Open Source Business Intelligence

Stefano Scamuzzo

Senior Technical Manager

Architecture & Consulting

Research & Innovation Division

Engineering Ingegneria Informatica

OW2 Europe Local Chapter

The Open Source Question

In many cases, the question is "when" to focus on open-source alternatives to traditional closed-source solutions, not "if" you should focus on them.

Gartner

Hype Cycle for Open-Source Software, 2005

The discovery of OSBI

Gartner

Research

Publication Date: 16 April 2008

ID Number: G00156326

Who's Who in Open-Source Business Intelligence

Andreas Bitterer

Many of the commercial business intelligence (BI) vendors have a long history and large marketing budgets, resulting in high visibility and mind share. Their lesser known open-source counterparts, such as Actuate BIRT, JasperSoft, Pentaho, or Spago have, however, started to gain traction in the market, beyond simple report writers for small shops and even larger enterprises are becoming aware of available open-source BI options.

Key Findings

- Open-source BI is here to stay.

OW2 Europe Local Chapter

Reasons to adopt OSBI

- According to Gartner Analysis
  - Reducing costs
  - Embed BI functionalities into existing applications
  - Complement the current BI infrastructure to extend BI usage to more users

- We should add to Gartner arguments …
  - Flexibility
  - Innovation
  - Better reactivity

The typical Business Intelligence layers

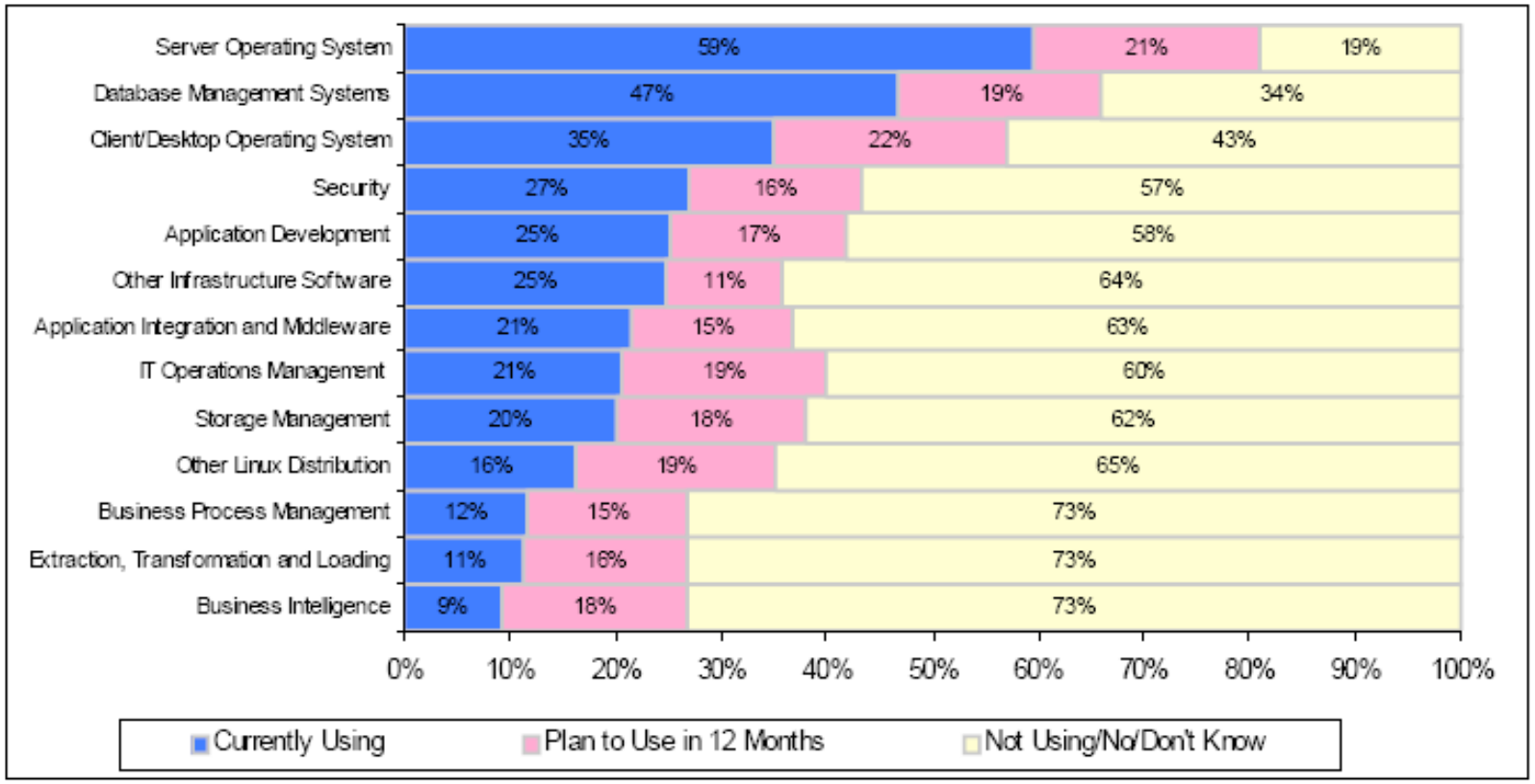
- Database Management Systems
- Data Ware House Platforms
- Extract Transfer Load (ETL) solutions
- Business Intelligence platforms:
  - Analytical tools
  - Document lifecycle management
  - Security
  - Integration

