Spatial Data Integrator (SDI) powered by Open Source Spatial ETL
Agenda

- Camptocamp and Talend presentation
- Why data integration in the geospatial domain?
- Talend overview
- Spatial Data Integrator (SDI) powered by Talend
- Tutorial & sample jobs
- Conclusion & questions
Camptocamp, an Open Source Base Camp!

- 35 employees
  - Switzerland & France
- About 50 to 70 % of growth per year since 2002
- 3 activity domains
  - Spatial solutions
  - Business solutions
  - Infrastructure solutions
- 4 services poles
  - Consulting
  - Engineering
  - Supporting
  - Training
**Talend overview**

- Talend is the first provider of open source data integration software
- Located in France, USA, Germany, China
  - VC-funded
  - 50 employees
- First product release: 2006
- Leader in open source data integration
  - Rival large established proprietary players
What is ETL? Extract / Transform / Load

- ETL is a process in Data Warehousing. « How to get data in? » is ETL process name.
- **Extract**: extract data from source system where data originates.
- **Transform**: apply series of rules or functions to the extracted data (selecting, translating, encoding, deriving, joining, summarizing, splitting, ... more on [http://en.wikipedia.org/wiki/Extract,_transform,_load](http://en.wikipedia.org/wiki/Extract,_transform,_load))
- **Load**: once data transformed and cleaned, load the data in a data warehouse.
Why Spatial Data Integration

- Data integration is a key process
  - Data volumes in exponential growth
  - Diversity and heterogeneity of data sources
  - Data processing plays a major role in implementing GIS projects
  - Consolidating and aggregating spatial data with data from other sources is often required

- GIS data integration situation
  - Use command or hand-made script from various tools and libraries
    - gdal/ogr commands, fwtools, postgis command, ...
  - Proprietary Spatial ETL such as FME
  - Lack of Open Source global geo-spatial data integrator

- **Spatial Data Integrator**, *Powered by Talend* is now available!
- Prototyped summer 2007, Presented in FOSS4G2007
Talend makes data integration solutions available to organizations of all sizes, and for all data integration needs.
Talend Data integration

Synchronize and check integrity of your applications data

Migrate legacy applications

Sales

Accounting

ERP/CRM

Finance

Production

Budgeting

Extract, Transform and Load Data

EDWH

Datamart

Datamart

Replicate subset of data into subject matter DM

Exchange / share data with customers or suppliers

eCommerce
eExchange
Spatial Data integration

- Synchronize and check integrity of your applications data

- Extract, Transform and Load Data

- Geospatial Database

- Migrate legacy applications

- Parcels

- Roads

- eCommerce

- Govt agency

- Datamart
Spatial Data Integrator

- Spatial Data Integrator is one component of the SDI useful for ...
  - Data manipulation (Extraction, Quality checking, Conversion, Projection)
  - Data & metadata production (vector and Raster analysis)
  - Data & metadata manager (Network files and database manipulation, archiving)
  - Data dissemination (WWW publication, Deploy jobs as webservice)
  - Data reporting (Indicators, Analysis, ...)
- ... End user tools to define common tasks (ie. Job, process, script) usually made by hand or scripting in desktop GIS.
Spatial Data Integrator in SDI

SDI Software Architecture

- **Thick (Desktop) Clients**
  - (ArcGIS, uDig, QGIS, gvSIG, Google Earth)

- **Thin (Web) Clients**
  - (MapBuilder, InterMap, Google Maps)

**Interfaces**

- **Direct Access**
  - WMS (GIF, PNG, JPG, KML)

- **WFS**
  - (GML, Shapefile)

- **WCS**
  - (TIFF, GeoTIFF, ...)

- **CSW**
  - (DC, FGDC, ISO19115/19119/19139)

**Servers**

- **GeoServer**
  - (T-WFS, WMS)

- **MapServer**
  - (WMS, WFS, WCS)

- **GeoNetwork**
  - (CSW)

**Databases**

- **PostGIS DBMS**
  - (Vector data, both archive and for editing, Metadata)

- **File system**
  - (Vector data & satellite images)
The Talend offering

- **Talend Integration Suite - Enterprise Edition**
  - Grid Conductor
  - CPU Balancer

- **Talend Integration Suite - Professional Edition**
  - Distant Run
  - Job Conductor Advanced
  - Activity Monitoring Dashboard

- **Talend Integration Suite - Team Edition**
  - Shared Repository
  - Job Conductor
  - Activity Monitoring Console

- **Talend On Demand**
  - Hosted Repository

- **Talend Open Studio**
  - Business Modeler
  - Job Designer
  - Metadata Manager

- **Spatial Data Integrator powered by Talend**
  - Input/output spatial data
  - Complex and simple spatial components

