



Cloud Computing *With MySQL and Pentaho Data Integration*

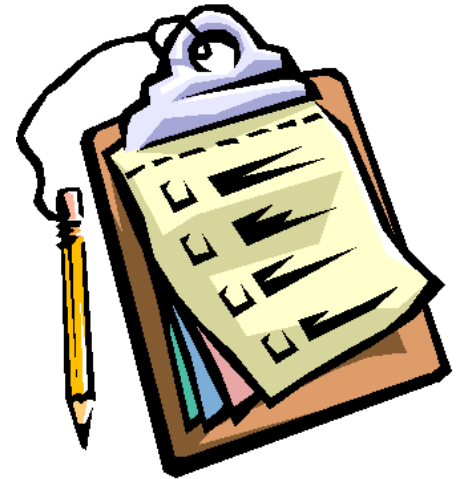
Matt Casters

Chief Data Integration at Pentaho

Kettle project founder

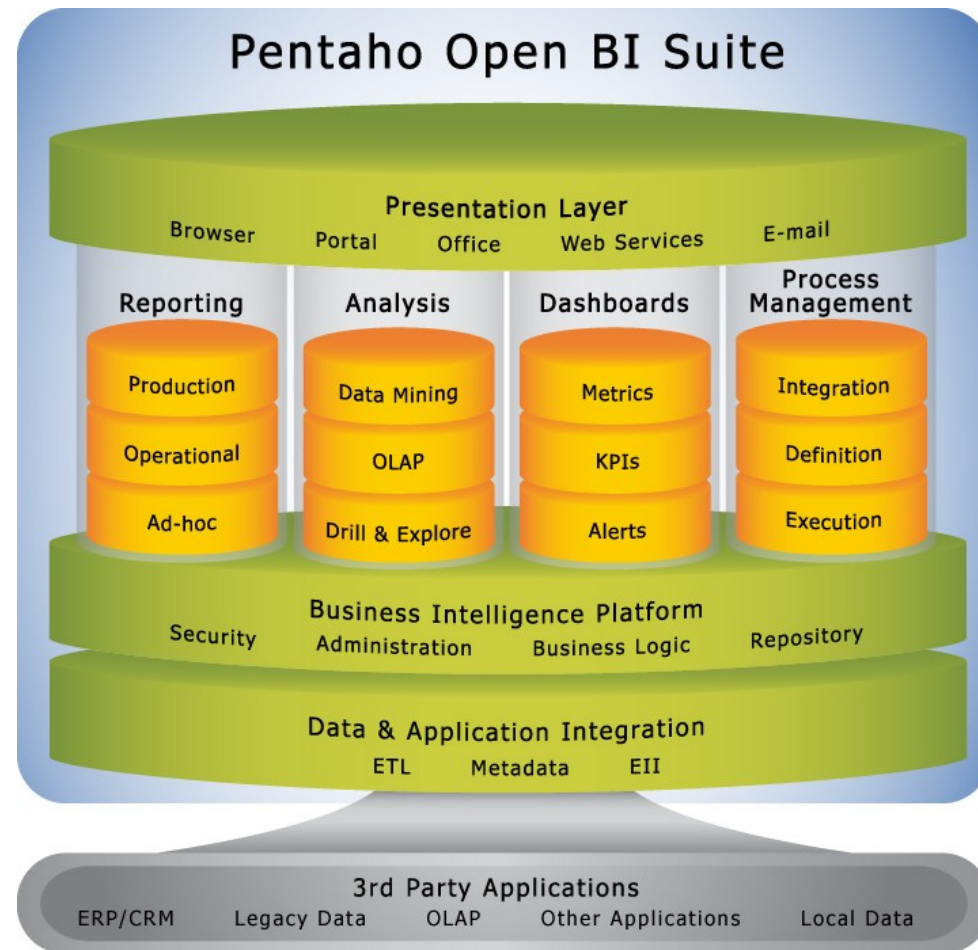
Agenda

- ***Introduction to Kettle***
 - Introduction
 - Use-cases + load demo
- Performance / Scalability
- Kettle Slave Servers
 - What & How
- Kettle Cluster Schemas
 - What & How
- The Cloud
- Cloud Examples
- Q & A



Pentaho Data Integration - Kettle

- PDI is the product associated with the KETTLE open source project
 - KETTLE provides open source software
 - PDI is a “whole product”
- Member of the Pentaho BI Suite
- Kettle = PDI CE
- PDI EE
 - Management Services Console
 - Knowledge Base
 - Documentation
 - Portal
 - Support
 - License Indemnification
 - ...