Database products

Open Source DBMS

- "From 2005 to 2006 open source vendor revenues grew 36,3% to $140 million… compared to the overall market growth of 12,2% … this growth will continue in the next five years at more than 40%" [Gartner]
- Gartner makes distinction between:
  - Mission critical vs non-mission critical
  - Supported by the community vs supported by a single vendor
- Four data management modes:
  - Content publishing
    - Write Once Read Many
  - Transactional
    - Highly normalized schema: ACID compliancy
  - Analytical
    - No transactions needed, but aggregation
  - Operational
    - Embedded (es. Smartphone)

Open Source DBMS and BI adoption

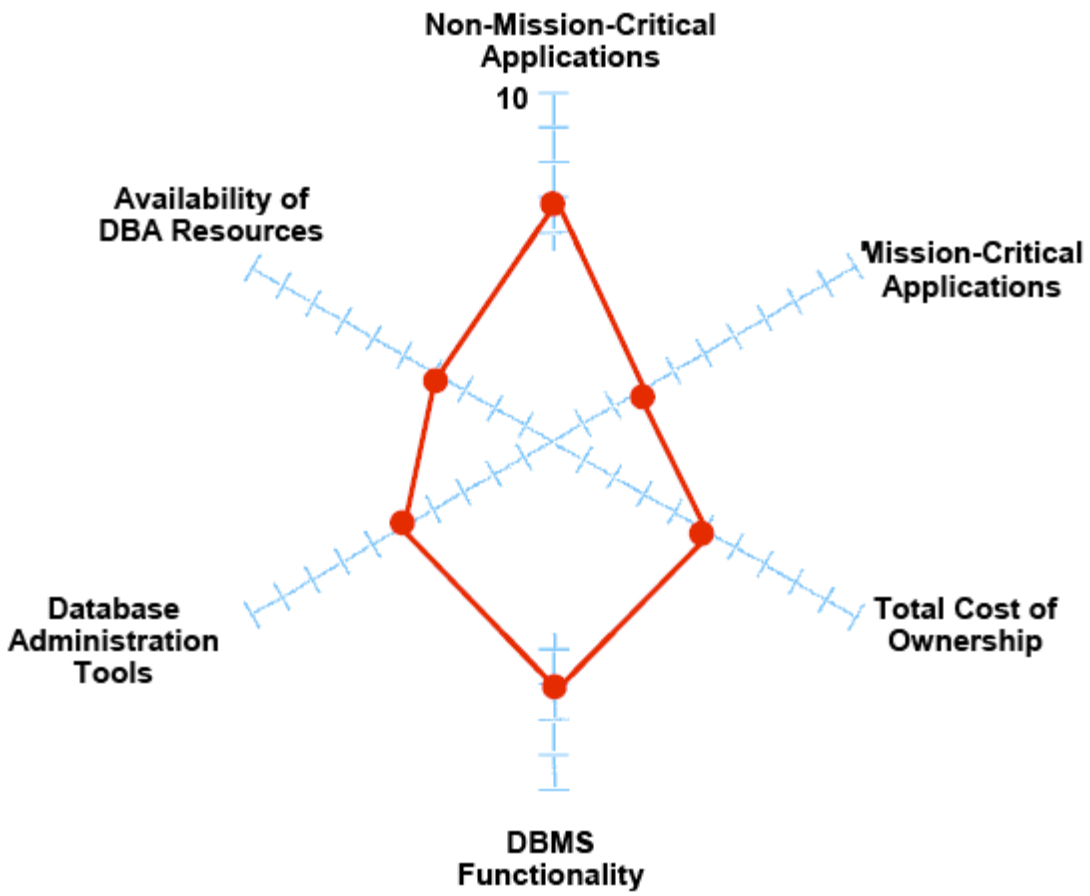


Source: Gartner (2007)

OW2 Europe Local Chapter

Maturity level of open source RDBMS (March 2008)

