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

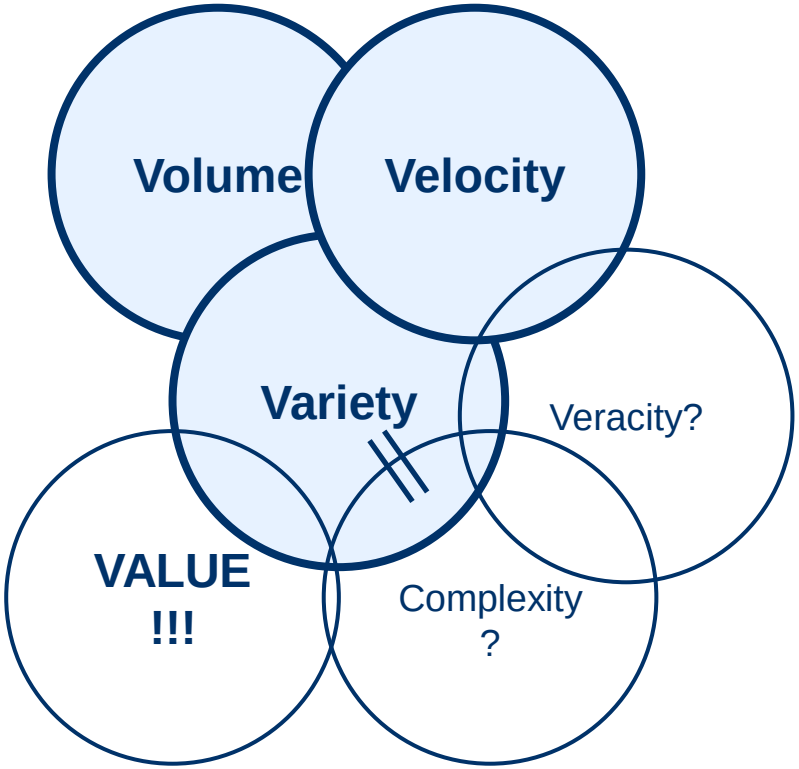
- What is Big Data?
- What are opportunities for eBay in Big Data world?
- What is Data Quality Management
- Does eBay have problems with Data Quality?
- What are specifics of Big Data Quality?
- Four viewpoints for Big Data Quality
- Support for Big Data Quality Managemen in tools