Introduction - Kettle : Kettle

Kettle

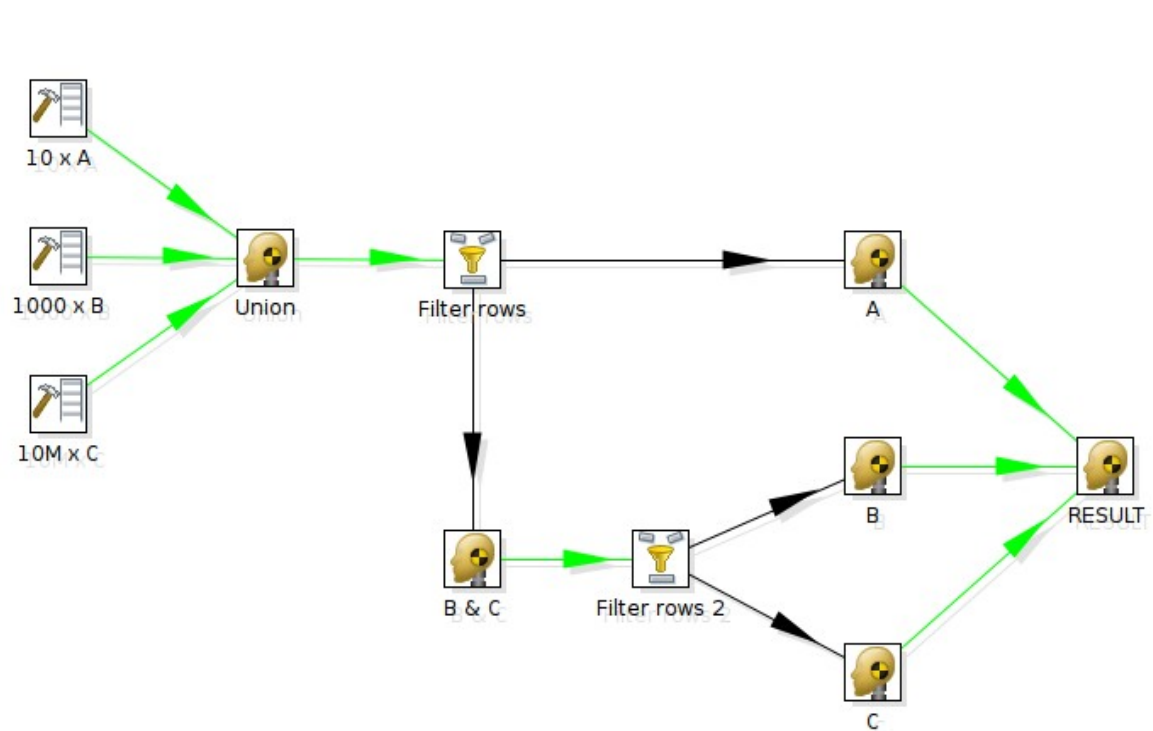
Extraction

Transportation

Transformation

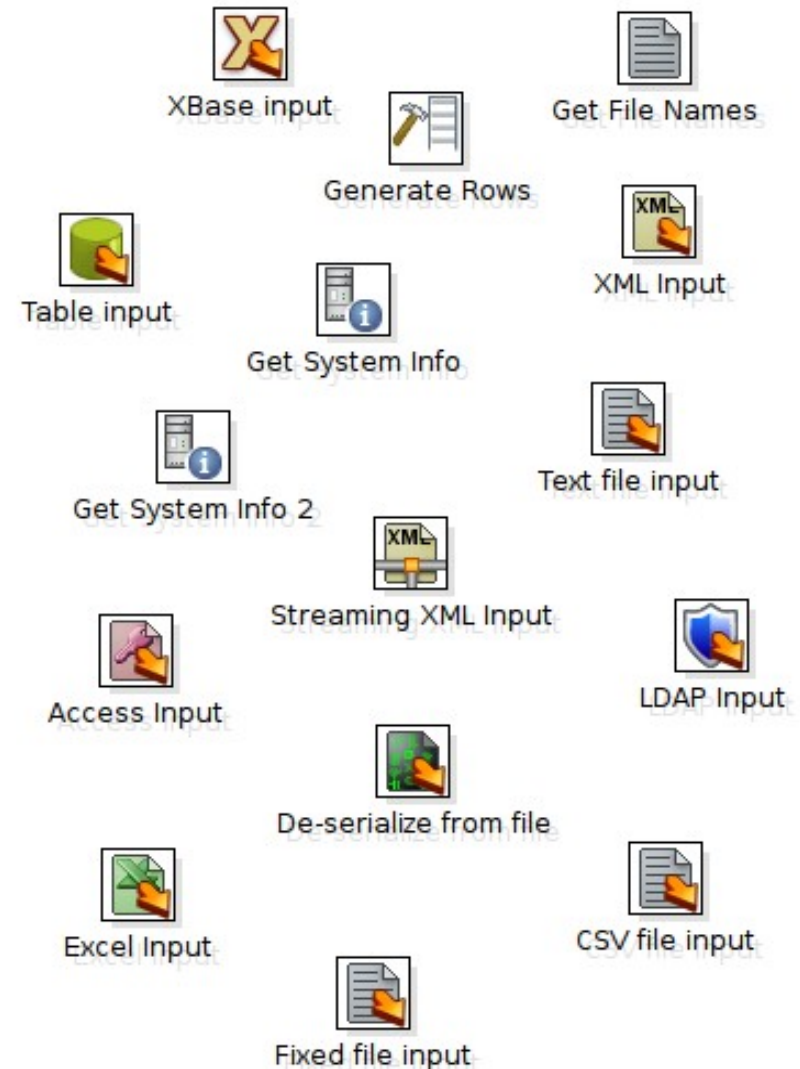
Loading

Environment



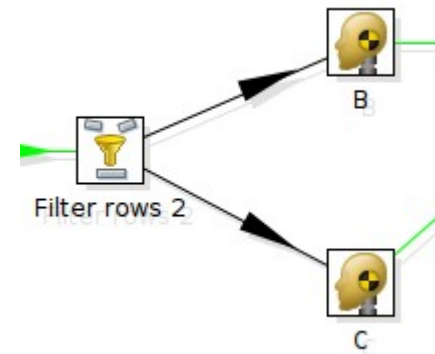
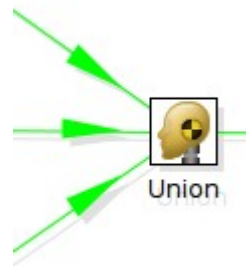
Introduction - Kettle : Extraction

- Extract data from :
 - 35+ database types
 - MySQL, PostgreSQL, SQLite, ...
 - Oracle, SQL Server, etc
 - Text files
 - XML files
 - XLS files
 - Xbase files (dBase, Foxpro, etc)
 - File systems information
 - Generated data
 - MS Access files
 - LDAP
 - Geo-data
 - ...