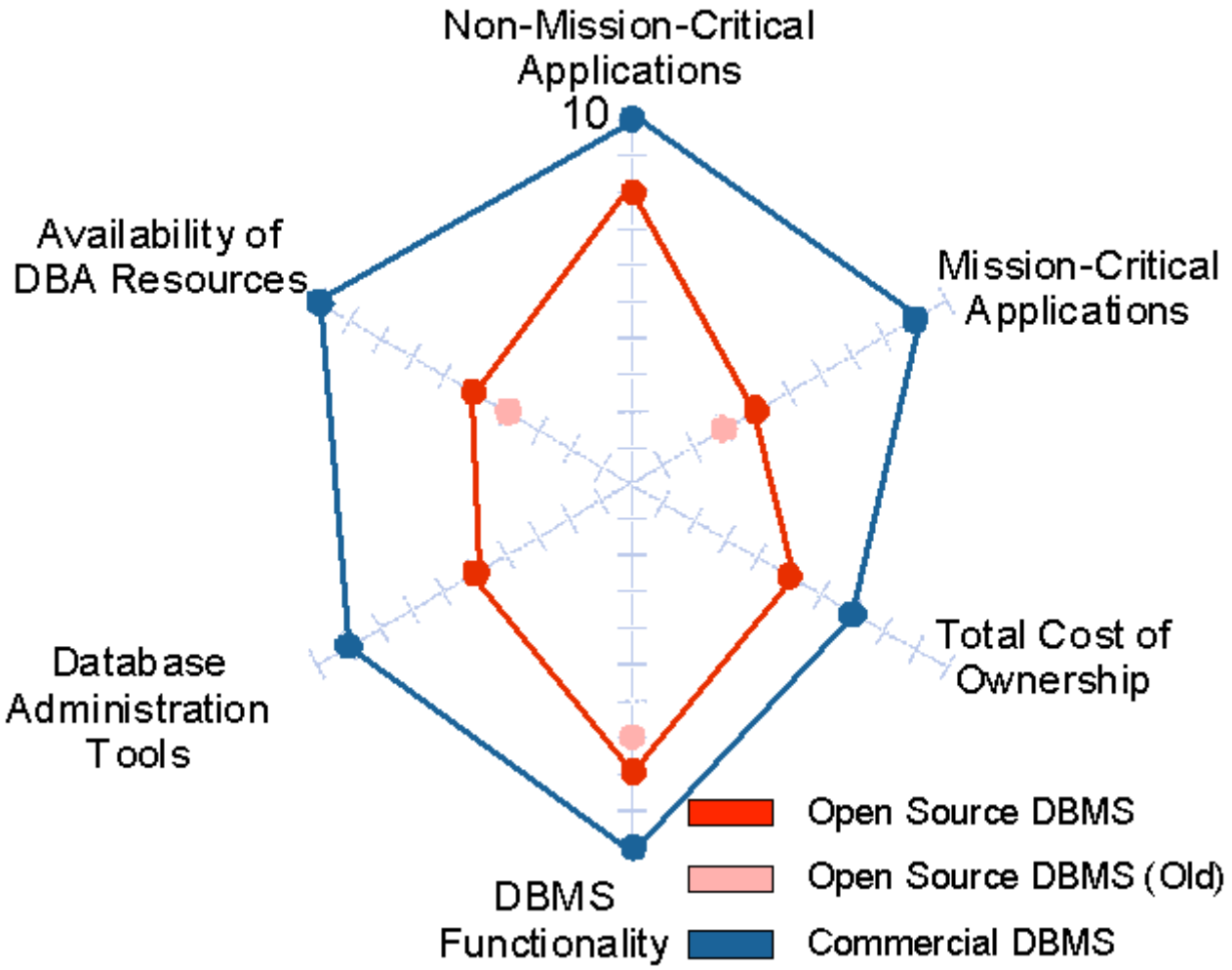
Figure 1. Maturity Level of Open Source DBMS



DBA = database administrator, DBMS = database management system

Source: Gartner (March 2008)

Maturity level of open source RDBMS (Nov 2008)



Open source OLTP database systems

- Traditional
  - Ingres
  - PostgreSQL
  - MySQL
  - Firebird
- Java
  - Apache Derby
  - HSQLDB
- Embedded
  - Oracle Berkeley DB

Open source analytical database systems

- C-Store
  - Commercialized as Vertica 3.0, company founded by M. Stonebraker, founder of Ingres
  - Columnar, MPP, data compression, HA
- MonetDB
  - High performance applications in data mining, OLAP, XML Query
- LucidDB
  - Designed for LucidEra BI (SaaS). Version 0.9.1 available
- Eigenbase
  - Used by SQLStream and LucidDB. Version 0.9.0

PostgreSQL

- Robust object-relational RDBMS
- BSD license
- Extensive community
- Runs stored procedures in many languages
- Interfaces for Java, ODBC, Perl, PL pgSQL…
- Triggers and stored procedures can be written in C and loaded as libraries
- It has a commercial version: EnterpriseDB (PostgresPlus Advanced Server)
- It has a parallelized version: Greenplum

MySQL

- The most popular open-source DBMS
  - Part of the LAMP stack
  - Highly used in web site hosting and development
  - Targets developers, ISVs, VARs, hardware vendors and network appliances
- Pluggable storage engines architecture
  - MyISAM, InnoDB, Falcon, …
- Cluster architecture

DBMS Tools

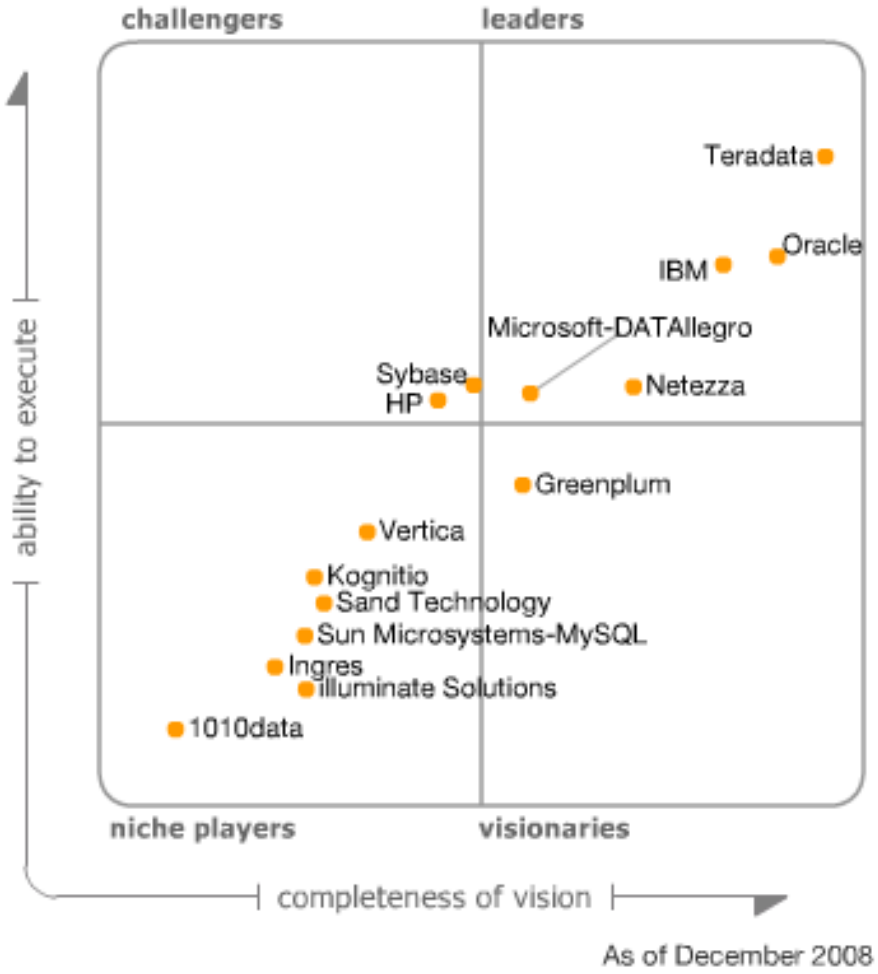
- MySQL GPL Tools
  - Administrator, Query browser, Migration utility
- SQuirrel SQL Client
  - GPL + LGPL license
  - JDBC access
- SQL Power
  - Power*Architect data modeling
  - Power*MatchMaker data cleansing
  - JDBC drivers for PostgreSQL, MS SQL Server, MySQL, HSQLDB
- TOAD for MySQL
  - Free, not open source

