
# Crazy NoSQL Data Integration with Pentaho

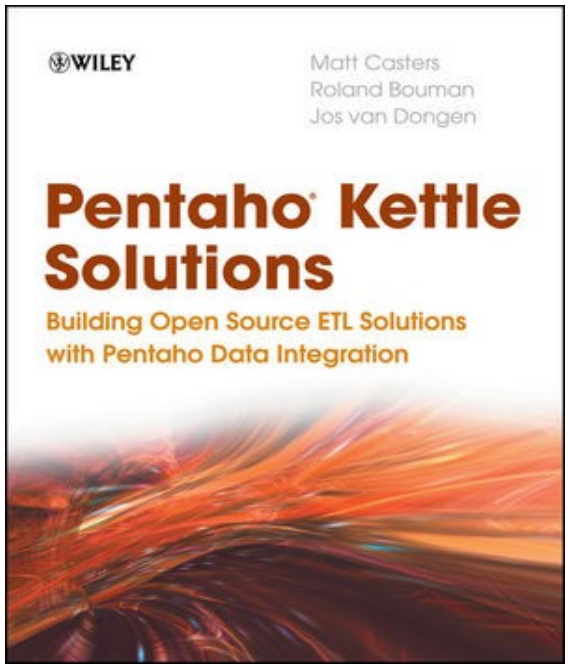
NoSQL Matters, Cologne Germany

May 30th, 2012

Matt Casters

# About Matt

- Chief of Data Integration at Pentaho
  - Lead Development
  - Project manager
  - Community contact
- Kettle Project Founder
- Author of Pentaho Kettle Solutions
  - Published by Wiley
  - 650 pages



# Pentaho Mission

**Delivering the future of analytics today: modern, unified data integration and business intelligence platform**

- Full business analytics & data integration
- Easy to incorporate big and diverse data
- Embeddable, cloud-ready analytics

Open source development model enables **fast and broad innovation**

**One Download Every**

30 Seconds

# A modern, unified embeddable platform built for the future of analytics, including big data and cloud-ready analytics

## ACCESS
*All Enterprise Data Sources*

- Relational Data Sources
- ERP / CRM / Enterprise Apps (e.g. SAP, Oracle)
- Cloud (e.g. Salesforce, Amazon, Dell)
- Hadoop & NoSQL Data
- Unstructured & semi-structured (XML, Excel, Files, etc.)