- **SDI Advanced Suite**

- **Subscription**

- **GPL**
Talend Open Studio

- Key features
  - Business-oriented process modeling
  - Graphical development
  - Robust and scalable execution
  - Broadest connectivity to support all systems
  - Project repository for design and execution
  - Real-time debugging

- A high adoption rate
  - 100,000 product downloads
  - 20% register as users

- Active community
  - 1,000 beta testers
  - 500 forum contributors
Productivity & Ease of Use

- Graphical development
  - Dramatically increased productivity & ramp up
  - Combined graphical & technical views
  - Drag-and-drop mapping interface
  - Large library of components & connectors
- Leverage industry-standard languages
  - Java, Perl, SQL
Performance and robustness

- Highest performance, robust and scalable execution
  - Grid-distributed processing
  - Industry-standard code generated (Java or Perl)
  - Leverage both ETL and ELT architectures
  - Process data closest to the source
Versatility through Connectivity

- Broadest connectivity to support all systems
  - 100+ connectors available out of the box
- RDBMS:
  - Oracle, PostgreSQL, MySQL, DB2, SQL Server, Sybase, Ingres, …
- Web:
  - Web Services, FTP, HTTP, POP, SMTP…
- Files:
  - Delimited, positional, XML, Excel…
- Business Applications:
  - SugarCRM, SalesForce.com, LDAP…
Talend Project repository

- The Repository
- shared code
- non-technical graphical representation of a business requirement
- graphical representation of the technical process
- Context variables
- Metadata (stream definitions)
- documentation
- recycle bin
Job designer

- Job: components connected together
Spatial Data Integrator - SDI

- Talend Open Studio with geo-spatial extensions
- SDI integrates a new family of vector and raster geo components
- Based on reliable open source tools:
  - Java Topology Suite (JTS)
  - GeoTools
  - GRASS
Spatial Data Integrator Flow Architecture

- Uses GeoTools / Java Topology Suite (JTS) librairies

**Files**
- Text files
- GIS (ESRI, MapInfo)
- RASTER (Gdal)

**Databases**
- All (JDBC) Databases
- GIS Postgis
- Talend **RowGenerator** (build input using user criteria)
Spatial Data Integrator Architecture

- SDI Components architecture
Geospatial components

Feature manipulation
- Calculators
  - AreaCalculator
  - BufferCalculator
  - CentroidCalculator
  - CoordinateFetcher
  - DecimalDegreesCalculator
  - EnvelopeCalculator
  - LengthCalculator
- Collectors
  - BoundingBoxAccumulator
  - ConvexHullAccumulator
  - NeighborFinder
- Geometric Operators
  - Dissolver
- Manipulators
  - 2DPointReplacer

Vector format
- Database
  - PostgisInput
  - PostgisOutput
- File
  - GeoRSSOutput
  - MapInfoInput
  - MapInfoOutput
  - ShapefileInput
  - ShapefileOutput

Raster processing (experimental)
- Grass
  - Grass
  - GrassRContour
  - GrassRLegend
  - GrassRWhat
  - GrassVOutOgr

Metadata management
- Metadata
  - CataloguePublisher
    - ESRIArcCatalog_to_ISO19115
    - ISO19115_to_ISO19139FR
    - ISO19139FR_to_ISO19115
    - ISO19139_to_ISO19139FR
Getting started with Spatial Data Integrator

Open Source Spatial ETL
Data used in the tutorial

Monitoring stations and rivers in the french part of the Alpes (mainly Rhône river basin)
Convert Textfile to common GIS format

- Start Talend
- Create a workspace named SDI
- Copy tutorial datasets in:
  TALEND_HOME/workspace/sdi/data
Getting started with Spatial Data Integrator

Tutorial n°1: Convert Textfile to common GIS format
Convert Textfile to common GIS format

Objectives of this job is to produce ESRI Shapefile, Mapinfo file from a text file describing monitoring stations and their geographic location.

- Input:
  - CSV file with x,y and attributes columns for id and name of monitoring stations

- Output:
  - Shapefile and Mapinfo file
  - (optional) PostGIS table

- Process:
  - Create a point geometry using the x and y column of the text file
Convert Textfile to common GIS format

- First step is to create a new job.
  1. Start Talend SDI
  2. On the repository Tab, in Job design, click on create a new Job
- This will create a new pane where the job will be drawn.
Convert Textfile to common GIS format

- Create metadata about the current job. Talend is able to produce metadata and versionning about jobs.
- The Name is mandatory.
Convert Textfile to common GIS format

- Open the « Palette » tab
- Open the « File/Input » family section
- Add a tFileInputCSV component to the job

In the name of the component, the first letter « t » stands for Talend initial components, « s » for Spatial ones, « u » for Users ones.