Data Ware House products

Data Warehousing

- Data Warehouse
  - A reference database structured for analysis
    - Non transactional
    - Contents harmonized and comprehensive
    - Partitioning, bitmap indexes, materialized views, SMP support
- DWH vendors
  - Teradata is the first DWH pure player
    - Followed by DW appliance vendors: DATAllegro, Netezza and Sun-Greenplum
  - Every DBMS vendor supports DWH
    - Oracle, Sybase, IBM, Microsoft
    - Specialized: Kalido, Kognitio
  - DW techniques are portable to any DBMS platform

Data Warehousing



Open Source Data Warehousing

- Three leading Open Source DBMS players:
  - Ingres
  - MySQL
  - PostgreSQL
- Ingres is possibly the most enterprise worthy
- MySQL, popular but limited DW capabilities before version 5.1
  - Strong point: multiengine architecture
  - Look at MyISAM and InfoBright
- PostgreSQL robust enterprise platform
  - Greenplum database designed for DWH
  - Truviso add streaming
  - Bizgres is dead
  - EnterpriseDB is the commercialization of PostgreSQL

DWH Recommendations

- Technological evolution
  - MPP
  - Column stores (InfoBright)
  - Search-reliant data warehouses
  - Data stream management (Truviso)
  - Appliances
- OS option
  - Ingres Icebreaker or Greenplum-Sun vs Netezza or DATAllegro
  - Adopt MySQL but evaluate performance and scalability, considering enhancements as InfoBright
  - Enterprises should consider supported RDBMS as Ingres and EnterpriseDB
  - Consider MonetDB

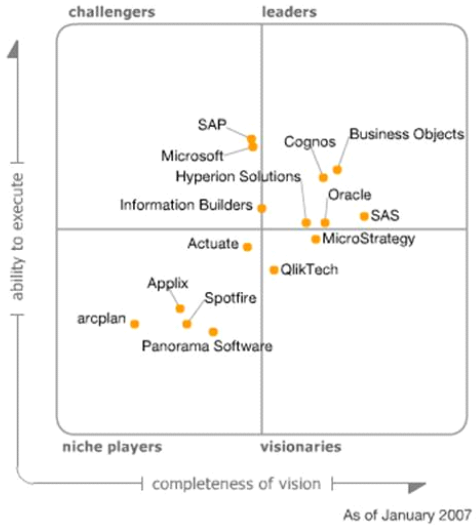
Business Intelligence tools and platforms

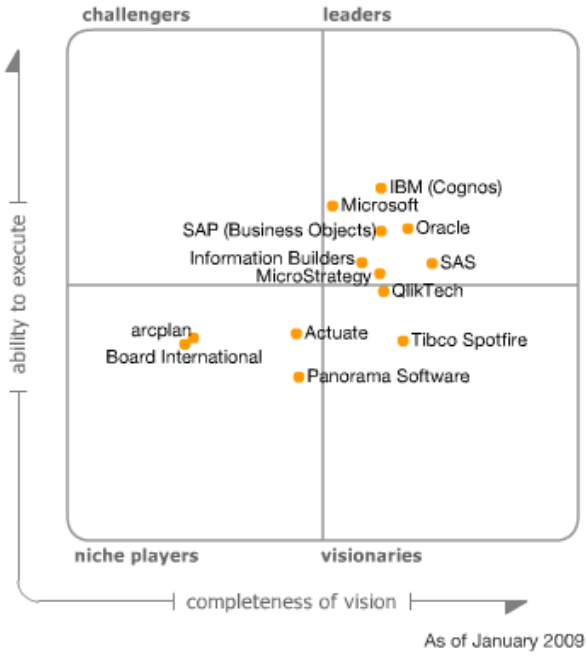
Business Intelligence

- More than just software
  - Integration with operational systems
  - Embedding analytics in business applications
  - Collaboration
- BI tools:
  - Reporting, dashboards, ad-hoc query
  - OLAP analysis
  - Advanced analytics (data mining, statistics, geospatial analytics)
  - Application integration

Business Intelligence Scene

- Many BI vendors
  - Dominators: SAP Business Objects, IBM Cognos, Oracle Hyperion, MicroSoft
  - Pure player: Microstrategy, SAS, SPSS
  - Visualization specialized: Actuate, TIBCO Spotfire, Tableau, QlickView





OW2 Europe Local Chapter

The OLAP Report (Nigel Pendse)

- "Current OS OLAP solutions are quite weak (at least a decade behind the current proprietary products), whereas the reporting solutions may be better …"
- "The proprietary BI software vendors seem to be genuinely unconcerned by open source BI. I guess they don't sell into OSW anyway and therefore aren't losing any business to OS BI that they are aware of."
- This is a "category error"
  - Open source does not succeed by replicating commercial proprietary software and processes
  - The most successful open source projects are innovative
  - OSBI as not aimed to replace closed-source, commercial solutions … YET!