## STREAMLINE
*Information Delivery*

- Direct Access

**INTEGRATE, CLEANSE, & ENRICH DATA**

- Graphical ETL Designer
- Enterprise Scalability
- Hadoop Clustering

**METADATA LAYER**

- Relational OLAP Cubes
- In Memory Caching
- High Performance

- Data Integration

## VISUALIZE
*& Report Information In Any Style*

**REPORTING**

- Interactive
- Operational
- Enterprise

**ANALYSIS**

- Ad hoc Exploration
- Multi-Dimensional

**DASHBOARDS**

- Interactive Metrics
- Rich Visualizations

**DATA MINING**

- Advanced & Predictive Analytics

## DELIVER
*When & Where Users Need It*

**STANDALONE**

- Web
- Mobile
- E-Mail
- Print

**EMBEDDED**

- ISV & Packaged Applications
- SaaS / Cloud Applications

**CENTRAL ADMINISTRATION, AUDITING & MONITORING**

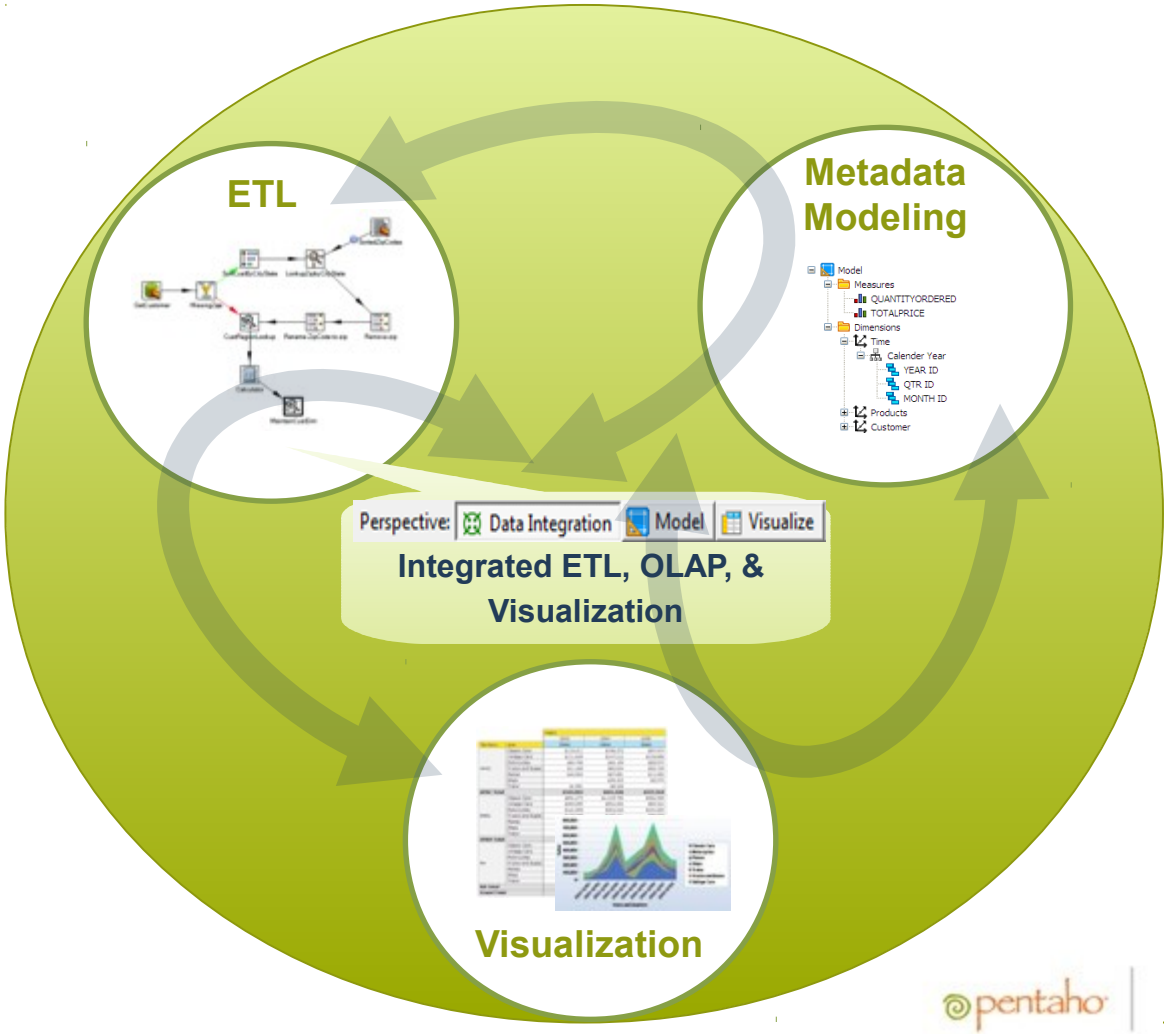
# Better Together: Purpose Built for Business Analytics

**Dramatically reduces time spent to get to useful insights**
**"data source to dashboard"**

### Graphical ETL Designer
- Integrate your data sources
- Auto-generate "Data Models" & "Data Visualizations" with just a right-click button

