

# DataConference 2013



**FUELING  
Commerce**

July 23 - 25

ebay inc

## **Big Data Quality: Just a New Buzzword or Serious Topic?**

David Pejcoch

July 25, 2013

# Outline



- What is Big Data?
- What are opportunities for eBay in Big Data world?
- What is Data Quality Management
- Does eBay have problems with Data Quality?
- What are specifics of Big Data Quality?
- Four viewpoints for Big Data Quality
- Support for Big Data Quality Management in tools

# What is Big Data

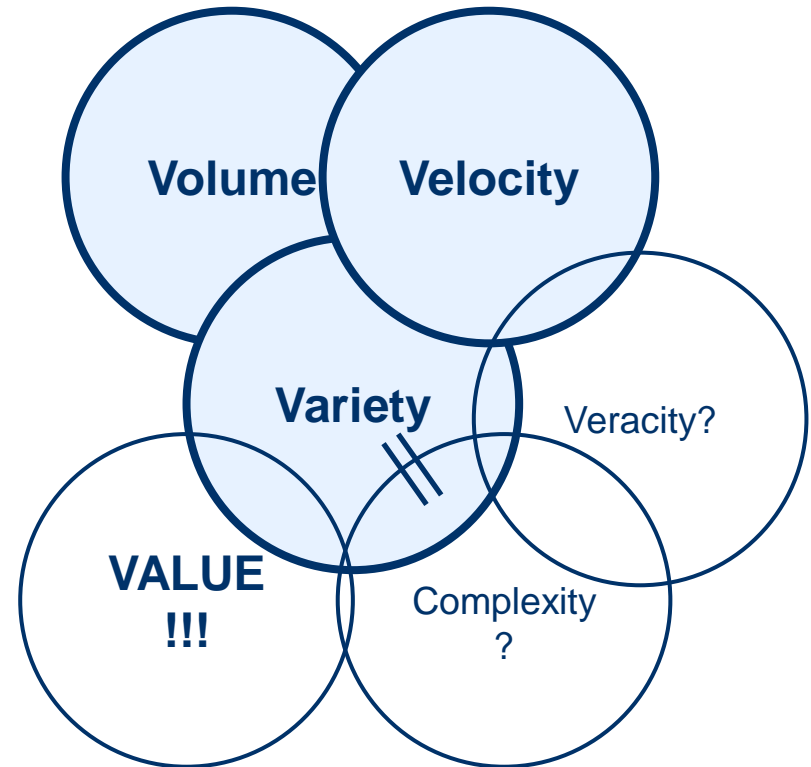
**DataConference 2013**

# Big Data

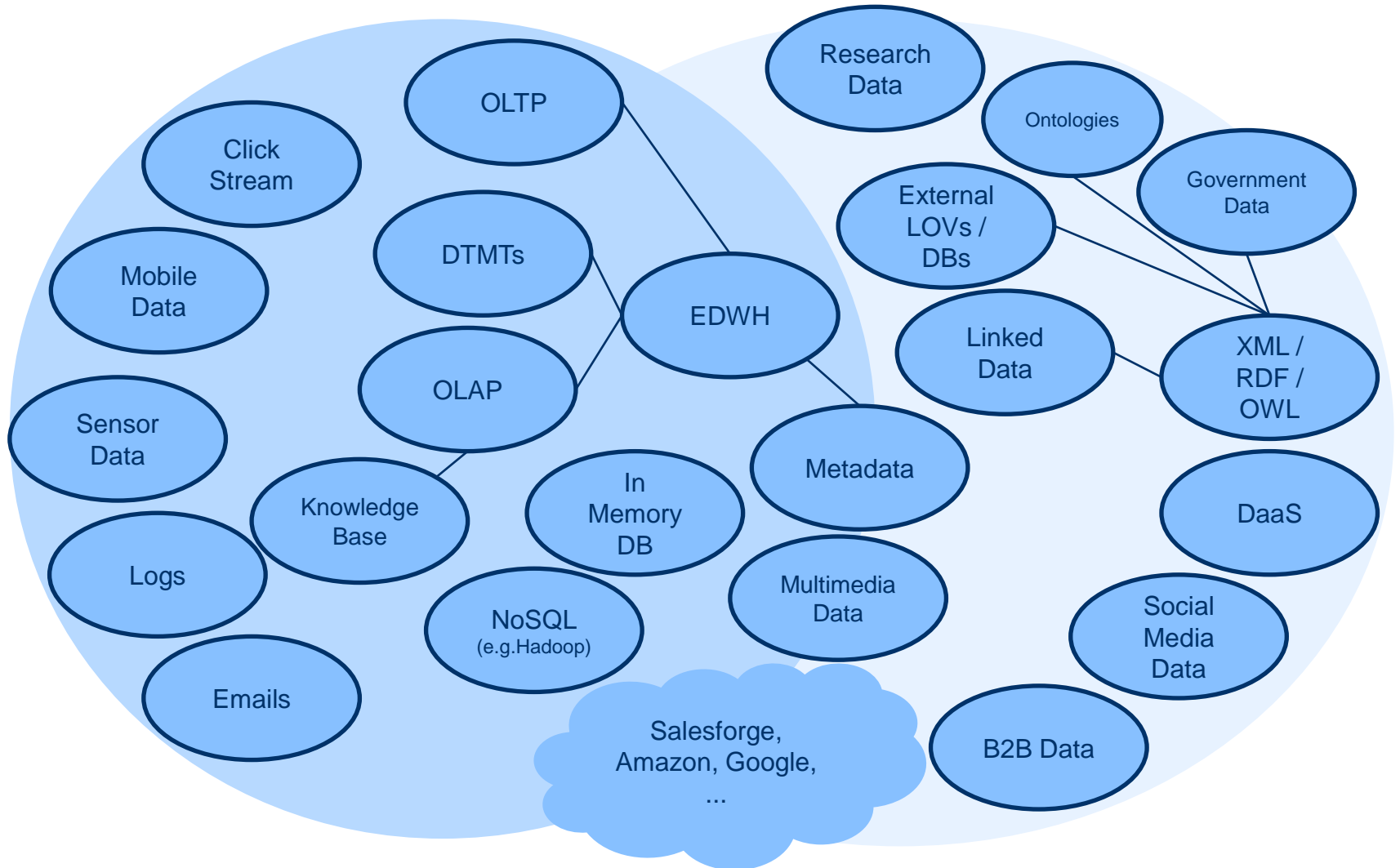


- At the beginning Man created „Small Data“
- How big data must be to become „Big Data“. Is Teradata big enough?
- S. Sarsfield: Small / Middle / Big Data
- Gartner: „Year 2013 is the year of Big Data“ => The Future?
- Hadoop (Google Map Reduce + Big Table), Hive, Pig, Zookeeper, Sqoop, ...
- 1st Stage: Big Data = Hadoop
- Sources of Big Data = bigger than Big Data