OS BI Analytical Tools

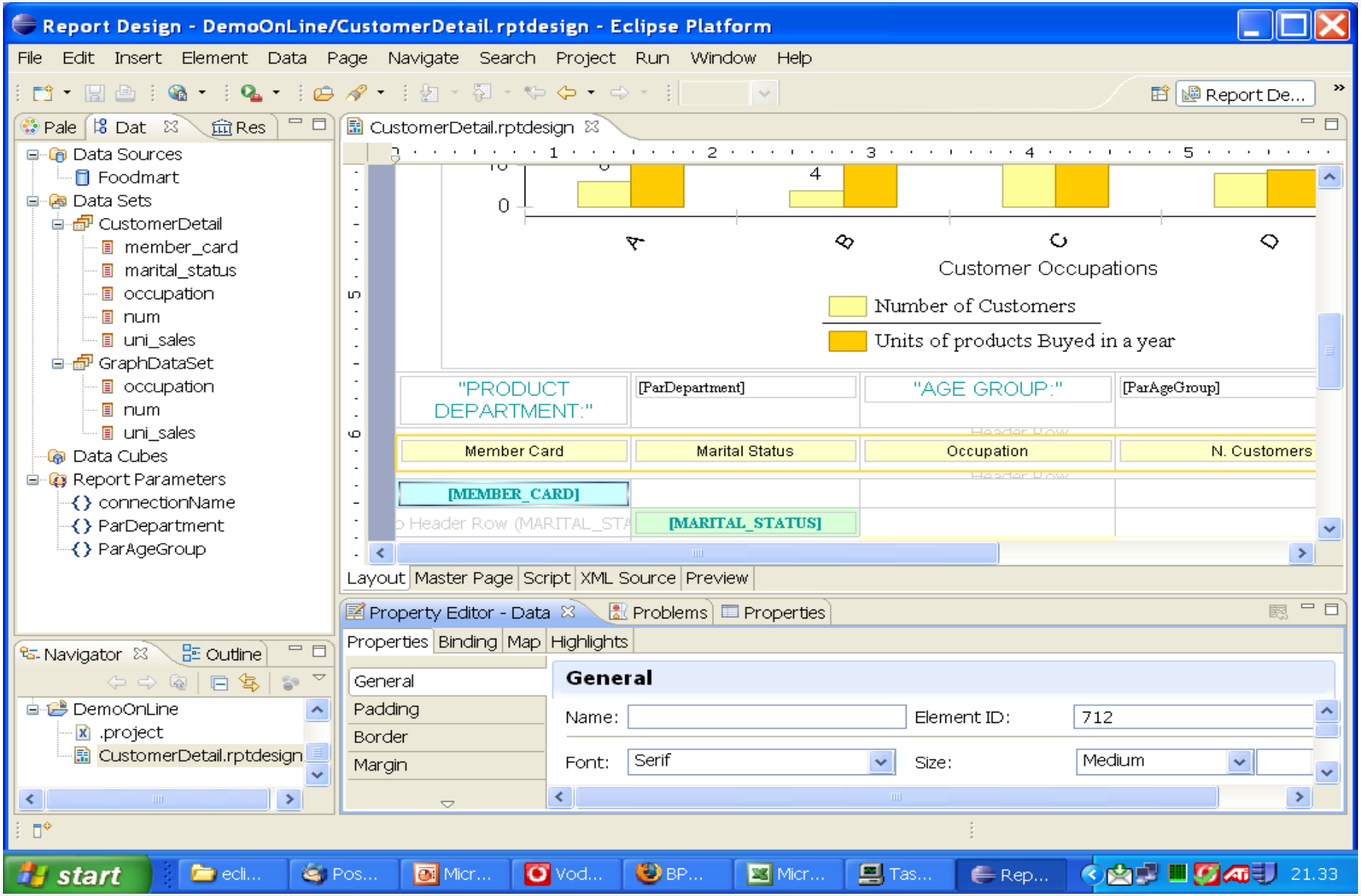
- Reporting
  - JasperReports
  - Eclipse BIRT from Actuate
  - JFreeReports
- OLAP
  - Mondrian Relational OLAP Server (ROLAP) + JPivot tag library
  - Palo Multidimensional OLAP Server (MOLAP)
- Data mining
  - R is an implementation of a statistical programming language
  - Weka is a machine learning tool

Reporting

BIRT Report Engine

- Eclipse project including
  - Graph generator
  - Report generator
  - Design environment (Eclipse based)
- Managed by Actuate that commercialize a BI offer whose only open source solution is BIRT
- Library allowing to generate reports in different format
- The report can mix data, graphics and images
- Can be integrated in any Java application