If a panel could not be find in the Talend workspace (eg. « Palette »), click on menu « Window>Show view », and then search for the « Palette ».
Convert Textfile to common GIS format

- Select the « Properties » tab, and select the file « TALEND_HOME/workspace/sdi/data/stations.txt »

**tFileInputCSV component**

**Properties:**
- filename
- separator
- ...
- schema
Convert Textfile to common GIS format

- The stations text file is composed of 4 columns:
  - Id : text
  - Name : text
  - X : double
  - Y : double
- ... where coordinates are in WGS84.
- ... and text separator is ',' and decimal is '.'.
Convert Textfile to common GIS format

- On the properties tab of the tFileInputCSV component, click on « edit schema »
- Then add 4 fields
- Change name and type for each column.

Schema could be import & export when used frequently.
Convert Textfile to common GIS format

- (Optional) add a tLogRow component (in the log & error family)
- Connect the tFileInputCSV to the tLogRow (right click the component, select « row>main » and connect to the output component)
- Run the job (F6)
- ... tLogRow is useful for debugging!
Convert Textfile to common GIS format

- **Objectives:** Create a point from X and Y column.
- Add a s2DPointReplacer to the job.
- Connect the tFileInputCSV component
- Move to the properties of 2DPointReplacer
- Select column to use to create the point geometry
- Try tLogRow
Convert Textfile to common GIS format

- **Objectives:** Add output components.
- Add a sShapefileOutput to the job.
- Connect the s2DPointReplacer component
- Display the properties of sShapefileOutput
- Define the file name
- Run the job (F6)
- ... Add a MapInfo output
Convert Textfile to common GIS format

- Run the job (F6)
- Test some options:
  - Turn statistics on
  - Turn Traces on
- Try to open the layers produced in a GIS

```
[statistics] connecting to socket on port 3445
[statistics] connected
[statistics] disconnected
Job ConvertTextToGIS ended at 12:36 30/10/2007. [exit code=0]
```
Convert Textfile to common GIS format

- In tutorial n°1, user learned how-to:
  - Create a new job
  - Add components to a job
  - Link main flow between components
  - Run a job (using statistics, traces and tLogRow for debugging)
Getting started with Spatial Data Integrator

Open Source Spatial ETL

Tutorial n°2: Publish GeoRSS feeds to the web
Publish GeoRSS feeds to the web

Objectives of this job is to define a mapping between element coming from a GIS layer to a GeoRSS output to be published to the web.

- **Input:**
  - Use previous job flow

- **Output:**
  - GeoRSS output

- **Process:**
  - Create a new attribute named link http://hydro.eaufrance.fr/+CODE
  - Rename attribute name to title
  - (optional) Filter station where id start with 06
Publish GeoRSS feeds to the web

- Output GeoRSS feed.
- Geometry format could be simple georss or gml point.
- Attributes will be output using attribute's name (ie. to set the "title" element of the item, set the name of that attribute as "title"). To do so, use a tMap component.
- This GeoRss output could be used in OpenLayers (for those of you who are going to attend to the OpenLayers lab ! ;)}
Publish GeoRSS feeds to the web

- In the previous job, ...
- Add a tMap and a sGeoRssOutput component
- Link the component
- Define the properties of the output georss file (File name, channel description)
- Then open the tMap interface ...
Publish GeoRSS feeds to the web

- tMap interface:
  
  **Input (one or more)**

  **Output (one or more)**

  **Filter**

  eg. Id starting with « W »

  **Expression**

  eg. create an attribute named link composed of 2 strings: « http://hydro.eaufrance.fr/stations/ »

  concatenate with id
Publish GeoRSS feeds to the web
Publish GeoRSS feeds to the web

- Run the job
- Open the GeoRSS feed
- ... try this feed later with OpenLayers
- ... add a tFtpPut component to publish the file to a webserver
Publish GeoRSS feeds to the web

- In tutorial n°2, user learned how-to:
  - Define a mapping between input/output columns
  - Filter data using a tMap component
  - Create new field using expression
Getting started with Spatial Data Integrator

Open Source Spatial ETL

Sample jobs:
- Nearest Neighbour
- Dissolve geometry
- Metadata convert existing xml file to/from ISO 19115, 19139, ArcCatalogue
Find nearest river for each station

Objectives of this job is to find the nearest river for each monitoring stations

- **Input:**
  - Monitoring stations (format Shapefile)
  - Rivers (format Shapefile)

- **Output:**
  - Updated monitoring stations
Dissolver

Dissolve geometry based on an attribute

- **Input:**
  - Catchments

- **Output:**
  - Main catchements
Dissolver

Dissolve geometry based on an attribute

- Input:
  - Catchments

- Output:
  - Main catchments
BoundingBox and ConvexHull aggregator

Compute boundingBox and convexHull polygon for a layer

- **Input:**
  - Monitoring stations

- **Output:**
  - BoundingBox
  - ConvexHull polygon
Metadata batch conversion

Convert metadata file from one format to another (ISO19115, ISO19139, ArcCatalogue)

- **Input:**
  - One or more files (tFileList is used to iterate over files in a specific directory)

- **Output:**
  - XML files

Thanks GeoNetwork & GeoSource projects for the XSL styleSheets!
... and more using Geospatial components!