## Modified Gartner Big Data Definition



# Sources of Big Data



# Does eBay have Big Data?



# Does eBay have all Big Data sources what we need?



**CLASSIFIED**

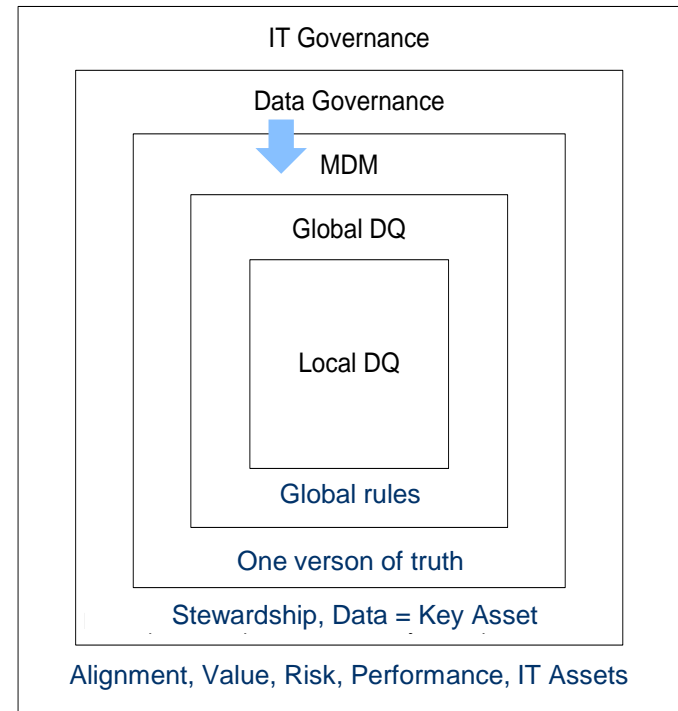
# What is Data Quality



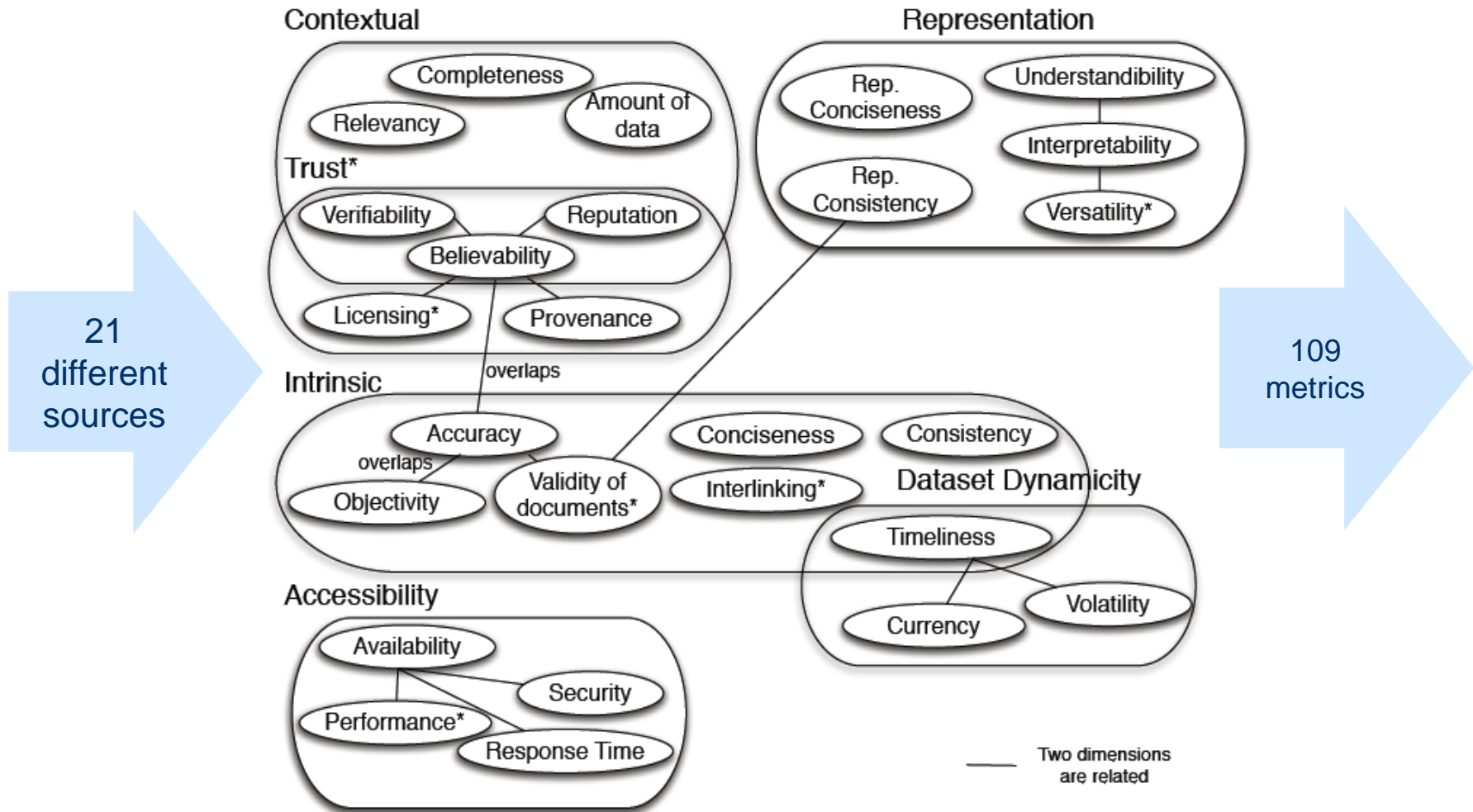


- Data are of high quality "if they are fit for their intended uses in operations, decision making and planning" (J. M. Juran). ... a lot of definitions. Practically all of them refer to some characteristics which are measured.
- Data Governance: data as an asset, principles, politics, rules, ownership (stewardship), necessary condition for MDM
- Focus on data lineage
- Modern approach is proactive instead of reactive
- Process analyses instead of technical assessment

## Hierarchy of DQM



# Zaveri: Data Characteristics



# What makes sense?



Dimensions	Data Characteristics
Intrinsic Dimension	Reliability
	Uniqueness
	Semantic Accuracy
	Syntactic Accuracy
Time Dimension	Currency
	Timeliness
	Volatility
	Time Synchronization
Contextual Dimension	Consistency
	Completeness
	Coverage
Dimension of Usage	Availability
	Comprehensibility
	Interoperability
	Security of Use
Economical Dimension	The cost of acquiring and updating data
	The costs of storing, sharing, distribution, bckp and archiving
	The costs of data protection