BIRT Report Engine



OW2 Europe Local Chapter

BIRT Report Engine

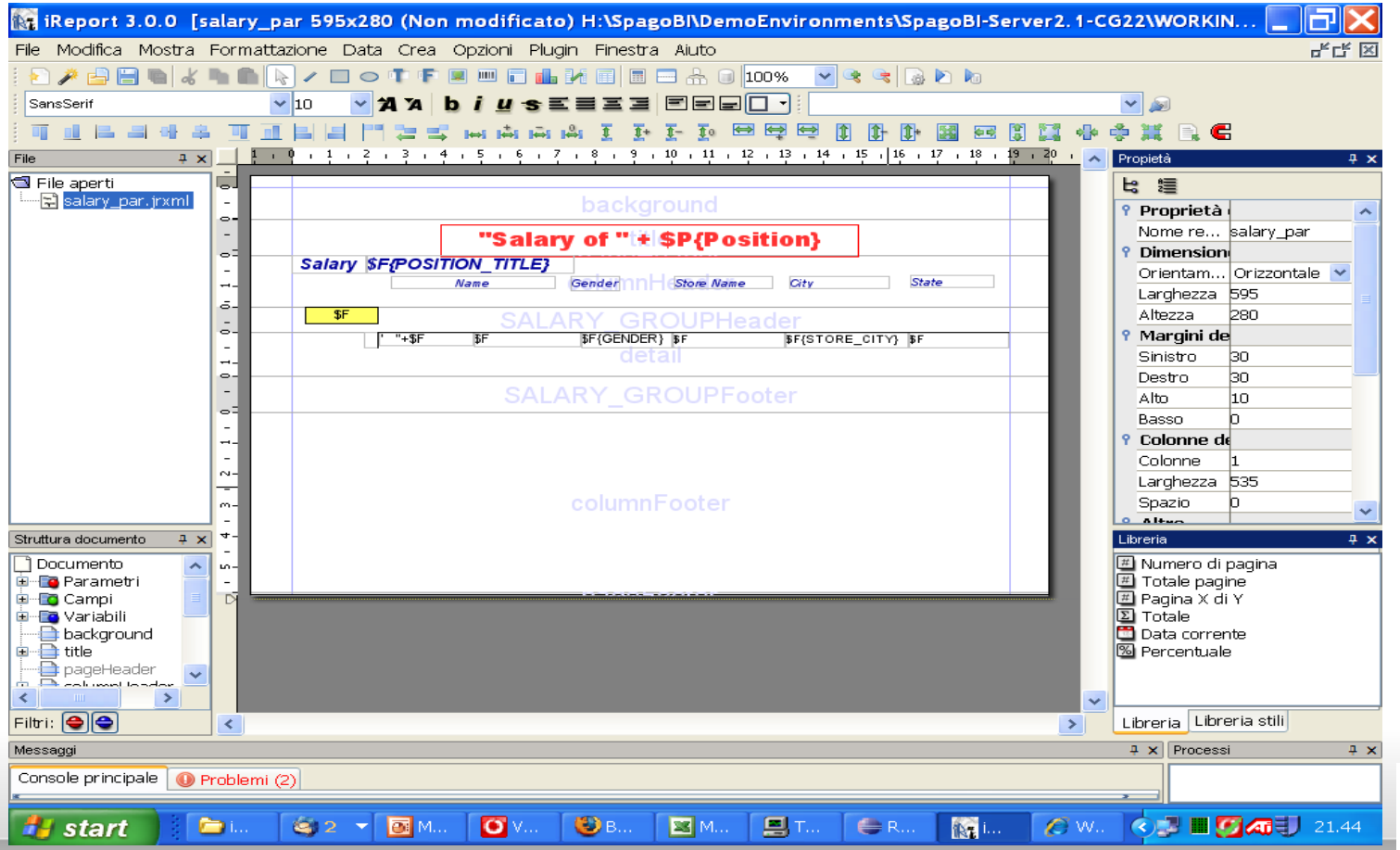
- Essentially oriented to developers, requests must be written in SQL
- It is possible to make BIRT accessible by less technical users
- It is possible to create resource libraries containing the basic elements to produce a report
- Strength
  - the Eclipse community
  - the ease of use

Jasper Reports

- Report engine developed by JasperSoft and distributes in open source
- Report are described as xml files that can be built:
  - Manually
  - Using ad-hoc tools (ex. iReport)
- Generates report in different formats:
  - HTML, PDF, XML, CSV
- The layout of the report is composed of layers:
  - Title, page header, column headings, details, column footers, page footer, last page, summary page
- It is possible to use subreports

iReport

- Tool to design Jasper reports
- Oriented to report developer
- Less intuitive than BIRT



OW2 Europe Local Chapter

Pentaho Report Designer

- Formerly known as JFreeReports
- Joined Pentaho in 2006
- It allows to directly deploy reports in the Pentaho platform
- It supports different formats:
  - PDF, HTML, CSV
- Reports are developed in layers, as in JasperReports
- Wizards are available

Multidimensional Analysis (OLAP)

Mondrian

- OLAP server
- It belongs to the ROLAP Category (Relational OLAP) since it access a relational data base
- Mondrian executes requests described in MDX language
- Mondrian can be used together with its client JPivot
- It also exposes XMLA interface allowing to be accessed by other clients (ex. JPalo)
- The Mondrian project has joined Pentaho and renamed ad Pentaho Analysis