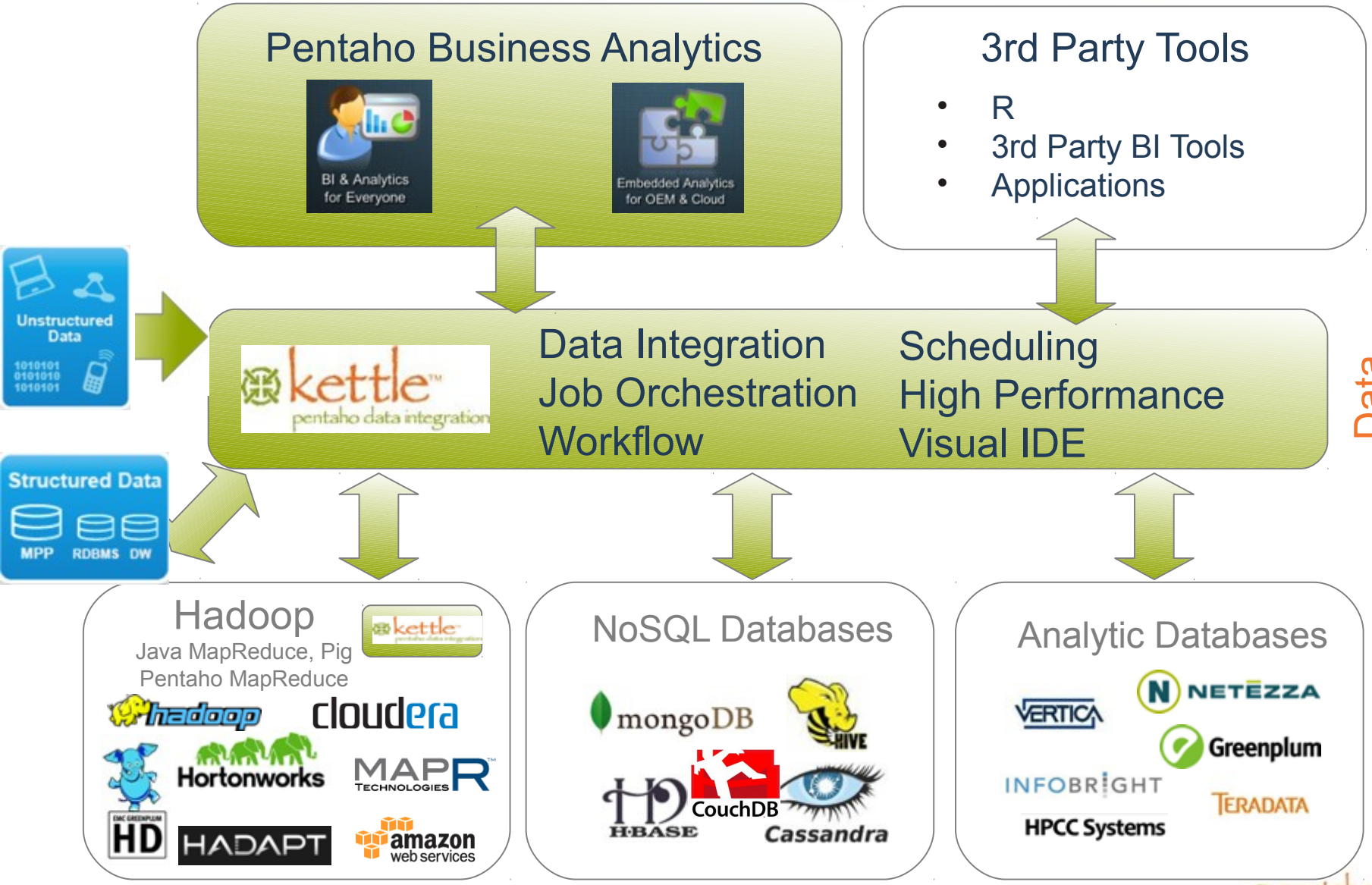
### Multi-Dimensional OLAP
- Model is auto-generated dynamically
- Tweak & automatically visualize results

### Interactive Data Analysis & Reporting
- Auto-generated dynamically
- Slice and dice until data is fit for purpose
- Iterate changing the model and integration flows
- Visualize results immediately




pentaho

# Pentaho in the Big Data Fabric



# Data Sources

Using the right tool for the job...

In what is the worlds data stored?

# Relational Data Sources

- Relational database examples:
  - *Apache Derby, AS/400, Borland Interbase, Calpont InfiniDB, dBase III-IV-5, ExtenDB, Firebird SQL, Generic database, Greenplum, Gupta SQL Base, H2, Hypersonic, IBM DB2, Infobright, Informix, Ingres, Ingres VectorWise, Intersystems Cache, KingbaseES, LucidDB, MaxDB (SAP DB), MonetDB, MS Access, MS SQL Server, MySQL, Native Mondrian, Neoview, Netezza, Oracle, Oracle RDB, PostgreSQL, Remedy Action Request System, SAP ERP System, SQLite, Sybase, SybaseIQ, Teradata, UniVerse database, Vertica*
- Split up into different categories:
  - **ISAM** like xBase variants, SQLite
  - **Relational** like MySQL, PostgreSQL, Oracle, SQL Server
  - **Columnar** like InfiniDB, InfoBright, LucidDB, VectorWise, Sybase IQ
- **All different** in support for the SQL standard
- All aiming for various use-cases
- All have specific advantages and disadvantages.