# Common DQM Techniques



- Data Quality Assessment:
  - Technical Profiling (pattern analysis + EDA)
  - Verification / Validation: syntax, LOVs, checksums, business rules (consistency), constraints (integrity + allowed values)
  - Root-cause analysis
  - Analyses of implemented controls
  - Process Analysis
- Unification / Standardization: schemas, rules
- Deduplication: clustering, fuzzy / crisp match-merge
- Imputation + Enrichment: using models (explicit / implicit), single values or external data sources
- Geocoding: linking to external sources
- Householding: identification of relationships among entities
- Stewardship: setting up ownership of data
- Implementation of policies, principles and controls
- Permanent Monitoring: business rules

# BP: Using DQ Knowledge Base



- Common „Semantic Data Element“ => Grammars, Syntax rules, LOVs, expected values, ... , business rules, additional knowledges
- Usage: data profiling, monitoring, standardization, validation, .... practically all steps in DQM cycle
- CDM (Common Data Model, Canonical Data Model, ...) usually used in online integration as a data model independent on individual application
- Examples of CDM:
  - ACORD (Association for Cooperative Operations Research and Development) for the insurance industry,
  - SID for telecommunications,
  - CIM (Common Information Model) for public services,
  - PPDM and MMDM for energetic industry,
  - OAGIS (Open Application Group Integration Specification) for production and supply chains,
  - HL7 (Health Level Seven International)
  - HIPAA for healthcare, ARTS (The Association for Retail Technology Standards) for sales and finally FPML and SWIFT for capital markets.

# BP: Using DQ Knowledge Base



- Forrester:

- 58% of respondents answered they use a conventional tool for Enterprise Architecture modelling,
- 21% of them use the modelling tool that is part of its SOA / BPM (Business Process Management) solution,
- 4% use the tool centred on XML schema
- 17% don't use any tools.
- No respondent considers semantic technologies such as RDF (Resource Description Format) or OWL (“Web Ontology Language “) as suitable solution for modelling and managing CDM

# Does eBay have problems with DQ?



**CLASSIFIED**

# What are the specifics of Big Data Quality?

Big Data = Small Data + Big Garbage?



# 4 Viewpoints for Big Data Quality



- Quality of Big Data
  - Retrospective: DQ in Hadoop
  - Proactive: DQ Management of Big Data sources
- Big Data as a source for Data Enrichment
- Big Data as a source for Data Validation
- Big Data techniques as a tool for effective Data Quality Management

# Potential Problems of Big Data sources



- We can find inspiration in Linked Data Quality (older topic)
- Especially:
  - Integrity
  - Compliance (is PayPal a bank? => Basel II/III)
  - Availability of Big Data
  - Security and Privacy (e.g. Facebook data)
  - Consistency between different sources
  - Lineage (History, Licence, Sustainability, ...)
  - Quality of architecture / metadata
  - Identity identification
  - Boundedness <= only relevant data with business reason
- Limitation of Map-Reduce:
  - Cleansing and transformation within single Map operation
  - Profiling & Matching of unstructured data
  - Matching of data in operations without inter-process communications
- Multimedia data quality
  - Deduplication (comparison of bit segments)
  - Consistency (proper images assigned to proper items)



# Proactive DQM: Complex Master Data Management

- Different sources of Master Data
- More requirements to MDM Hub
  - Communication with DaaS
  - Structured / Semistructured / Unstructured data
  - SOA
  - Canonical Data Model adopted within ESB and QKB
- However:
  - More agile methodologies of implementation needed
  - Risk of very, very long and expensive implementation
  - Implementing „Space Program“ where it is not necessary

# Proactive DQM: Faster Matching



- Phonetic Algorithms: don't work + depends on national environment (Soundex for English, Daitch-Mokotoff for German and Slovenian languages)
- Similarity metrics (token-based, costs-based) are inefficient
- Simple Match Codes (extra-fast, stored / indexed) don't respect semantic meaning
- Complex Match Codes based on QKB (Quality Knowledge Base) with predefined sensitivity => too much new attributes in Big Data from different domains and sources => coordination among Data Stewards
- Blocking Strategy
- Machine Learning for automatically building match classifiers
- Similarity metrics + Match Codes don't work with DaaS
- Optimum: combination of all of them