JPivot

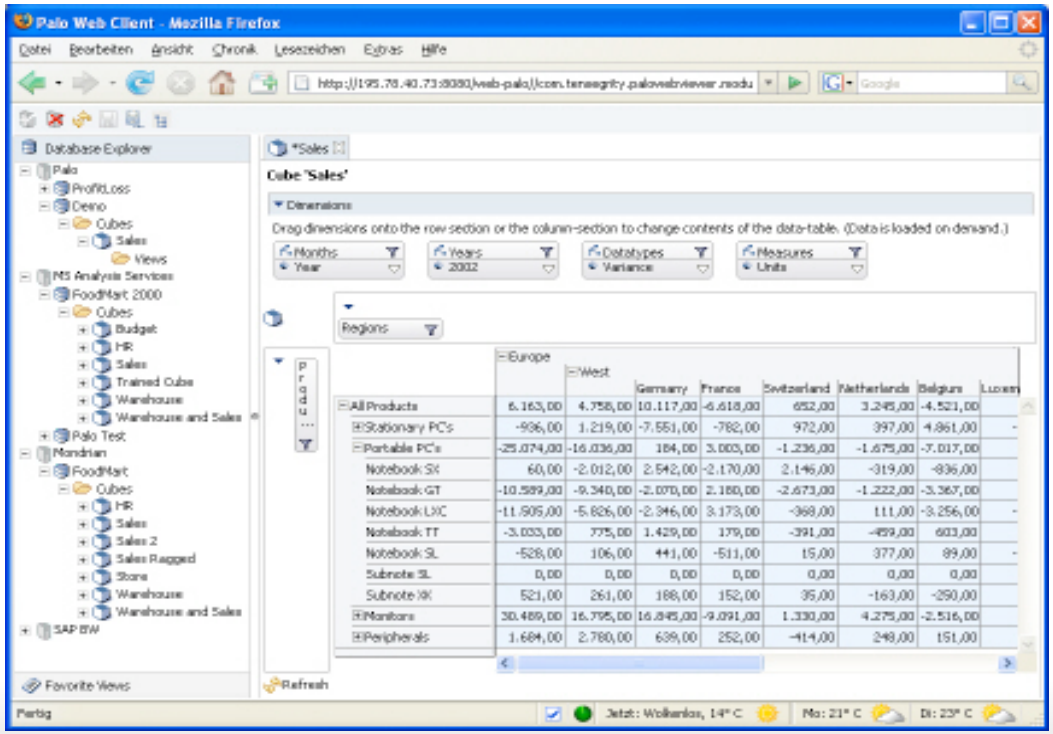
- OLAP client
- It allows to represent a OLAP cube and to navigate it
  - Drill down, drill up
  - Drill across, drill through
  - Slice and dice
- It allows to associate a graph to the dimensional table
- It exports in PDF or Excel
- The user interface can be customized using style sheets

Palo

- OLAP server
- It belongs to the MOLAP Category (Multidimensional OLAP) since it load data in a dedicated structure
- A plugin is available to access Palo server from Excel
- It can be accessed by a JPalo client
- In the commercial version it is possible to select and change the values and to spread aggregated data trough the details

JPalo

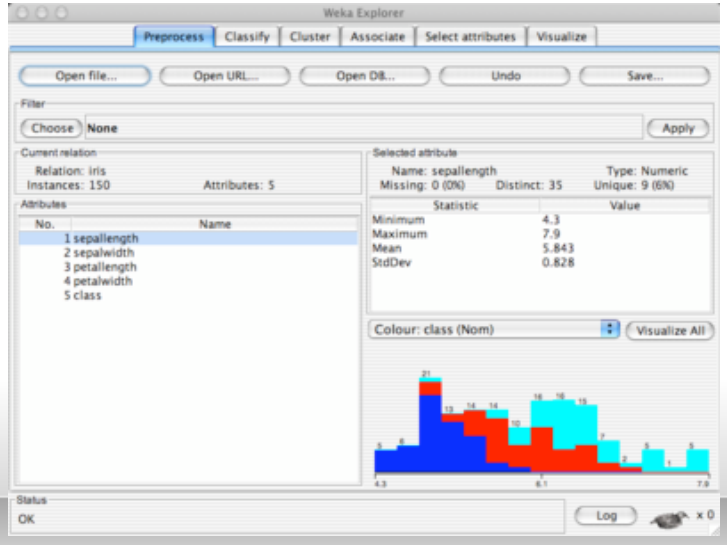
- OLAP client
- Web interface to access both Palo and Mondrian
- As an alternative you can user Palo Eclipse Client, a thick client based on Eclipse

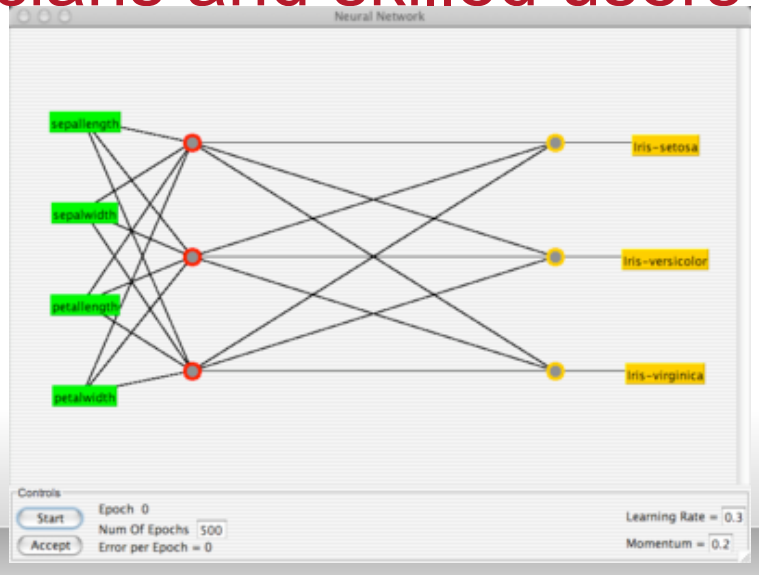


Data Mining

Weka

- Tool allowing to execute data mining algorithms
- It has its own user interface
  - Graphical
  - Command line
- It allows to use the single algorithms or to chain them in a workflow process
- Oriented to statisticians and skilled users