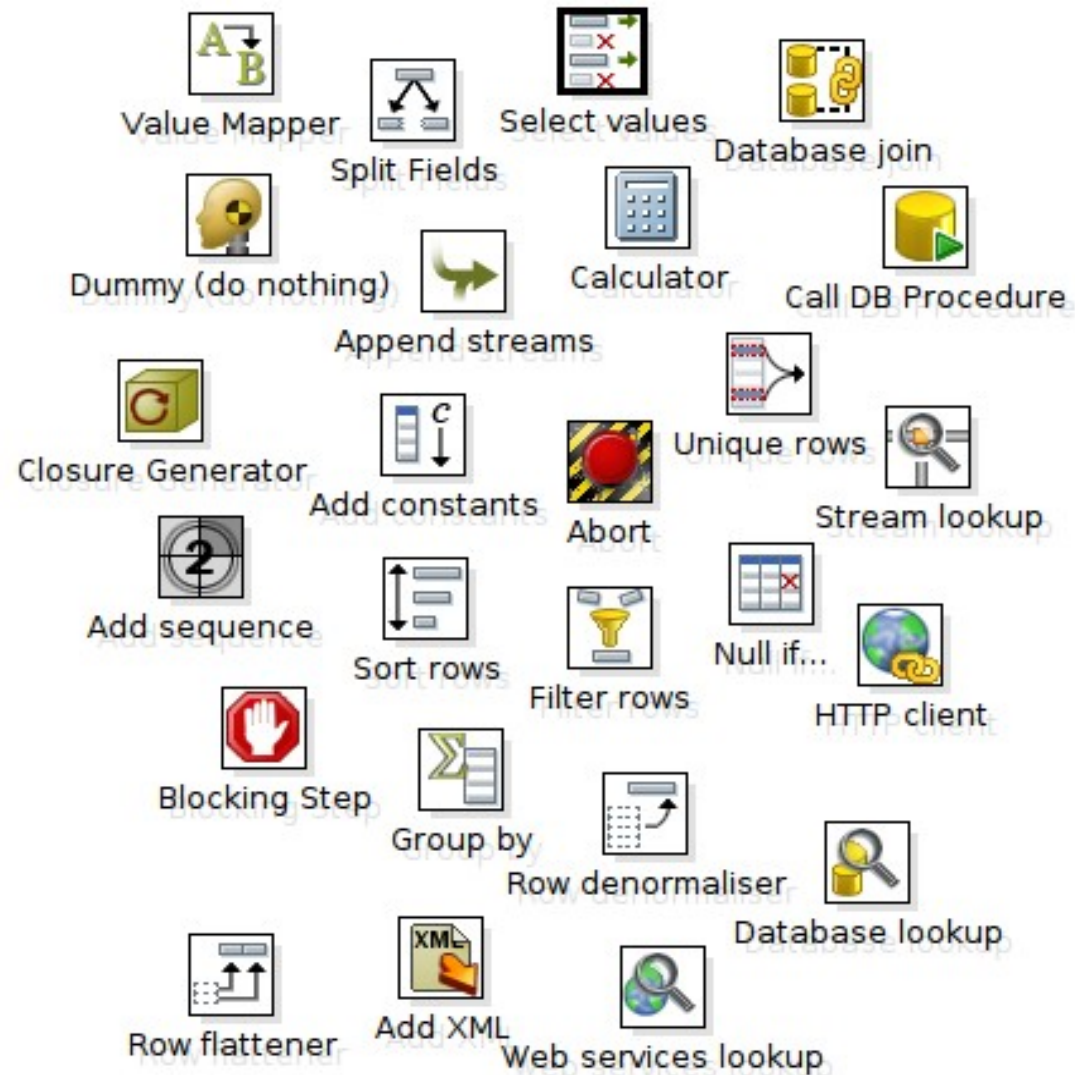
Introduction - Kettle : Transportation

- Transportation of data
 - Engine based data transfer (no code generator)
 - Very flexible pathways:
 - splitting
 - partitioning
 - merging
 - joining
 - duplicating
 - clustering (MPP)



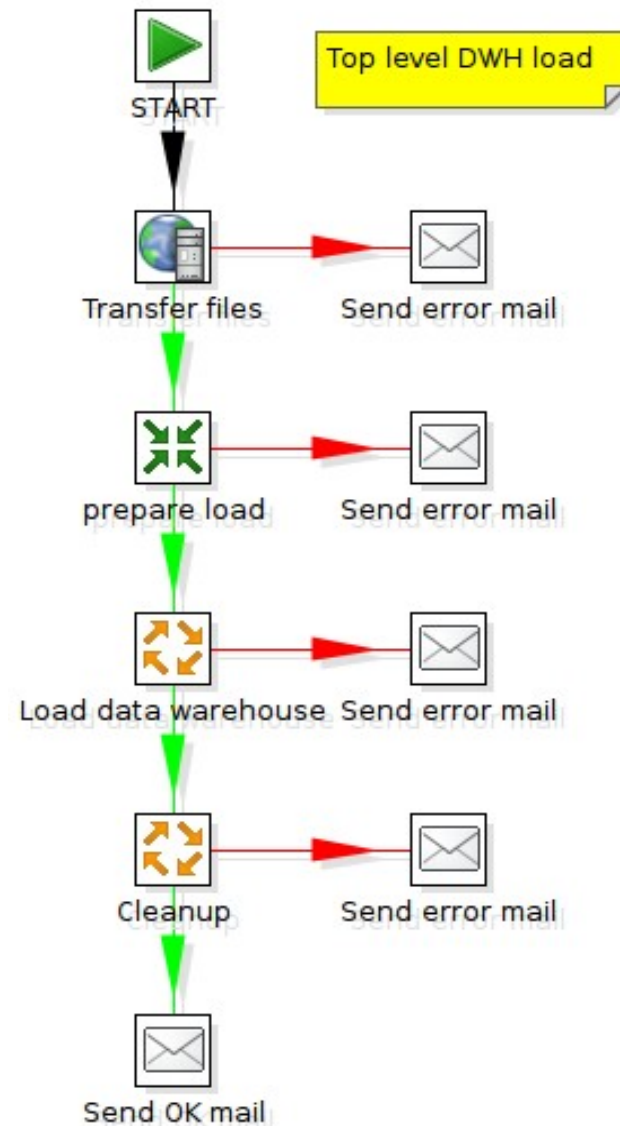
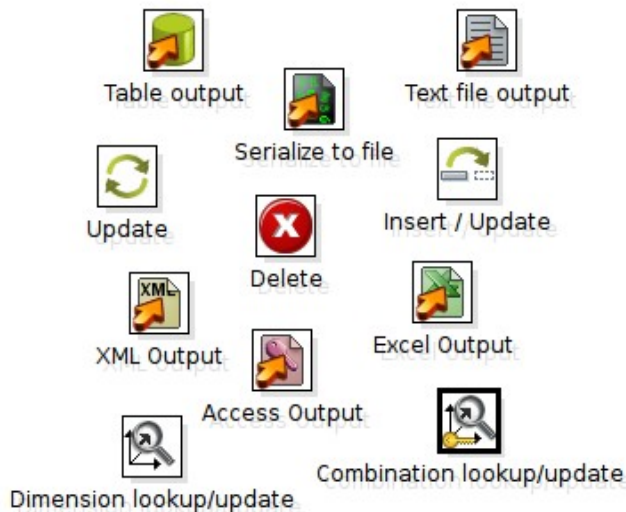
Introduction - Kettle : Transformation

- Flexibly transform data
 - Looking up data
 - databases
 - files
 - memory...
 - Calculating
 - Scripting
 - JavaScript, SQL, RegExp
 - Splitting
 - Mapping
 - Selecting
 - Filtering
 - Pivotting ...