# Text File Data Sources

- “Comma Separated Values”
- Fixed Width
- Varied width (Cobol redefines)
- Complicating matters:
  - (binary) delimiters, line separators, enclosures
  - Single, double, triple & quadruple byte encoding
  - Various compression formats (.z, .gz, .zip, .rar, ...)

# Other File Data Sources

- XML files : any kind of content
- INI files : key-value pairs
- Properties files : key-value pairs
- xBase files : rows of data
- JSON
- SOAP/WSDL
- MS Access : binary ISAM
- Spreadsheets: xls, xlsx, ods
- LDAP / LDIF
- SAS
- YAML

# Non-relational Data Sources

- SAP R/3 application server
- LDAP
- Message Queues (JMS, MQSeries, ...)
- HL7
- EDIFACT
- Windows Messaging
- SNMP
- ...

# Cloud Data Sources

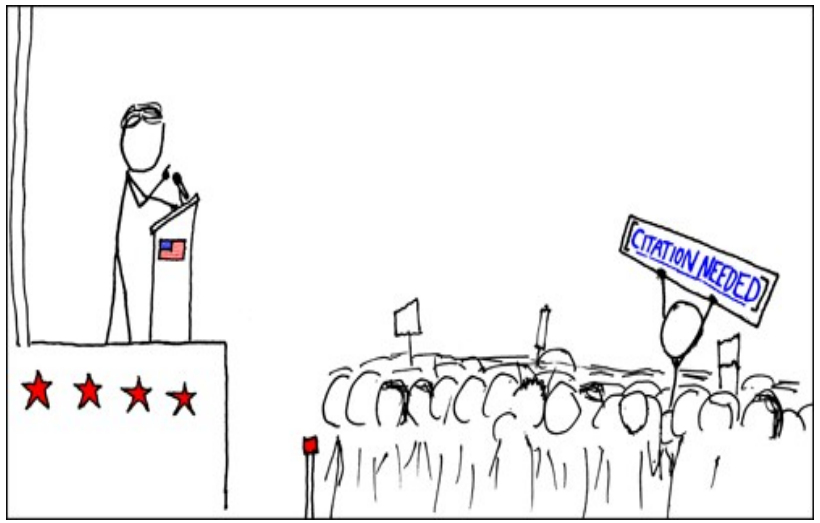
- SalesForce, SugarCRM, ...
- Amazon S3
- Google Analytics
- Twitter
- ...

→ **All different** in format without a standard

→ Various Web Services and authentication methods

# NoSQL Data Sources

- Large and diverse group of different types of data sources
- Document store
  - Apache CouchDB, Jackrabbit, MongoDB, SimpleDB, ...
- Graph databases
  - Neo4J, HyperGraphDB
- Key Value stores (cont.)




*http://en.wikipedia.org/wiki/Wikipedia:Citation_needed*

# NoSQL Data Sources (more confusion)

- Key-Value stores:
  - **Eventually consistent**: Apache Cassandra, Dynamo, Project Voldemort
  - **Hierarchical**: GT.M
  - **Hosted**: Freebase
  - **RAM Cached**: memcached, Velocity, Citrusleaf
  - **Disk stored**: BigTable, MemcacheDB, Citrusleaf, MongoDB
  - **Ordered**: Berkeley DB, Informix C-ISAM, MemcacheDB, NDBM
  - **Multivalue**: Intersystems Caché, ESE/NT, OpenQM
  - **Object databases**: Caché, JADE, ObjectDB, ObjectStore, db4o
  - **Tabular**: BigTable, Apache Hadoop, Apache Hbase, Hypertable, Mnesia
  - **Tuple store**: Apache river

**Source**: http://en.wikipedia.org/wiki/NoSQL

# Other Encountered Data Sources

- Home brew data formats...
  - Legacy stuff
  - Because there just aren't enough options out there!
- Non-structured data formats
  - PDF
  - Images
  - Video
  - Web content
  - …

# But...

Now the integration goes crazy!