OW2 Europe Local Chapter

Business Intelligence Platforms

Pentaho BI Suite

- Product suite to distribute analytical functionalities and documents through
  - portals (JBoss portal)
  - web application
- It has a double-license model
  - Community edition: free open source
  - Enterprise edition: license fee
- Open Source Modules
  - Pentaho reporting
  - Pentaho analysis
  - Pentaho dashboard
  - Pentaho data integration
  - Pentaho data mining

Pentaho Enterprise Edition

- The main modules are “certified”
- Professional support
- Software maintenance
- Main enhanced functionalities:
  - Console
  - Dashboard designer
  - SSO
  - Lifecycle management
  - Audit reports
  - Clustering
  - Performance monitoring
  - ETL management and monitoring
- Product expertise
- Software assurance

Pentaho: main components

- Workflow engine
  - Based on Shark
  - It allows to structure a decision process by means of action
  - Each action is described in a XML file
  - The XML files are created in the Pentaho Studio environment, an eclipse based user interface
- Task Scheduler
  - Based on Quartz
  - It allows to schedule any Pentaho action
  - It allows to periodically send reports by mail
  - The task control can be manual or linked to an action

Pentaho: user interface

- Web application
  - It manages user roles in accessing functionalities
  - It is the preferred way to access Pentaho
- Portal
  - It manages portlets in JBoss Portal
    - EmbeddedReportPortlet
    - ChartPortlet
  - The security is managed by the portal

SpagoBI

- Integration Platform
- Totally open source, only one version and one license
- It has a open architecture allowing to integrate new components both open source and proprietary
- It integrates open source solution and provide some original ones

SpagoBI: modules

- SpagoBI Server
  - SpagoBI Reporting
  - SpagoBI OLAP
  - SpagoBI Free Inquiry (QbE)
  - SpagoBI GEO
  - SpagoBI KPI
  - SpagoBI Dashboards
  - SpagoBI Data Mining
  - SpagoBI ETL – Talend
- SpagoBI Studio
- SpagoBI Metadata
- SpagoBI SDK
- SpagoBI Applications

SpagoBI

- Analytical model
  - Set of different solutions for different analytical areas
- Behavioural model
  - Manages user roles
  - Associate functionality to user roles
  - Associate data visibility to user roles
- Cross-navigation
  - Allows to link analytical documents between them

SpagoBI: the user interface

- Web application
  - Can be deployed on any Web Container as: Tomcat, JBoss, WebSphere
  - Security is managed by the integrated CAS module
- Portal
  - Can be deployed on any Portal Container compliant to the JSR 168 standard as: eXo WebOS, Liferay
  - Security is managed by the portal
  - The source code is the same: deploying as web application or portal is a matter of configuration

Jasper Intelligence

- The BI platform of JasperSoft
- Main modules
  - Jasper Server
  - Jasper Analysis
  - Jasper Reports
  - Jasper ETL
  - iReport
- It is available under two licenses:
  - GPL (BI for Everyone or JasperSoft Community)
  - Commercial (JasperSoft Professional Edition)
- Users can build their reports
- The user interface is based on a specific web application, no use of portal

Jasper Intelligence: commercial version and ETL

- The commercial version includes:
  - Certified support
  - Release cycle management
  - Support guarantees
  - Legal matters
- Commercial version added functionalities
  - Jasper Server
    - Ad-hoc query and reporting, dashboard and mush-up designer, additional installers, comprehensive sample reports and analysis
  - Jasper Analysis
    - Drag and drop user interface, interactive charts, OLAP server management utility
  - Jasper ETL
    - Job monitoring tool, team development, slowly changing dimensions

ETL products

ETL

- Tools allowing to extract, transform (format, normalisation) and load data in the target database
- ETL manages different sources of data, both in input and in output (databases, XML files, CSV files, fixed format record files)
- ETL jobs are usually scheduled
- Open source ETL solutions:
  - Talend
  - Pentaho Data Integration (ex Kettle)

Talend

- Open source ETL belonging to the “code generators” category
- It allows to graphically design ETL processes and to generate code to be compiled and deployed to a target system
- Talend can generate Perl and Java code
- Used in SpagoBI and in Jasper BI Suite, where it has been renamed as JasperETL
- Talend Open Studio is the product to design the job and to generate the code

Talend Open Studio

- ETL processes are designed using a friendly user interface
- Native connectors exist to read and write from the most diffused data sources:
  - Almost all existing DBMS
  - XML files
  - Flat files (CSV or fixed record format)
- New interfaces and components can be added to the product
- It manages metadata and allows to build a Business Model of the process

Talend Integration Suite

- Features matrix

	Team Edition	Professional Edition	Enterprise Edition
Advanced Studio	✓	✓	✓
Shared Repository	✓	✓	✓
Activity Monitoring Console	✓	✓	✓
Job Conductor	✓	✓	✓
Activity Monitoring Dashboard		✓	✓
Job Conductor Advanced		✓	✓
Distant Run		✓	✓
Grid Conductor			✓
CPU Balancer			✓
Support & Maintenance	✓	✓	✓

Pentaho Data Integration (ex Kettle)

- Graphical model, based on steps
- Two types of processes:
  - Transformations: simple data management
  - Tasks: complex activities (mail, transformation execution, file download, …)
- Can be used in multi-user mode
- It provides several connectors
  - Databases
  - PALO cubes
  - LDAP
- It contains wizards to assist in creating read and write requests

Pentaho Data Integration (ex Kettle)

- 3 applications
  - Spoon: to create and execute transformations and tasks
  - Pan: command-line application to launch a transformation
  - Kitchen: command-line application to launch a task
- Scheduler
  - Based on external systems (cron …)

Thanks

Thank you for your attention !