# Proactive DQM: Examples of Match Codes



String	Using QKB	Without QKB
Jim Goodnight	F8B~\$\$\$\$\$\$C&B_4\$\$\$\$	CB4F8P~\$\$\$\$\$\$
Jim Goodbride	F8MY~\$\$\$\$\$\$C&B_4\$\$\$\$	CB4F8MY~\$\$\$\$\$\$
James Goodnight	F8B~\$\$\$\$\$\$C&B_4\$\$\$\$	CB4F8P~\$\$\$\$\$\$
James Good Knight	P~\$\$\$\$\$\$C&B_4\$\$\$\$	CB4F83P~\$\$\$\$\$\$
Jim Goodnite	F8B~\$\$\$\$\$\$C&B_4\$\$\$\$	CB4F8P~\$\$\$\$\$\$
Good Night James	CB4\$\$\$\$\$\$F~\$\$\$\$\$\$	F8P~CB4\$\$\$\$\$\$
Jim Nightgood	P~F~\$\$\$\$\$\$C&B_4\$\$\$\$	CB4P~F~\$\$\$\$\$\$

Knowledge: the string most probably consists of First Name and Last Name



# Proactive DQM: More Complex Data Governance Strategy

- Soares: „Big data is part of a broader information governance program“
- Too many different data sources => how to assign Data Stewards (x-domains, x-sources, x-technologies) => Steward = Data Scientist?
- All data should be governed but one governance strategy doesn't fit to all data
- Some data could be worse than another => „*Hadoop data could be bad*“
- Different approaches according to usage: Compliance, Risk, CRM, ... => Bus Matrix needed
- Linking data to business value
- No single version of truth
- Missing connection to central metadata
- Information Lifecycle Management: delete / archive Big Data which are not used

# Reactive DQM



- Using standard DQM tools: Talend, SAS, Informatica, ..., Datameer
- Map-reduce DQ: spaghetti code
- Hive:
  - Hive QL = „SQL“: projection, equi-joins, group by, sampling, order by, ... not enough
  - inclusion of e.g. Python code to MapReduce
- Cloudera: Impala



# Hadoop as a DQM tool: Dedooop

- Still the prototype => Use on your own risk
- Compatible with Hadoop 0.20.2 and Debian-based OS
- Servlet container for Apache Tomcat => web based interface
- Blocking based entity matching in parallel
- Automatically transforms the specification into MapReduce workflow
- Several map tasks, several reduce tasks based on blocking key
- Multiuser system
- Load balancing strategies
- Graphical HDFS and S3 file manager
- University of Leipzig
- Link: <http://dbs.uni-leipzig.de/dedooop>

The screenshot displays the Dedooop web interface for defining a workflow. The 'Workflow Definition' section is active, showing the following configuration:

- DBLP\_id**: [Dropdown]
- Attribute Mapping**: DBLP\_title: [Dropdown] GS\_title, DBLP\_authors: [Dropdown] GS\_authors
- Normalize attribute values
- Output Directory**: hdfs://master/output

The 'Blocking' section is also visible, showing:

- Blocking Strategy**: Standard Blocking (BlockSplit)
- Collect skew metadata
- Key**: TokenizingBlockingKeyOf
- Generator**: [Dropdown]
- Tokenizer**: LuceneTokenizer
- Avoid redundant computation
- Tokens the blocking attribute. Each token is regarded as a blocking key for this entity.

The 'Data Source definition & File Viewer' section shows a table of data sources:

Data Source	Size		
hdfs://master/input_data/DBLP.txt	362.37KB		
hdfs://master/input_data/GoogleScholar.txt	8.83MB		
GS_id	GS_title	GS_authors	GS_year
0H8M-YLH4BJ	Too Much Middleware	M Stonebraker	2002
rgzK3G-mGJ	A Correctness Proof for a Practical Byzantine-Fault-Tolerant Replication	M Castro, B Liskov	1999
r3cCE4v4GJ	On a stochastic optimization algorithm using IPAs which updates after evi	EKP Chong, PJ Ramadge	
7B7KCLJu4BJ	Flight to Objectivity: Essays on Cartesianism and Culture	S Bordo	
wGT0B7miLJY	Capturing Design Rationale in Concurrent Engineering Teams	M Klein	
d20IKJZ3z0J	Chromosome banding in X-linked mental retardation	C Fonatsch	



# Hadoop as a DQM tool: Dedoop



**Dedoop - Efficient Deduplication with MapReduce**

Experiment 1 + Expert Mode

**Hadoop Cluster**

Running Cluster Launch EC2 Cluster

Namenode:

Jobtracker:

WebUI port:

Disconnect

**Workflow Definition**

DBLP\_id:

Attribute Mapping: DBLP\_title:

DBLP\_authors:

Normalize attribute values

Output Directory:

**Blocking**

Blocking Strategy:

Collect skew metadata

Key Generator:  Attributes:  Tokenizer:

Avoid redundant comparisons

Tokenizes the blocking attribute. Each token is rearded as a blocking key for this entity.