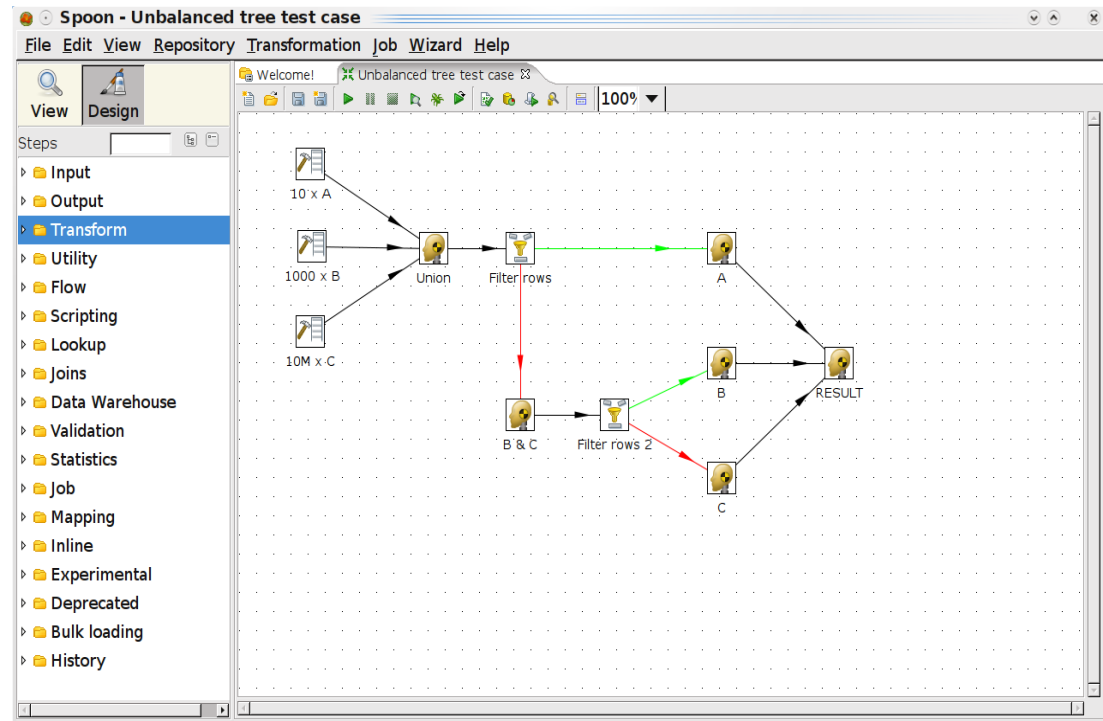
Introduction - Kettle : Loading

- Load data into a target format
 - Database loads
 - Data warehouse population
 - Partitioned loading
 - Bulk loading
 - Parallel loading
 - Clustering

