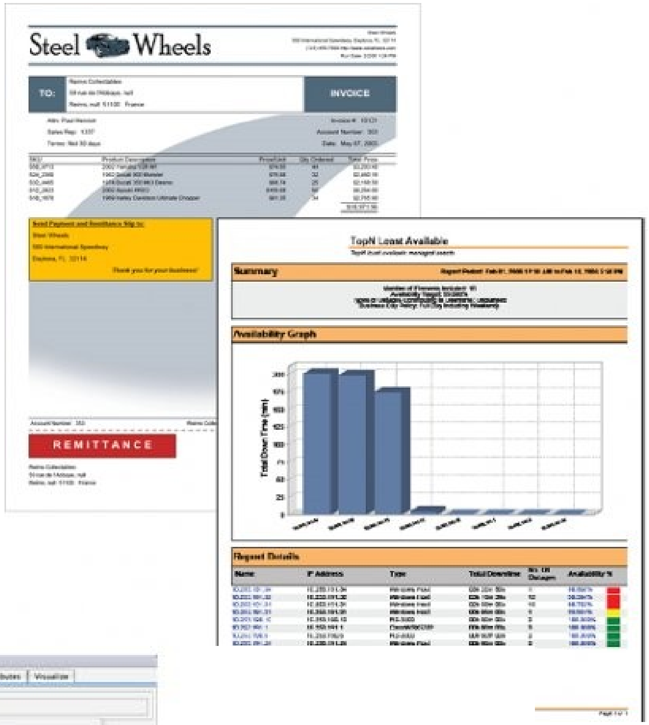
Headcount Costs

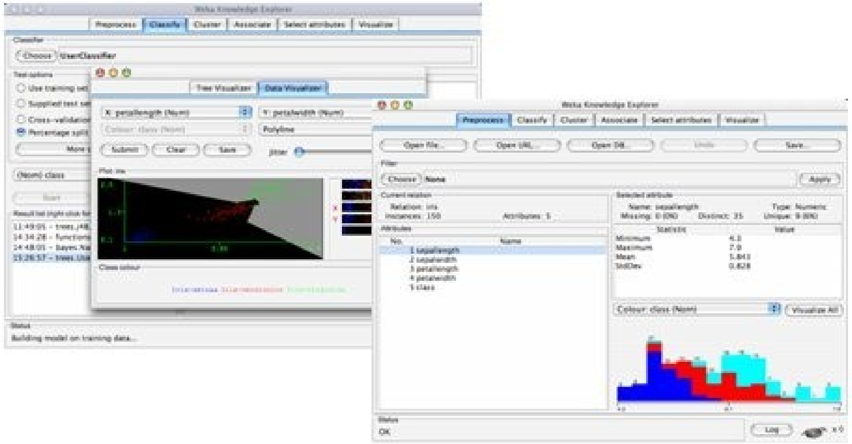
Payroll = 226,000
IS = 531,780
Controller = 531,200
Administrative Assistant = 918,400
CFO = 831,800

Headcount Data

Position	Actual	Budget	Variance
Payroll	$226,000	$236,395	$10,395
Controller	$531,200	$528,624	-$2,576
IS	$531,780	$529,179	-$2,601
Administrative Assisitant	$918,400	$899,368	-$19,032
CFO	$831,800	$816,449	-$15,351
Total	$3,039,180	$3,010,015	-$29,165

Powered by JBoss Portal





As you probably know …

- Business Intelligence applications are usually used to better understand historical, current and future aspects of business operations in a company.
- The applications typically offer ways to mine database- and spreadsheet-centric data, and produce graphical, table-based and other types of analytics regarding business operations.
- They support the decision process and allow to take more informed decision!
- Rely on an architecture with robust components and applications:
  - ETL tools & data warehousing (DW)
  - On-line Analytical Processing (OLAP) servers and clients
  - Reporting tools & dashboards
  - Data mining

So, an ETL tool is …

- A type of software used to populate databases or data warehouses from heterogeneous data sources.
- ETL stands for:
  - **Extract** – Extract data from data sources
  - **Transform** – Transformation of data in order to correct errors, make some data cleansing, change the data structure, make them compliant to defined standards, etc.
  - **Load** – Load transformed data into the target DBMS
- An ETL tool should manage the insertion of new data and the updating of existing data.
- Should be able to perform transformations from :
  - An OLTP system to another OLTP system
  - An OLTP system to an analytical data warehouse

Why use an ETL tool?

- Automation of complex and repetitive data processing without producing any specific code
- Conversion between various data formats
- Migration of data from a DBMS to another
- Data feeding into various DBMS
- Population of analytical data warehouses for decision support purposes
- etc.