**Data Source definition & File Viewer**

Data Source	Size
hdfs://master/input_data/DBLP.txt	362.37KB
hdfs://master/input_data/GoogleScholar.txt	8.83MB

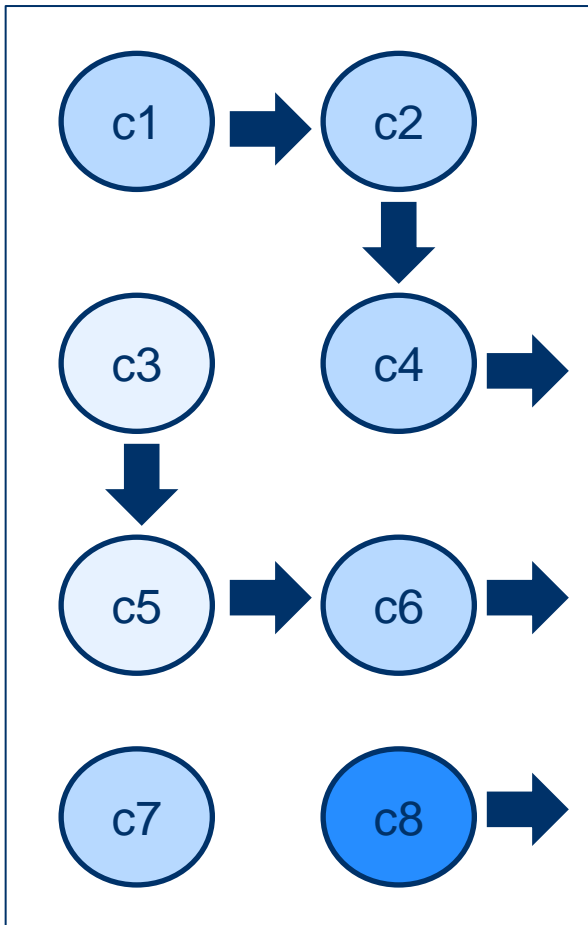
GS_id	GS_title	GS_authors	GS_year
0HMk-YUh4i8J	Too Much Middleware	M Stonebraker	2002
rgzK3sG-rmQJ	A Correctness Proof for a Practical Byzantine-Fault-Tolerant Replication	M Castro, B Liskov	1999
r3sCE4vukG0J	On a stochastic optimization algorithm using IPA which updates after evi	EKP Chong, PJ Ramadge	
7B7KCnJu4βJ	Flight to Objectivity: Essays on Cartesianism and Culture	S Bordo	
wGTOB7lmLYJ	Capturing Design Rationale in Concurrent Engineering Teams	M Klein	
d2QifUxZ3zQJ	Chromosome banding in X-linked mental retardation	C Fonatsch	

Entity Resolution workflow definition - Selection of the Blocking key generation function

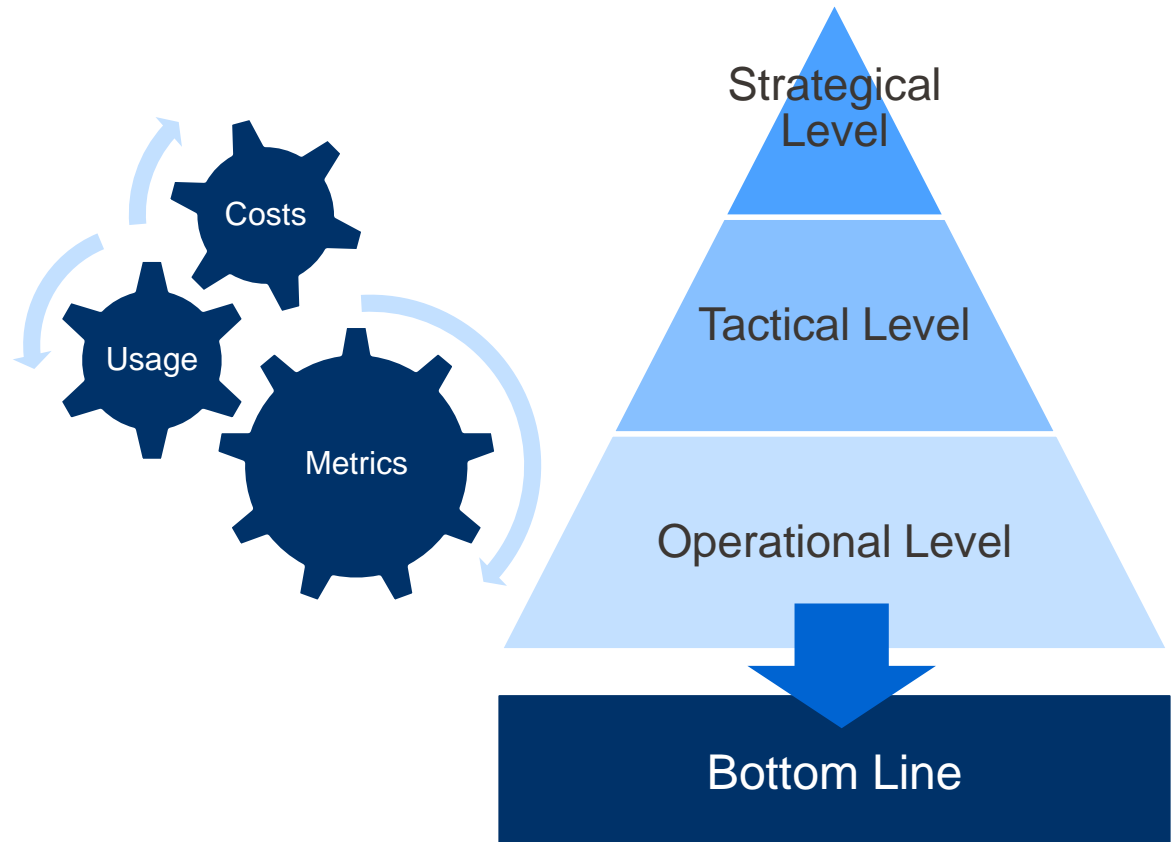
# Proactive DQM: My Causal Simulative Model