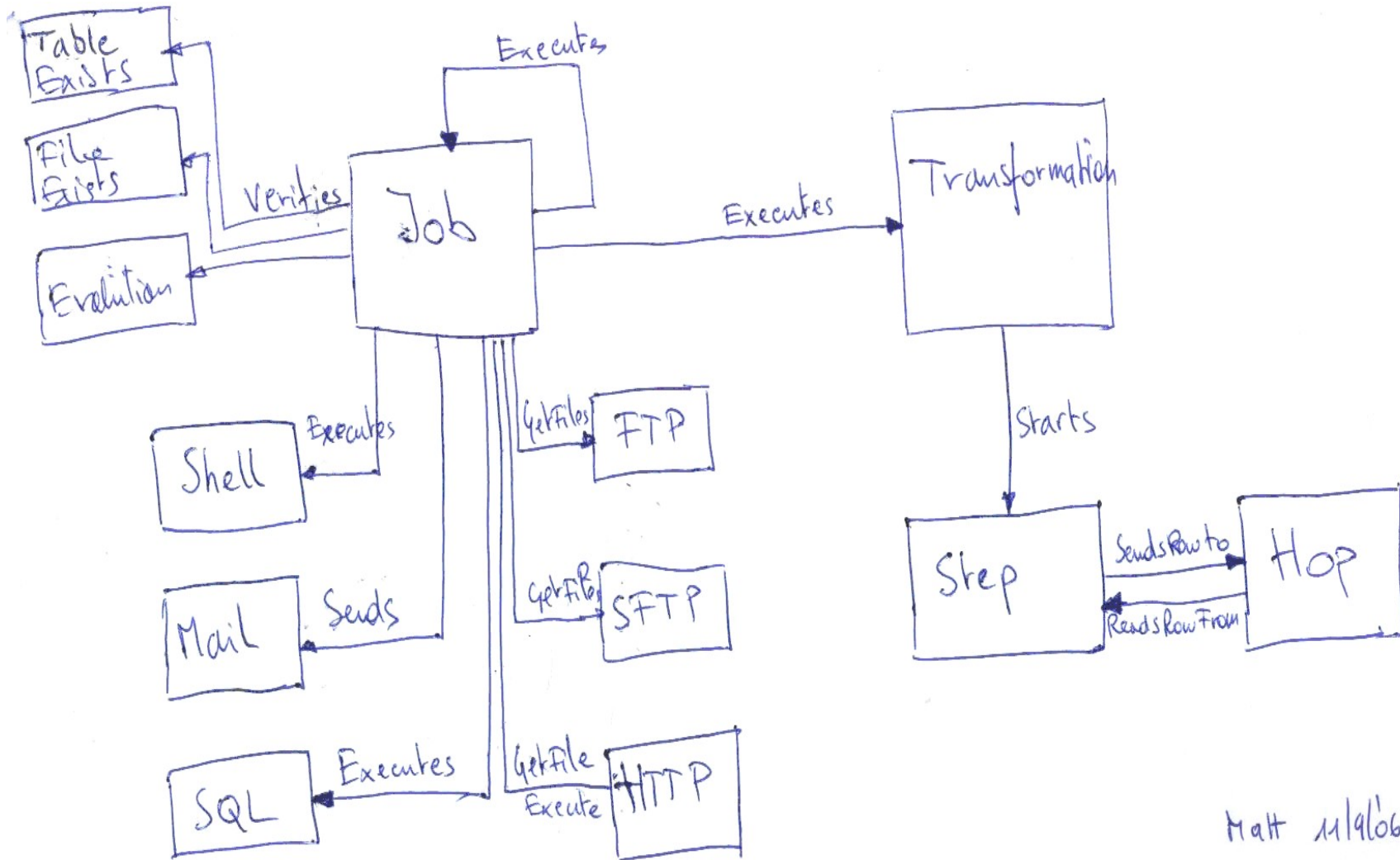


Introduction - Kettle : Environment

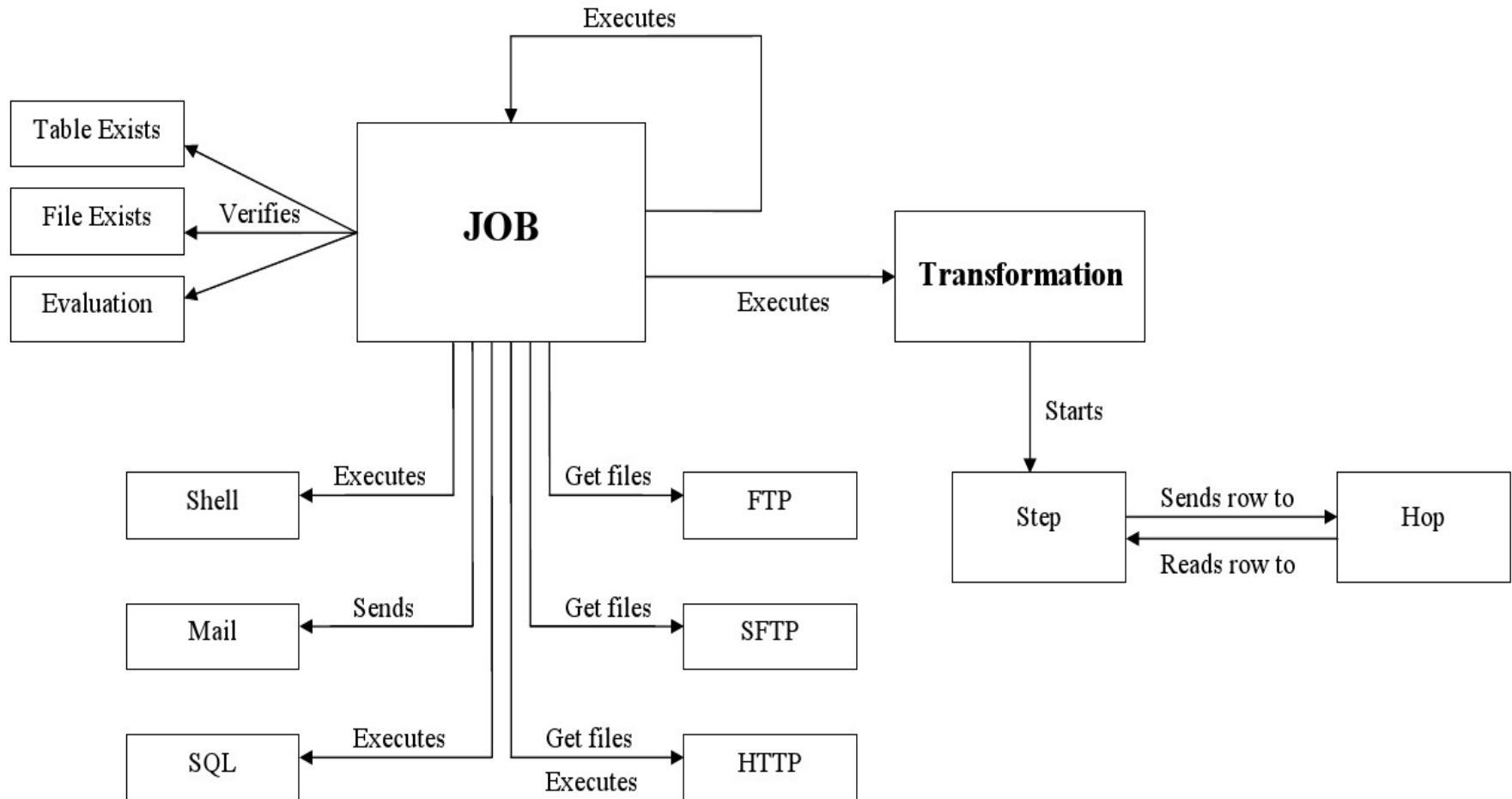
- Full GUI called “Spoon” to edit every option in Kettle
 - Drag & Drop
 - Debugger
 - Rich GUI
- Command line tools
 - execute jobs
 - execute transformations
- Web server
 - clustering
 - remote execution
- Programming API for Java
- Plugin eco-system
- ...



Introduction - Kettle : Conceptual model



Introduction - Kettle : Conceptual model



Introduction - Kettle : User community

- Paying Pentaho customers
- Large and small corporations
 - All possible sectors
- Lone rangers & Hobbyists
- All regions on Earth
- Meet on our Forum : +30,000 posts in 3 years
- Use our JIRA case tracking systems
- Download more than 10,000 copies of Kettle per month



<http://www.ohloh.net/projects/3624?p=Kettle>



<http://www.softpedia.com/progClean/Kettle-Clean-80094.html>

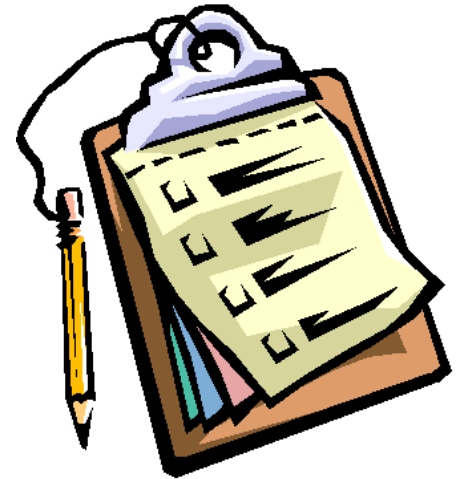
Typical use-cases

- Load data from text files and store it into a database [demo]
- Export data from database to text-file or more other databases
- Data migration between database applications
- Exploration of data in existing databases (tables, views, etc.)
- Information improvement using lookups
- Data cleaning
- Application integration
- Data warehouse population
- Application integration
- Report data generation
- ...



Agenda

- Introduction to Kettle
 - Introduction
 - Use-cases + load demo
- ***Performance / Scalability***
- Kettle Slave Servers
 - What & How
- Kettle Cluster Schemas
 - What & How
- The Cloud
- Cloud Examples
- Q & A



Party time!!!!

- You're preparing for a big event
- You bought plenty of food:
 - 500 hot-dogs
 - 500 burgers
- Distaster strikes: invitees are not showing up!
- What do you do with all that food?



Call Joey Chestnut!!