What is Big Data

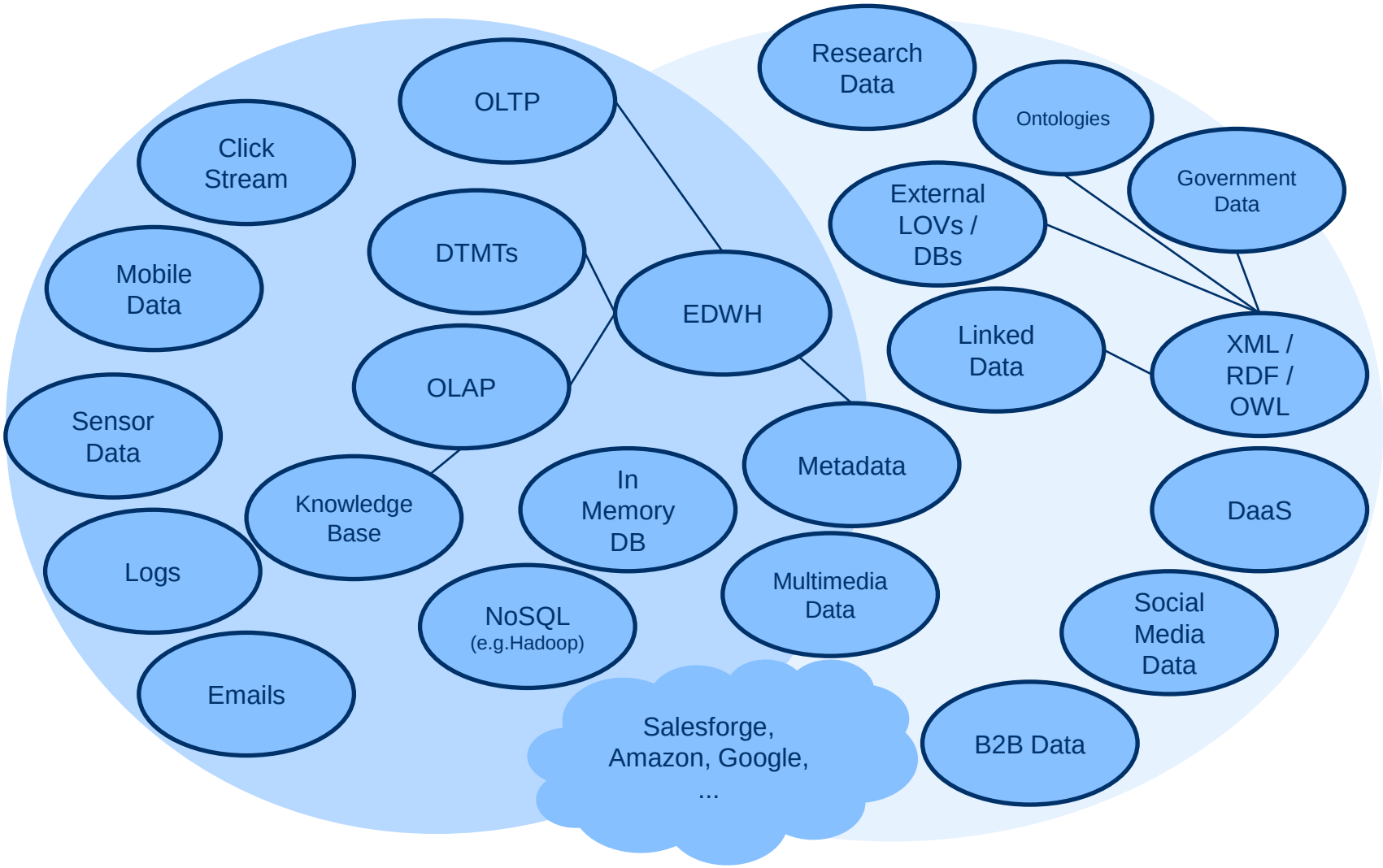
Big Data

- At the begining Man created „Small Data“
- How big data must be to became „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



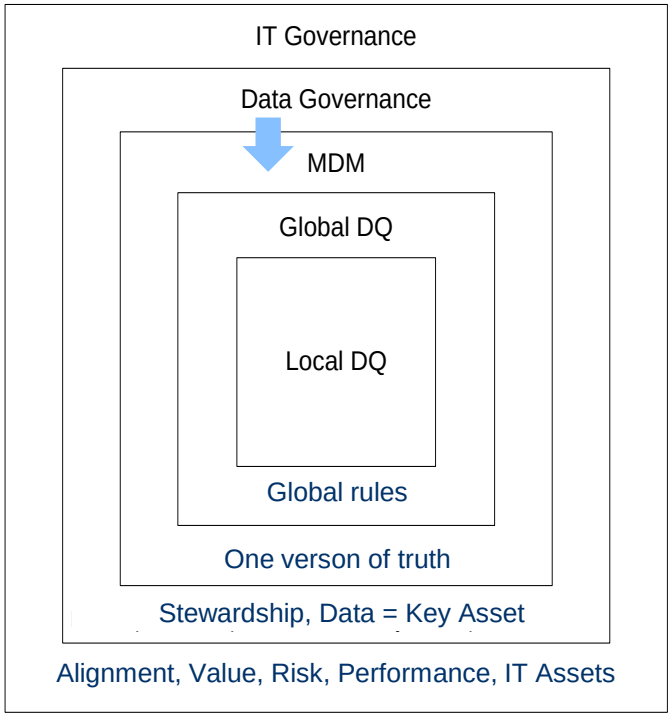
Does eBay have Big Data?

Does eBay have all Big Data sources what we need?