# Adding complexity by mixing data sources

- Reading from … and writing to ...
- Performing lookups
- Joining data
- Calculating
- Scripting
- ...

EXPLODING COMPLEXITY

- X systems, Y architectures, Z operations

- Trashing it all when new technology surfaces!

# Classical Solutions...

- Standardize on software stacks
- Limit the number of talking systems

Good advice but...

Perceived as **limiting**, **stifling** and **preventing innovation**.

# Consequences!

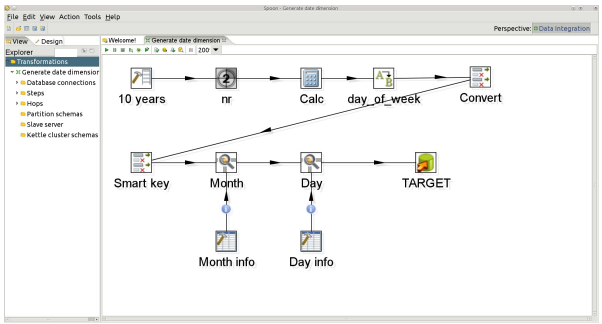
- Sooner or later you're going to be using any of these data sources
- ... to read from or to write to
- ... with more other data sources coming after it
- Get ready to learn more and more APIs
- Rate of change seems to be increasing

# Solving it the Kettle way... with abstraction layers

Task design

Save work

Execution




***Java API***

**`XML generation`**

**`<XML/>`**

***Repositories***

Graphical

Batch

Local

Remote

Java API

M/R

Reporting

Embedded

# Demo

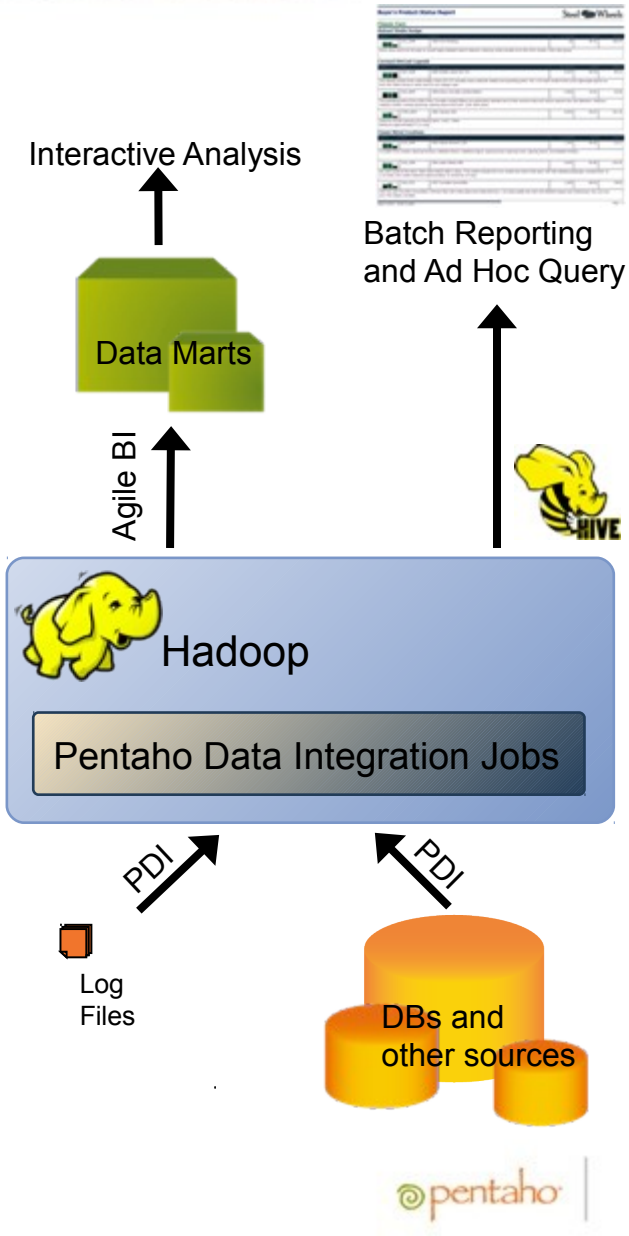
- The design environment
- Load some data into MySQL