- Fastest eater in the world **
- Local boy from Vallejo
- 103 burgers in 8 minutes
- 66 hot dogs in 12 minutes



**** Ranked #1 by the International Federation of Competitive Eating**

Problems with the chosen solution

- You would need at least 10 Joey Chestnuts to get the “job” done
- Getting Joeys to show up costs money \$
- It's more fun with a crowd!

Performance / Scalability

- Demo: what about MySQL limits? **[loading data continued]**

Performance / Scalability

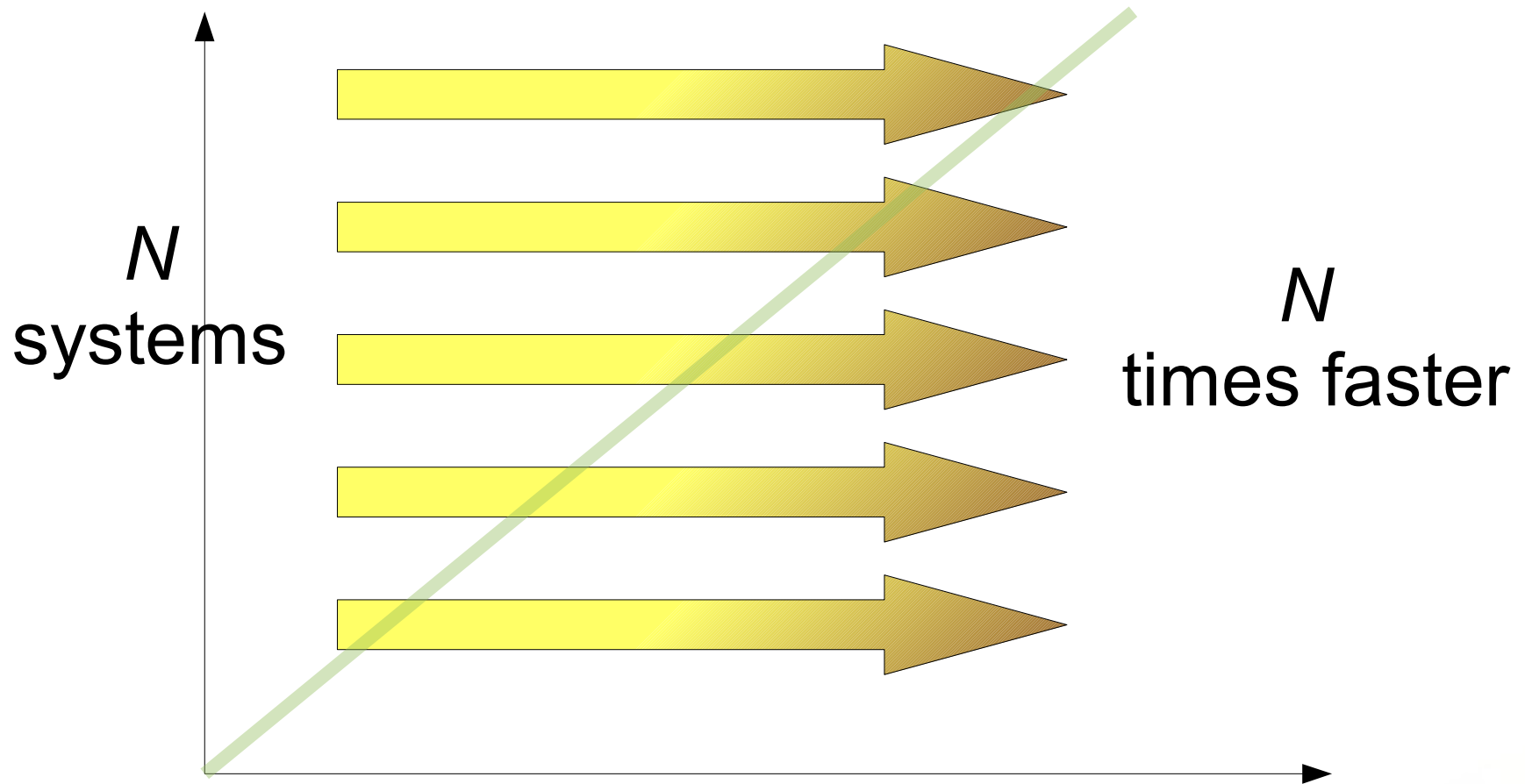
- MySQL bulk loading limitations...
 - Memory backed B-tree grows and need to swap out to disk
 - Kills performance
 - MySQL Performance Blog (Percona)
 - Predicting how long data load would take
 - Predicting performance improvements from memory increase

Performance / Scalability

- Limits
 - Single threaded limits
 - Multi-threaded limits
 - The weakest link
- Solutions?
 - Optimizing
 - Tweaking
 - Prodding
 - Removing bottlenecks
 - ...
 - Scaling out
- Scaling out with Pentaho Data Integration
 - Clustering
 - Partitioning
 - Database sharding/partitioning