What is Data Quality

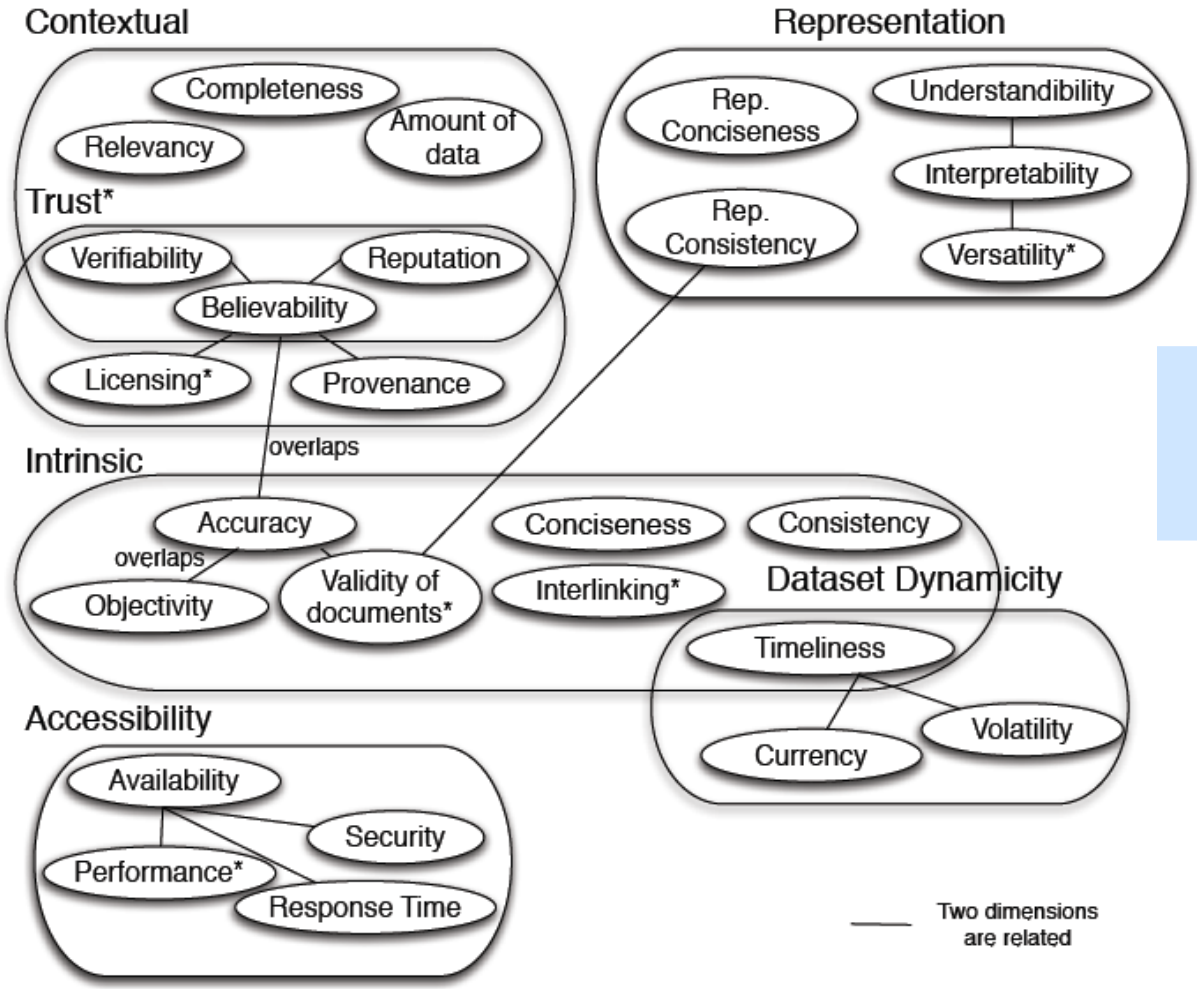
- Data are of high quality "*if they are fit for their intended uses in operations, decision making and planning*" (J. M. Juran). ... a lot of definitions. Practicaly all of them refer to some characteristics which are measured.
- Data Governance: data as an asset, principles, politics, rules, ownership (stewardship), necesary 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

21 different sources



109 metrics

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 profilling, 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?

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 PayPall 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 (comparation 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
  - Cannonical 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 Slovanian languages)
- Simmilarity 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
- Simmilarity 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