**Feature transformation**
- Calculators
  - sAreaCalculator
  - sBufferCalculator
  - sCentroidCalculator
  - sCoordinateFetcher
  - sDecimalDegreesCalculator
  - sEnvelopeCalculator
  - sLengthCalculator

- Collectors
  - sBoundingBoxAccumulator
  - sConvexHullAccumulator
  - sNeighborFinder

- Geometric Operators
  - sDissolve

- Manipulators
  - s2DPointReplacer

**Vector format**
- Database
  - sPostgisInput
  - sPostgisOutput

- File
  - sGeoRSSOutput
  - sMapInfoInput
  - sMapInfoOutput
  - sShapefileInput
  - sShapefileOutput

**Raster processing (experimental)**
- Grass
  - sGrass
  - sGrassRContour
  - sGrassRInGdal
  - sGrassRWhat
  - sGrassVOutOgr

**Metadata management**
- Metadata
  - sCataloguePublisher
    - sESRIArcCatalog_to_ISO19115
    - sISO19115_to_ISO19139FR
    - sISO19139_to_ISO19139FR

... and all others components!
In the community, user components are also available (eg. geolocalize)
What's up for the future?

Open Source Spatial ETL
What's up for the future? Raster components

- Raster components use GRASS tools
- GRASS components:

```
- Grass
  - sGrass
  - sGrassRContour
  - sGrassRInGdal
  - sGrassRWhat
  - sGrassVOutOgr
```

Diagram:

1. Set Grass env
2. Read shapefile and get first point coords
3. Get pixel (DEM) value at current location
4. Out & load
What's up for the future? Metadata

- **Objectives:** Establish a closer link between data & metadata during production step.
  - Quick metadata entry (title and abstract + automatic fields)
  - Do not create metadata after data creation / better improving metadata

- **New component to compute metadata during the job:**
  - User editor: title / abstract / purpose / category
  - SDI generate: Bbox / Number of objects / Geometry Type

- **Support for ISO & DCLITE4G standards**

- **Component to publish the metadata into an existing catalogue** (support only GeoNetwork catalogue)

- **Status:** Beta version
What's up for the future? Metadata

Start working on metadata during the creation of a new dataset.

- A GIS layer is describe at least by a title, an abstract and a bounding box. This is, in most case, enough to enable search in a catalogue and be able to know if that layer matchs users needs.
- In all SDI output component a form to create metadata is available
  - Using metadata templates as in GeoNetwork
  - Analysis the use of tag (like DATE) to have consistent metadata
What's up for the future? Metadata

Published metadata produced in a job in a catalogue (eg. GeoNetwork).

- Metadata publication steps:
  - Login to the GeoNetwork node
  - Select group and category
  - Publish the metadata
What's up for the future? Metadata

Create a shapefile and generate metadata.

zip, publish the Shapefile to a FTP and the metadata to a GeoNetwork node.
Spatial Data Integrator strengths

- Fast and efficient
- User-friendly Interface
- Easily customizable jobs (code generation)
- Benefits of « classical » ETL features
- Fully Open Source (GPL licence)
- Scalable
- High level of support by Camptocamp and Talend
Perspectives

- Development of new components:
  - Simple and complex components
  - New input and output formats
  - Community contribution very welcomed
- Spatial data viewer (uDig)
- Raster components optimization (Jgrass)
- Metadata components
- Integration of high-end Talend features:
  - Load balancing, Job conductor, Grid conductor
- Integration in Entreprise Service Bus (ESB) systems (PEtALS)
Spatial Data Integrator project

- Community infrastructure is being set up (mailing list, forum, wiki, download area, tutorial, ...).
- Register your interest to be informed:
  - http://www.camptocamp.com/sdi
Contacts

- Camptocamp:
  - François-Xavier Prunayre
    francois-xavier.prunayre@camptocamp.com
  - David Jonglez, david.jonglez@camptocamp.com
  - Claude Philipona, claude.philipona@camptocamp
  - http://www.camptocamp.com/sdi