# What about NoSQL & Big Data

- Big Data “lakes”
- Danger of creating isolated data silos
- Connecting with classical tech is important
- Similar abstraction solutions are needed

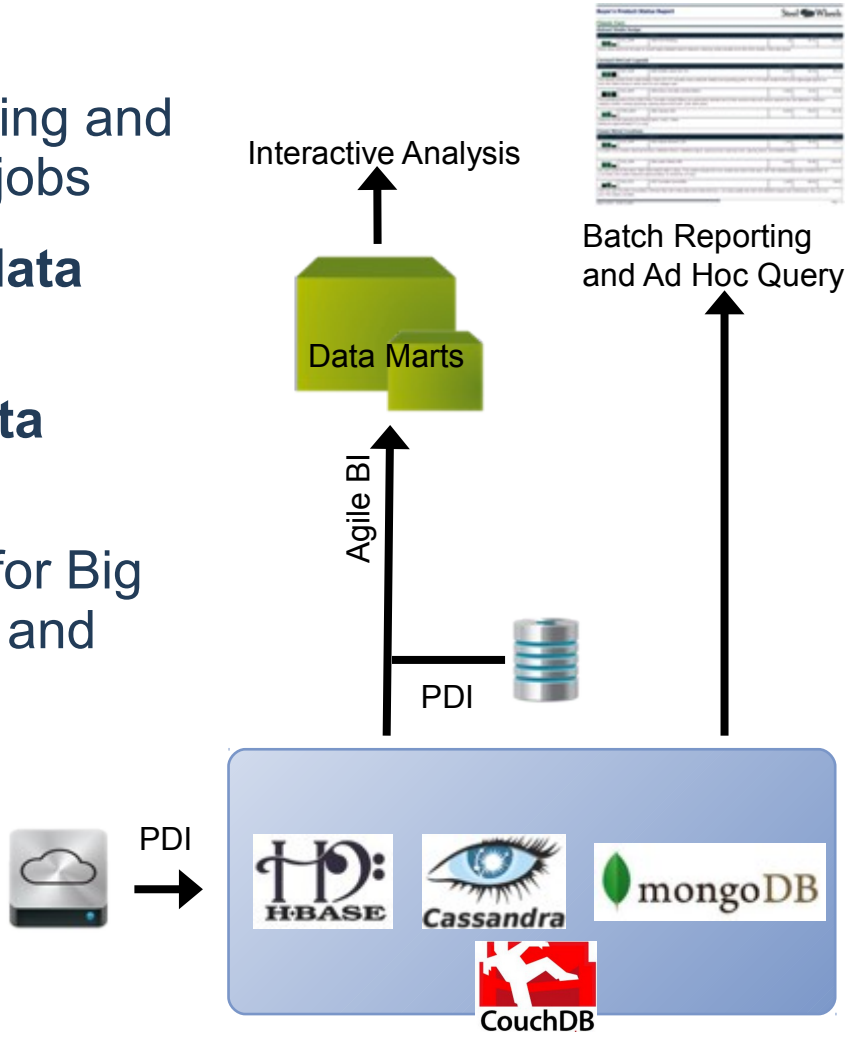
# Pentaho for Hadoop – Value Proposition

- **Lowers technical barriers** through a graphical design environment for creating and managing MapReduce jobs
- **Automatic ETL scalability** through deployment across the Hadoop cluster
- **Easily integrate external reference data** with data from Hadoop
- **Easily spin-off high performance data marts** for interactive analysis
- **Provides an end-to-end BI solution** for Big Data including reporting, ad hoc query and interactive analysis