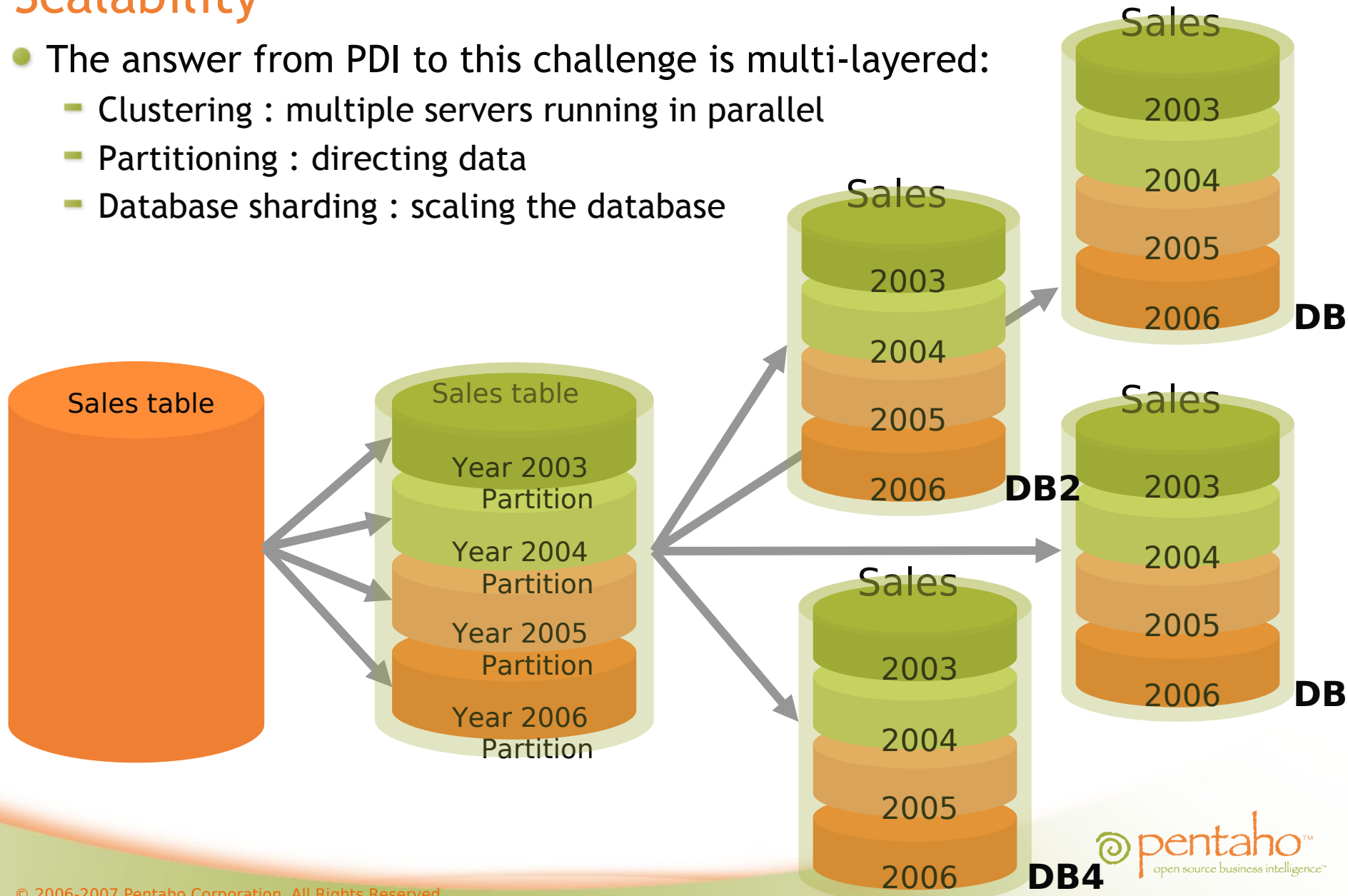
Performance / Scalability

- The “Ideal” architecture has linear scalability



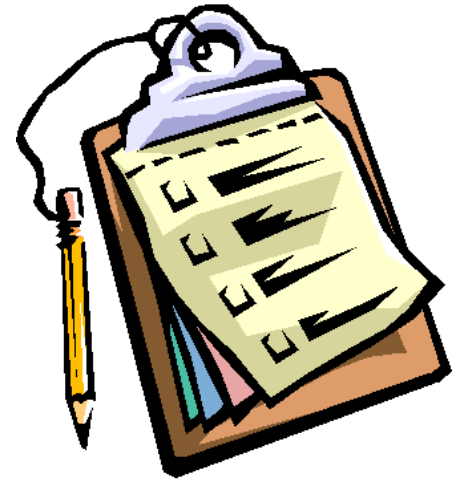
Scalability

- The answer from PDI to this challenge is multi-layered:
 - Clustering : multiple servers running in parallel
 - Partitioning : directing data
 - Database sharding : scaling the database



Agenda

- Introduction to Kettle
 - Introduction
 - Use-cases + load demo
- Scalability
- ***Kettle Slave Servers***
 - What & How
- Kettle Cluster Schemas
 - What & How
- The Cloud
- Cloud Examples
- Q & A



Slave Servers

- Building block of the PDI clustering offering
- Small embedded webserver (Jetty)
- Controlled over HTTP
- Spits out XML or HTTP **[demo]**
- Easy to start / configure / use
- Available HTTP services:
 - Start / Stop transformation or Job
 - Pause transformation
 - Adding (posting) transformation or job
 - Get status of server, transformation or job
 - Cleanup of transformation
 - Allocate a socket port
 - Register slave
 - Get list of slaves

Slave Server Configuration

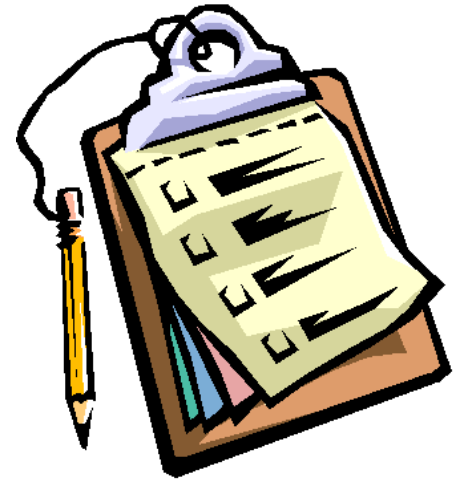
- Simple configuration : Hostname & HTTP Control port

```
sh carte.sh localhost 8080
```

- XML configuration:
 - Optionally look at network interface to grab address
 - Optionally report to a central server
 - Etc.

