Javlin, a.s.
Den s průmyslovými partnery FI 10. 12. 2009

December 9, 2009
Javlin a.s.

• Founded in 2005

• Global HQ in Prague, Offices in Washington, DC and UK

• ~45 Employees

• Developer of the CloverETL Solution

• Javlin solutions for mass data processing include
  – Data Quality
  – Master Data Management
  – Data Integration
  – Data Warehousing
A few of Javlin’s Worldwide Customers
What is CloverETL

• Dominate your data: huge batches or lots of small ad-hoc jobs
  ▶ Read from one or more data sources of almost any type
  ▶ Transform data in almost any way imaginable
  ▶ Write data out to any number of data targets of almost any type

• Common Approach: Programming or heavy DBA work

• Rapid Setup. Very fast execution. Scalable.

• Runs on PC, Mac OS X, Linux, Unix, AS/400

• GUI Driven
CloverETL - Dominate Your Data

- Master Data Management
- Data Migration
- Data Warehouse
- Data Consolidation
- Data Synchronization
- Data Integration
CloverETL - Continuous Product Line

- CloverETL Desktop Designer
- CloverETL Server
- CloverETL Cluster
Clover Designer - Quick Demo
# CloverETL - Server

![CloverETL screenshot](image)

The screenshot above shows the interface of CloverETL Server. It displays a table with the following columns:

- **run ID**
- **Node**
- **sandbox / graph**
- **executed by**
- **status**
- **started**

Each row represents a run with details such as:

- `node01`
- `default/graph/graphMountainsCSV.grf`
- `clover`
- `FINISHED_OK`
- `2009-11-10 02:51:39.0`

This interface allows monitoring, executing, and scheduling tasks, providing a comprehensive view of the ETL processes. The `sandbox` column indicates the specific sandbox used for each run, while the `graph` column specifies the graph file associated with the run.
CloverETL Cluster

**Increased processing power & throughput**
- Parallel execution of transformations over multiple servers / nodes
- Load balancing based on individual nodes utilization

**Increased fault tolerance**
- Fail over in case of problems with particular node(s)
- Data replication

**Increased flexibility**
- Cluster can be dynamically reconfigured by adding or removing nodes on the fly

![Diagram of CloverETL Cluster]

- Large quantity of data loaded
- Processed in parallel in Cluster
- Written out in parallel
Cloud Computing

• Amazon EC2, Voxel, Rackspace, MS Azure / Google App Engine, Salesforce

• Rent as many servers as you like by the minute / hour /month

• Highly configurable and rapidly deployable
  ▶ Server CPU / RAM / Hard Disk setup fully configurable
  ▶ Offline storage (costs extra / GB)
  ▶ Data snapshot capability and Machine Images
  ▶ Provider server farms and storage available West Coast, East Coast, Europe (Amazon)
  ▶ Content Delivery Network
  ▶ Autoscaling, Monitoring, Load-Balancing
  ▶ Custom and automated deployments
Amazon EC2 Cloud Case Study

The Washington Post

- 20,000 pages of non-searchable images released on paper
- Needed to convert to text format using OCR within 24 hour news cycle
- Data had to be scanned, formatted, image cleaned, resized and OCR'd
- 200 Amazon EC2 server instances used for 9 hours
- Total cost $144.62
CloverETL Cluster - In the Cloud

• Clover in the Sky with Diamonds!

• CloverETL Server Cluster will often run on In-House Servers

• But can be deployed very rapidly on Cloud servers

• Will offer substantial cost savings when load is controlled and predictable
CloverETL Cluster In the Cloud

- Processing of 6 million records of data
  - Calculates discount offered and date reformat
  - Sort all 6 million records
  - Aggregate total Price, Discount and Quantity
  - Calculate shipping period
  - Sort by Ship Time
  - Write out slowest 100 orders to be fulfilled

<table>
<thead>
<tr>
<th>Notebook</th>
<th>2 Nodes</th>
<th>3 Nodes</th>
<th>4 Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>150 seconds</td>
<td>60 seconds</td>
<td>47 seconds</td>
<td>31 seconds</td>
</tr>
</tbody>
</table>
Possible Cluster / Cloud Configuration

In House

Data read in here ....

Internet

FTP

is processed in parallel in the Cloud here ....

Amazon EC2 Cluster (US)

and is written out simultaneously here ....
## Cloud Computing - Yearly Costs

<table>
<thead>
<tr>
<th>Utilization</th>
<th>Cloud Cost (Amazon)</th>
<th>In-House Servers</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>$340</td>
<td>$8'700</td>
</tr>
<tr>
<td>10%</td>
<td>$680</td>
<td>$8'700</td>
</tr>
<tr>
<td>25%</td>
<td>$1'700</td>
<td>$8'700</td>
</tr>
<tr>
<td>50%</td>
<td>$3'400</td>
<td>$8'700</td>
</tr>
</tbody>
</table>
CloverETL Cluster Scalability

- Ditch versus hole
- Excellent scalability for data that can be “chunked”
- Each Node reads a separate “chunk” of the source data
- Even better performance can be realized if source data can be pre-split by the source application
Who are we looking for?

• Java/C programmers
  • Challenge-oriented programming
  • Rapidly developing family of products
  • Dbs, UNIX, data management experience is an advantage
  • English
Who are we looking for?

- **Consultants**
  - You have the knowledge
  - We offer the experience
  - Work abroad (international customers)
  - Problem solving skills
  - General knowledge in IT
  - Soft skills (customer oriented work)
  - Willingness to accept challenges and learn
  - English