## Data Characteristics



## Performance Metrics



# Support for Big Data Quality in DQM Tools

**DataConference 2013**

# Talend Open Studio for Big Data

The screenshot shows the Talend Open Studio for Big Data (5.3.0.r101800) interface. The main workspace displays a job design with three rows:

- row1 (Main):** A 'Load from Oracle' row containing a 'tOracleInput\_1' component connected to a 'tHDFSPut\_1' component.
- row2 (Pig):** A 'Pig Load' row containing three components: 'tPigLoad\_1', 'tPigAggregate\_1', and 'tPigStoreResult\_1', connected in sequence.
- row4 (Main):** An 'Extract to MySQL' row containing a 'tHDFSGet\_1' component connected to a 'tMysqlOutput\_1' component.

The interface includes a menu bar (File, Edit, View, Window, Help), a toolbar with various icons, and several panels:

- Repository:** Shows a tree view of components under 'LOCAL: BigDataTest', including 'Routines', 'Pig UDF', 'SQL Templates', 'Hive', and 'UserDefined'.
- Palette:** A search bar 'Find component...' and a list of 'Big Data' components such as Cassandra, CouchDB, Couchbase, Google BigQuery, HBase, HCatalog, HDFS, Hive, MongoDB, Neo4j, Pig, and Sqoop.
- Outline:** Lists the components in the job: tHDFSGet\_1, tHDFSPut\_1, tMysqlOutput\_1, tOracleInput\_1, tPigAggregate\_1, tPigLoad\_1, and tPigStoreResult\_1.
- Code Viewer:** Currently shows 'Properties not available.'

# SAS



- SAS/ACCESS Interface to „Whatever“ .... e.g. Hadoop
  - Supports Map-Reduce, Pig, HDFS commands, HIVE queries
- Possibility to use standard SAS Tools:
  - EMiner, DI Studio, Data Flux
- Unified data management platform for all data (structured, semi-structured, unstructured)
- Quality Knowledge Base based on Semantic Data Type
  - Could be similar to ESB model based on industry standard

# Data Cleaner



Analysis job | DataCleaner

File Reference data Write data Window Help

Save Save As... Visualize Transform Analyze Execute

Source Metadata

Select datastore for analysis

Create a new datastore:

CSV Excel Access SQL Server dBase XML Oracle Hadoop Greenplum Redshift MySQL PostgreSQL

more

Analyze an existing datastore: Search/filter datastores

**CSV** **Country codes**  
Country reference data, courtesy of Graham Rhind

**orderdb**  
DataCleaner example database

Analyze!