GeoKettle

- GeoKettle is a "spatially-enabled" version of Pentaho Data Integration (Kettle)
- Kettle is a metadata-driven ETL with direct execution of transformations
  - No intermediate code generation!
- Support of several DBMS and file formats
  - DBMS support: MySQL, PostgreSQL, Oracle, DB2, MS SQL Server, ... (total of 37)
  - Read/write support of various data file formats: text, Excel, Access, DBF, XML, ...
- Numerous transformation steps
- Support of methods for the updating of DW

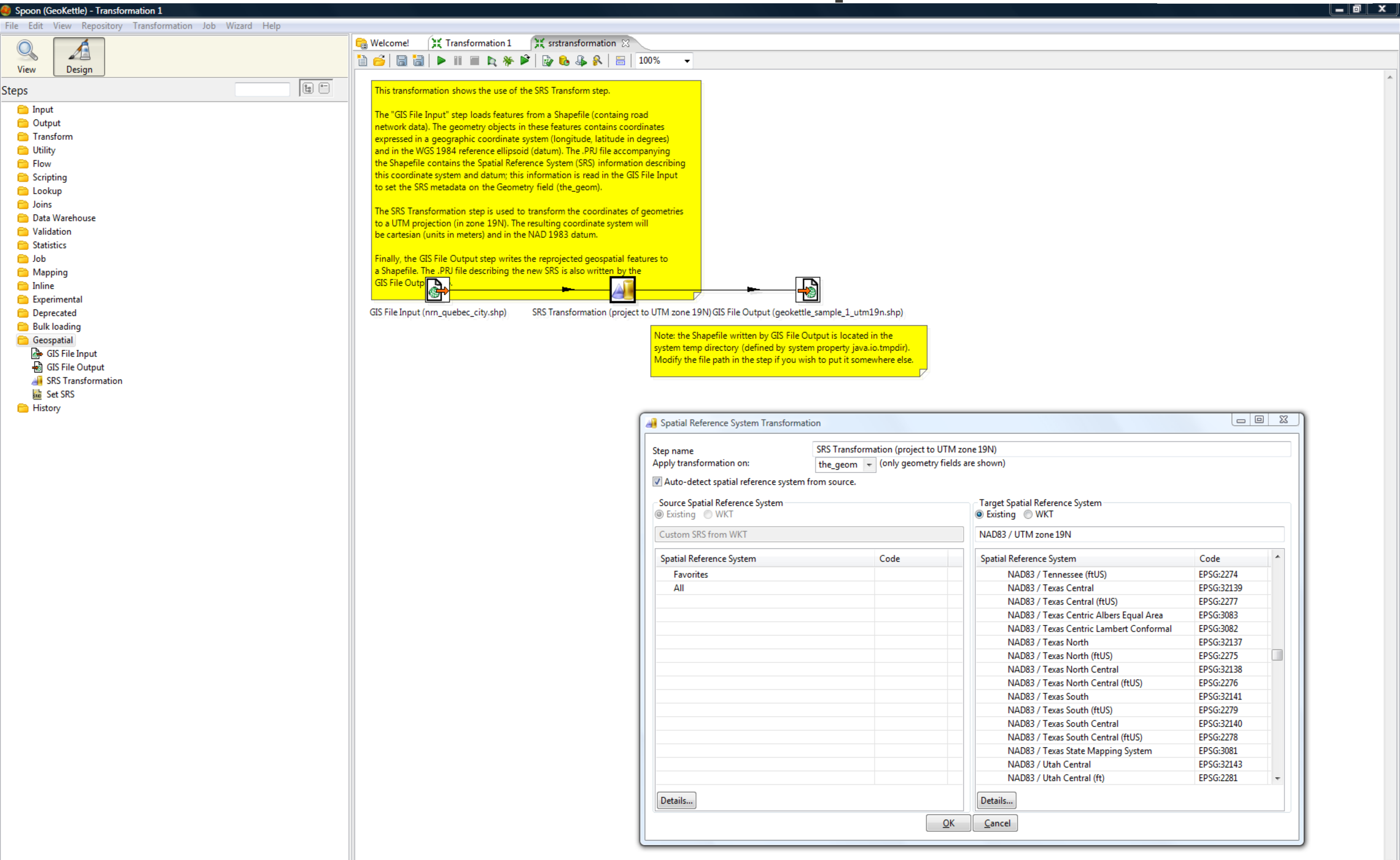
GeoKettle

- GeoKettle provides a 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
- First release in May 2008: 2.5.2-20080531
- Current stable version: 3.2.0-r188-20090706
- To be released shortly: GeoKettle 2.0 with many new features!
- Released under LGPL at http://www.geokettle.org
- Used in different organizations and countries:
  - Some ministries, bank, insurance, integrators, …
  - E.g. GeoETL from Inova is in fact GeoKettle! :-)
- A growing community of users and developers

GeoKettle

- Transformations vs. Jobs:
  - Running in parallel vs. running sequentially
- All can be stored in a central repository (database)
  - But each transformation or job could also be saved in a simple XML file!
- Offers different interfaces:
  - Spoon: GUI for the edition of transformations and jobs
  - Pan: command line interface for running transformations
  - Kitchen: command line interface for running jobs
  - Carte: Web service for the remote execution of transformations and jobs

GeoKettle - Spoon



This transformation shows the use of the SRS Transform step.