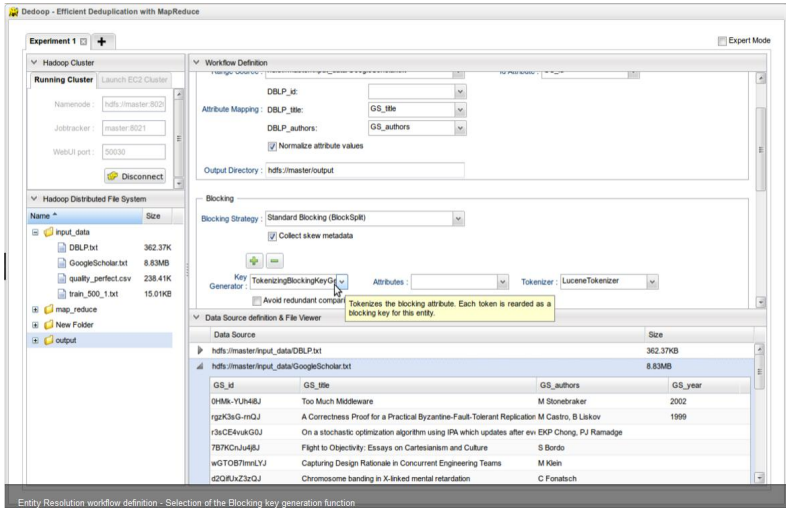
- Soares: „Big data is part of a broader information governance program“
- Too many different data sources => how to assign Data Stewards (x-domains, x-sources, x-technologies) => Steward = Data Scientist?
- All data should be governed but one governance strategy doesn‘t fit to all data
- Some data could 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 conection to central metadata
- Information Lifecycle Management: delete / archive Big Data which are not used

Reactive DQM

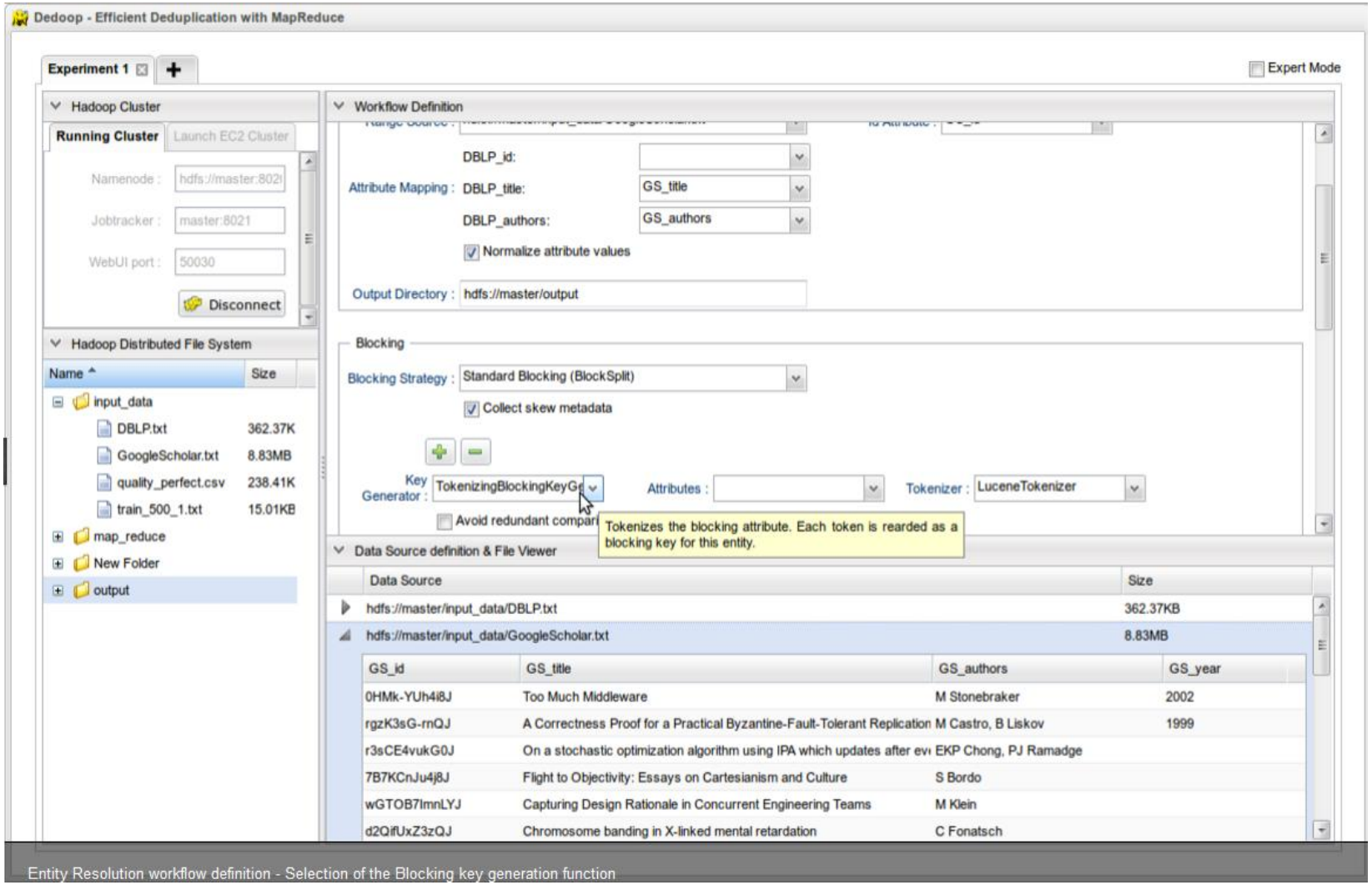
- Using standard DQM tools: Talend, SAS, Informatica, ..., Datameer
- Map-reduce DQ: spaghetti code
- Hive:
  - Hive QL = „SQL“: projection, equi-joins, group by, sampling, order by, ... not enough
  - inclusion of e.g. Python code to MapReduce
- Cloudera: Impala