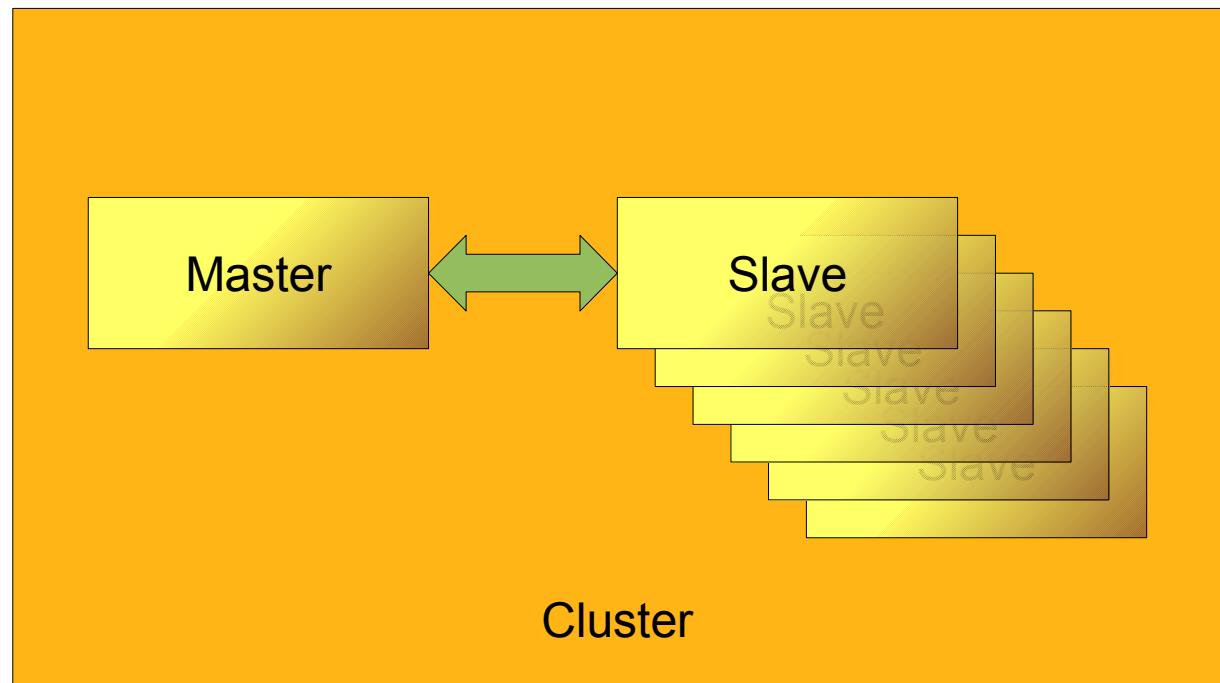
Agenda

- Introduction to Kettle
 - Introduction
 - Use-cases + load demo
- Scalability
- Kettle Slave Servers
 - What & How
- ***Kettle Cluster Schemas***
 - What & How
- The Cloud
- Cloud Examples
- Q & A



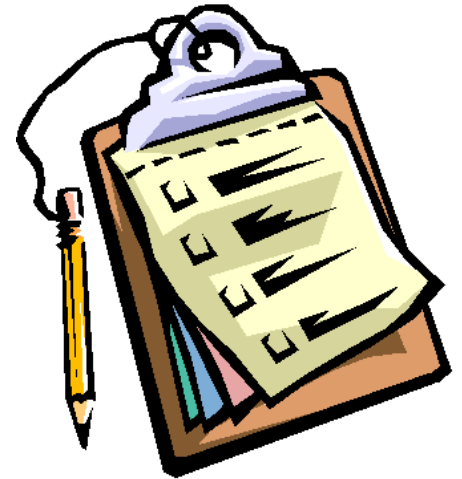
Clustering Schema

- Consists of a collection of one or more slave servers
- Is made up of
 - Master : at least one per cluster
 - Slaves : parallel worker nodes
- Simple local sample **[demo]**



Agenda

- Introduction to Kettle
 - Introduction
 - Use-cases + load demo
- Scalability
- Kettle Slave Servers
 - What & How
- Kettle Cluster Schemas
 - What & How
- *The Cloud*
- Cloud Examples
- Q & A



The Cloud

- Wikipedia:

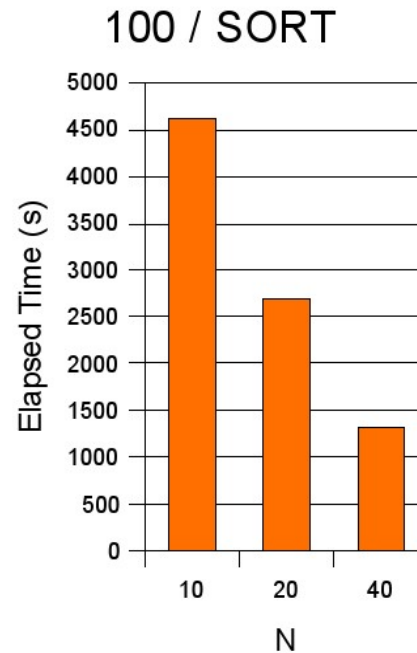
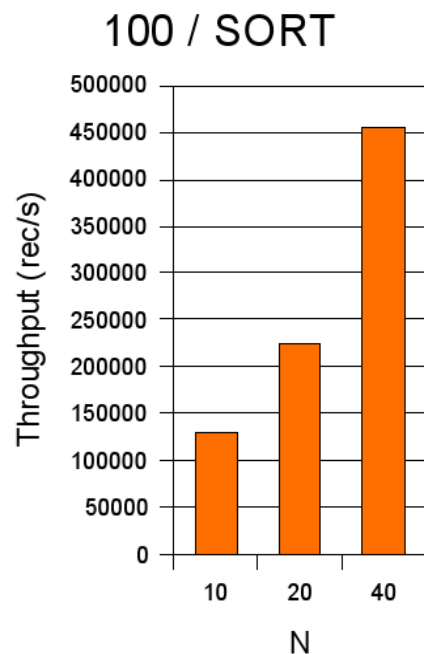
“Cloud computing is a style of computing in which dynamically scalable and often virtualised resources are provided as a service over the Internet”

The Cloud : types

- **Infrastructure as a Service (IaaS)** : Amazon EC2, Eucalyptus, GoGrid, Nymbus, ...
- **Platform as a Service (PaaS)** : Amazon Web Services, AppJet (ex Google guys), Azure Services Platform (MS), Force.com (SalesForce), ...
- **Software as a Service (SaaS)** : SalesForce, Google, Lucidera and *many* others

The Cloud : white paper

- Bayon Technologies (Pentaho Partner)
- <http://www.bayontechnologies.com>
- Sorting 600M line-item rows from TCP-H



The Cloud : Demo time

- Start a dynamic cluster on EC2
- Job:
 - Execute DDL on all slave servers
- Design a first transformation :
 - Load a file on the cloud in parallel in 10 MySQL dabases in 10 tables, split the rows

The Kettle Cloud : links

- My blog : <http://www.ibridge.be>
- <http://wiki.pentaho.com/display/EAI/Dynamic+cluster>

Q & A

- Our homepage: <http://kettle.pentaho.org>
- Our Forum: <http://forums.pentaho.org/forumdisplay.php?f=69>
- Our case tracker: <http://jira.pentaho.org/browse/PDI>
- Our wiki : <http://wiki.pentaho.org/>
 - <http://wiki.pentaho.com/display/EAI>
- Our IRC Channel: ##pentaho (on Freenode)
- Developers mailing list:
<http://groups.google.com/group/kettle-developers>
- My humble blog: <http://www.ibridge.be>
- My coordinates: mcasters@pentaho.org