The "GIS File Input" step loads features from a Shapefile (containg road network data). The geometry objects in these features contains coordinates expressed in a geographic coordinate system (longitude, latitude in degrees) and in the WGS 1984 reference ellipsoid (datum). The .PRJ file accompanying the Shapefile contains the Spatial Reference System (SRS) information describing this coordinate system and datum; this information is read in the GIS File Input to set the SRS metadata on the Geometry field (the_geom).

The SRS Transformation step is used to transform the coordinates of geometries to a UTM projection (in zone 19N). The resulting coordinate system will be cartesian (units in meters) and in the NAD 1983 datum.

Finally, the GIS File Output step writes the reprojected geospatial features to a Shapefile. The .PRJ file describing the new SRS is also written by the GIS File Outp

Note: the Shapefile written by GIS File Output is located in the system temp directory (defined by system property java.io.tmpdir). Modify the file path in the step if you wish to put it somewhere else.

Step name: SRS Transformation (project to UTM zone 19N)

Apply transformation on: the_geom (only geometry fields are shown)

☑ Auto-detect spatial reference system from source.

Source Spatial Reference System

● Existing ○ WKT

Custom SRS from WKT

Spatial Reference System	Code
Favorites	
All	

Target Spatial Reference System

● Existing ○ WKT

NAD83 / UTM zone 19N

Spatial Reference System	Code
NAD83 / Tennessee (ftUS)	EPSG:2274
NAD83 / Texas Central	EPSG:32139
NAD83 / Texas Central (ftUS)	EPSG:2277
NAD83 / Texas Centric Albers Equal Area	EPSG:3083
NAD83 / Texas Centric Lambert Conformal	EPSG:3082
NAD83 / Texas North	EPSG:32137
NAD83 / Texas North (ftUS)	EPSG:2275
NAD83 / Texas North Central	EPSG:32138
NAD83 / Texas North Central (ftUS)	EPSG:2276
NAD83 / Texas South	EPSG:32141
NAD83 / Texas South (ftUS)	EPSG:2279
NAD83 / Texas South Central	EPSG:32140
NAD83 / Texas South Central (ftUS)	EPSG:2278
NAD83 / Texas State Mapping System	EPSG:3081
NAD83 / Utah Central	EPSG:32143
NAD83 / Utah Central (ft)	EPSG:2281

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.) and aggregation operators (envelope, union, geometry collection, ...)
  - SRS definition and transformations
  - Input / Output with some spatial DBMS
    - Native support for Oracle, PostGIS and MySQL
    - MS SQL Server 2008 and IBM DB2 can be used but it requires some tricks
  - GIS file Input / Output: Shapefile, GML 3, KML 2.2 and OGR support (~33 vector data formats and DBMS)
  - Cartographic preview

GeoKettle

- GeoKettle releases are aligned with the ones of Pentaho Data Integration (Kettle),
  - GeoKettle then benefits all new features provided by PDI (Kettle).
- Kettle is natively designed to be deployed in cluster and web service environments.
  - It makes GeoKettle a perfect software component to be deployed as a service (SaaS) in cloud computing environments as those provided by Amazon EC2.
  - It enables then the scalable, distributed and on demand processing of large and complex volumes of geospatial data in minutes for critical applications and without requiring a company to invest in an expensive IT infrastructure of servers, networks and software.

GeoKettle – Requirements and install

- Very simple installation procedure
- All you need is a Java Runtime Environment
  - Version 5 or higher
- Just unzip the binary archive of GeoKettle ...
- And let’s go !
  - Run spoon.sh (UNIX/Linux/Mac)
or spoon.bat (Windows)
- Need help, please visit our wiki:
  - http://wiki.spatialytics.org

GeoKettle

GeoKettle
Spatialytics.org ETL Tool

- Demo -

GeoKettle

GeoKettle
Spatialytics.org ETL Tool

- Upcoming features:
  - Implementation of data matching and conflation steps/jobs in order to allow geometric data cleansing and comparison of geospatial datasets (*results of a Google Summer of Code, should be available in version 2.x*)
  - Read/write support for other DBMS, GIS file formats and services
    - LAS (LiDAR), ...
    - Native support for MS SQL Server 2008, ...
    - WFS-T, Sensor Web (TML, SensorML, SOS, ...), ...
    - GIS metadata and CSW
  - Implementation of a “Spatial analysis” step with a GUI
  - Dedicated steps for social media (Twitter, ...), OSM, generalization, ...
  - Support of the third dimension
  - Raster support: development in progress of a plugin to integrate all capabilities provided by the Sextante library (BeETLe)

Questions?

- Thanks for your attention and do not hesitate to ask for more demos!
- Contact:

Dr. Thierry Badard, CTO
Spatialytics inc.
Quebec, Canada
Email: tbadard@spatialytics.com
Web: http://www.spatialytics.org
http://www.spatialytics.com
Twitter: tbadard & spatialytics

GeoKettle Spatialytics.org ETL Tool	http://www.geokettle.org	Twitter : geokettle
GeoMondrian Spatialytics.org SOLAP Server	http://www.geo-mondrian.org	Twitter : geomondrian
SOLAPLayers Spatialytics.org Map Component	http://www.solaplayers.org	Twitter : solaplayers