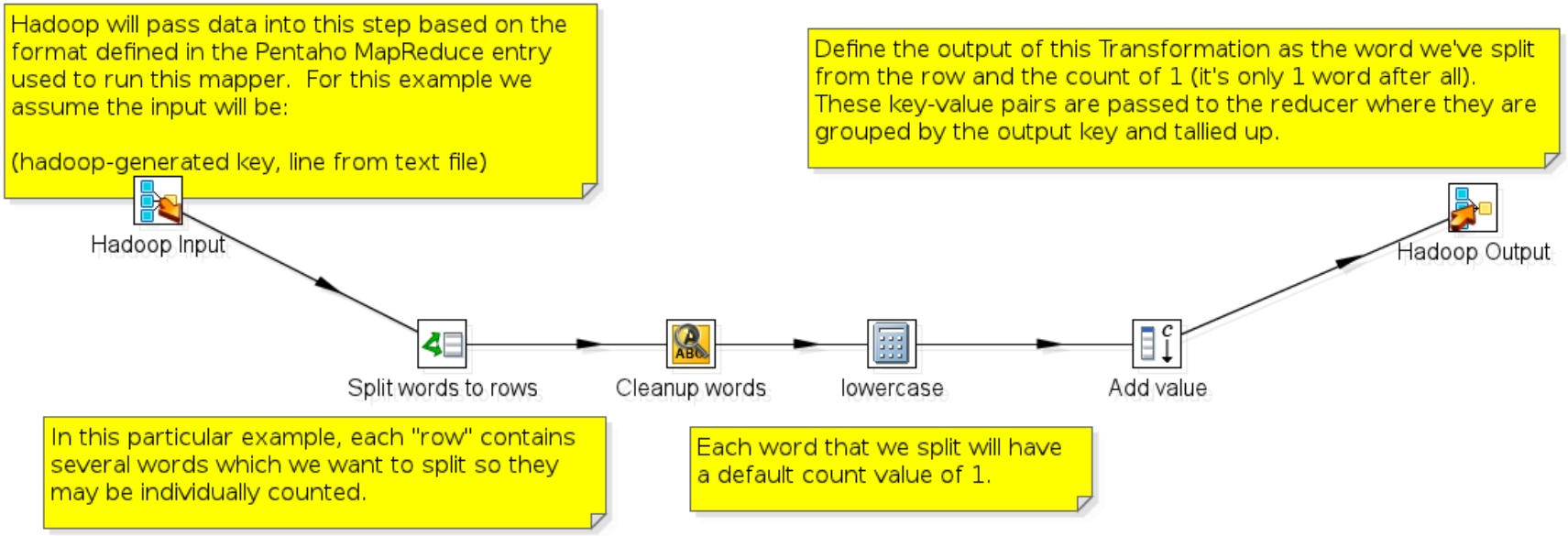
# Pentaho for NoSQL – Value Proposition

- **Lowers technical barriers** through a graphical design environment for creating and managing NoSQL load and extraction jobs
- **Easily integrate external reference data** with data from NoSQL
- **Easily spin-off high performance data marts** for interactive analysis
- **Provides an end-to-end BI solution** for Big Data including reporting, ad hoc query and interactive analysis



# Demo

- Execution of a Hadoop **Map/Reduce** job
- Retrieval of wordcount results, **copy to MySQL**
- Parallel execution of **partitioned MySQL table copy** to HDFS
- Insert MySQL data into MongoDB and CouchDB
- Read data from MongoDB & CouchDB and bulk load into MySQL



# Open Source Software

- Complete stack of BI and Data Integration software
- Software under Open source license(s)
  - *Kettle 4.3 is Apache Licensed*
- Free to download and use by each and everyone
- Projects can be integrated or separately used


**Don't re-invent the wheel!**

**Spent time on the fun stuff!**

# Pentaho Kettle for Big Data Benefits


- ✓ Visual tools delivers 10x boost in productivity for developers
- ✓ Makes big data platforms usable for a huge breadth of developers
- ✓ Enables easy visual orchestration of big data tasks
- ✓ Fully leverages the full capabilities of each big data platform
- ✓ Provides an easy on-ramp to Pentaho Business Analytics


# Thank You

## Join the conversation. You can find us on:


http://blog.pentaho.com


@Pentaho


Facebook.com/Pentaho


Pentaho Business Analytics

http://www.ibridge.be

@mattcasters