Welcome to DataCleaner 3.5

Offline

more

- H2
- Hsqldb/HyperSQL
- JDBC-ODBC bridge
- Sybase
- Other database
- Manage database drivers...
- Composite datastore

Clustered big data analysis in DataCleaner

YouTube

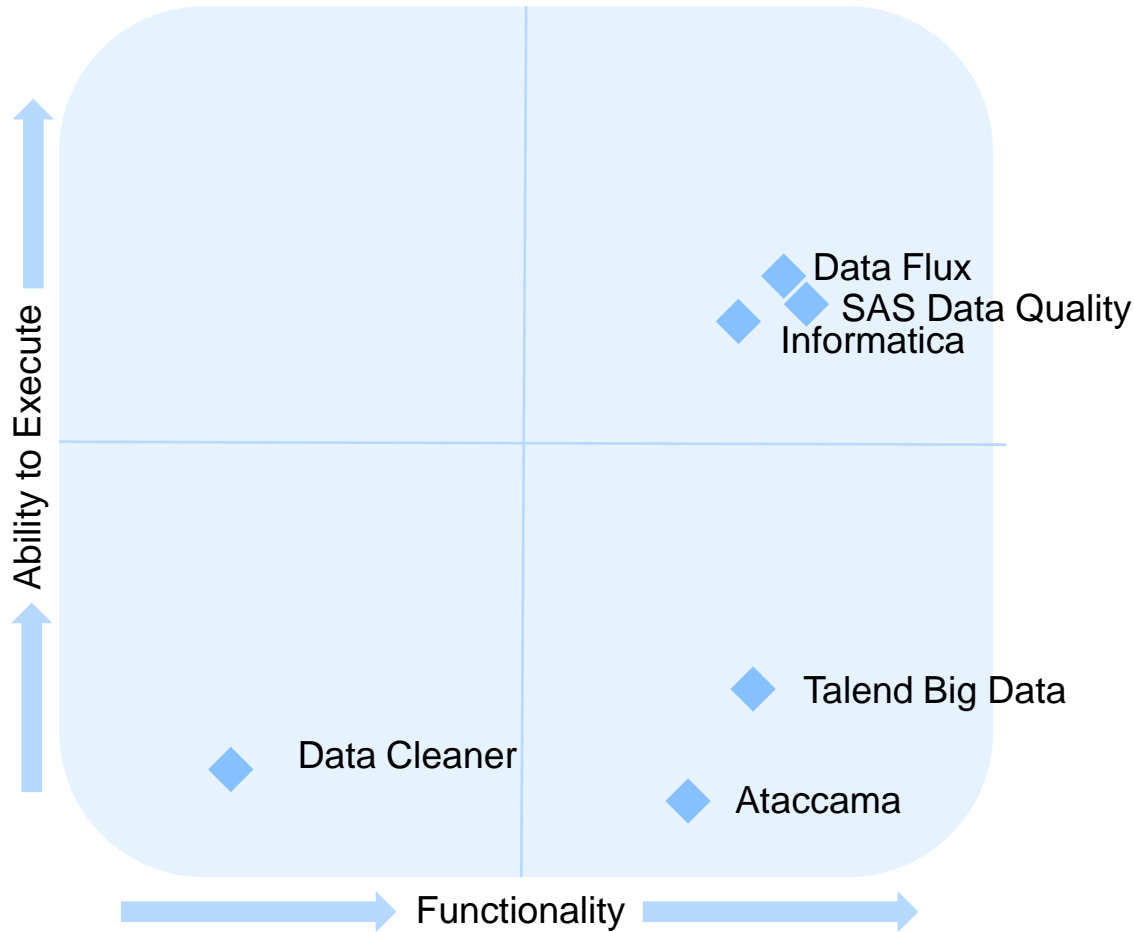
[Raspberry.com]

# Ahoth Tools



- Ataccama:
  - results stored in Hadoop cluster, connection to different sources (Hadoop included)
  - tasks run in cluster using map-reduce
- Informatica PowerCenter Big Data Edition
  - DBMS, OLTP, OLAP, ERP, CRM,
  - mainframe, cloud, and others. You can also
  - access all types of big interaction data, including
    - social media data, log files, machine sensor data,
    - Web sites, blogs, documents, emails, and other
    - unstructured or multi-structured data
- IBM InfoSphere Quality Stage
- Oracle Enterprise Data Quality
- Trillium Software BQuality
- SAP Business Objects Data Quality Management

# „Magic Quadrants“ for Big Data QM Tools



## Methodology:

- Comparison of Gartner Magic Quadrants for DQM and Data Integration
- Results of my own evaluation



... so „New Buzzword“ or serious  
topic?

# Big Data Quality: Does It Matter?



- Old DQM techniques used in complex environment
- Focus on performance of tools and techniques
- Renaissance of sampling
- Old governance styles but applied differently for different groups of data
- Different levels of required quality
- Necessity of metadata
- Complex skills of Data Stewards => hire Data Scientists
- Old tools but with additional integration functionality
- Conclusion: Big Data Quality = Complex Data Quality = set of disciplines previously focused on respective kind of data and now integrated to one big environment

**Any Questions?**

**DataConference 2013**