Hadoop as a DQM tool: Dedoop

- Still the prototype => Use on your own risk
- Compatibile with Hadoop 0.20.2 and Debian-based OS
- Servlet container for Apache Tomcat => web based interface
- Blocking based entity matching in parallel
- Automatically tranforms 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/dedoop

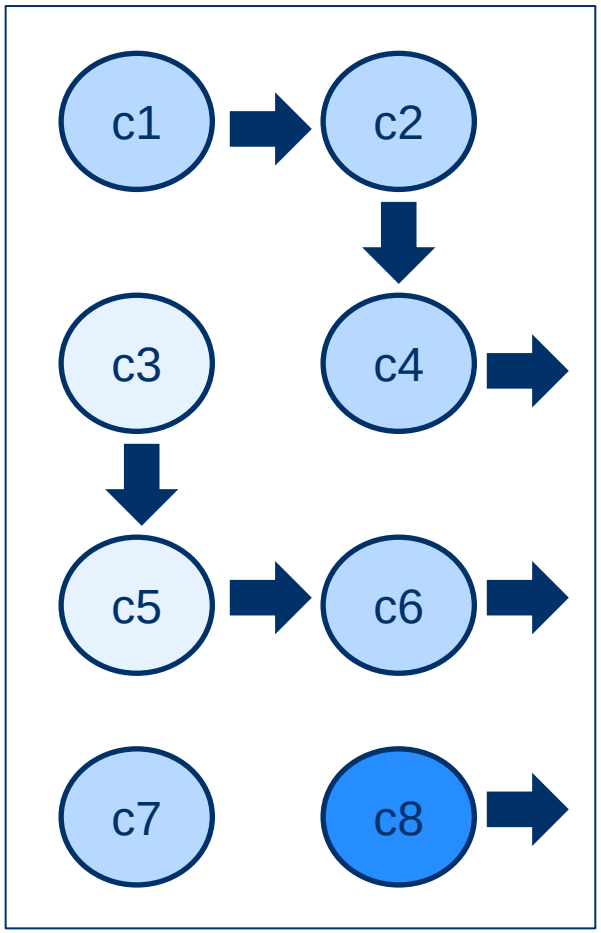


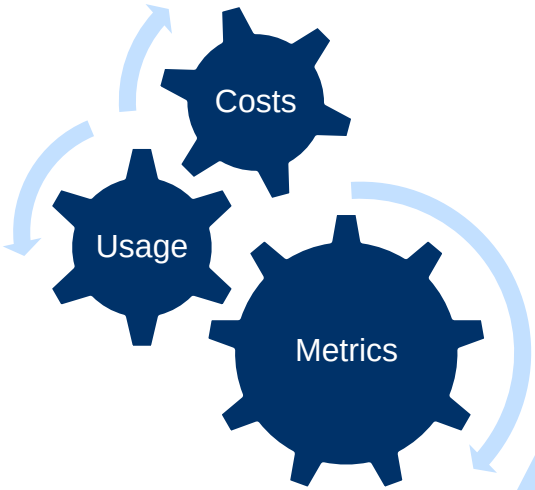
Hadoop as a DQM tool: Dedoop

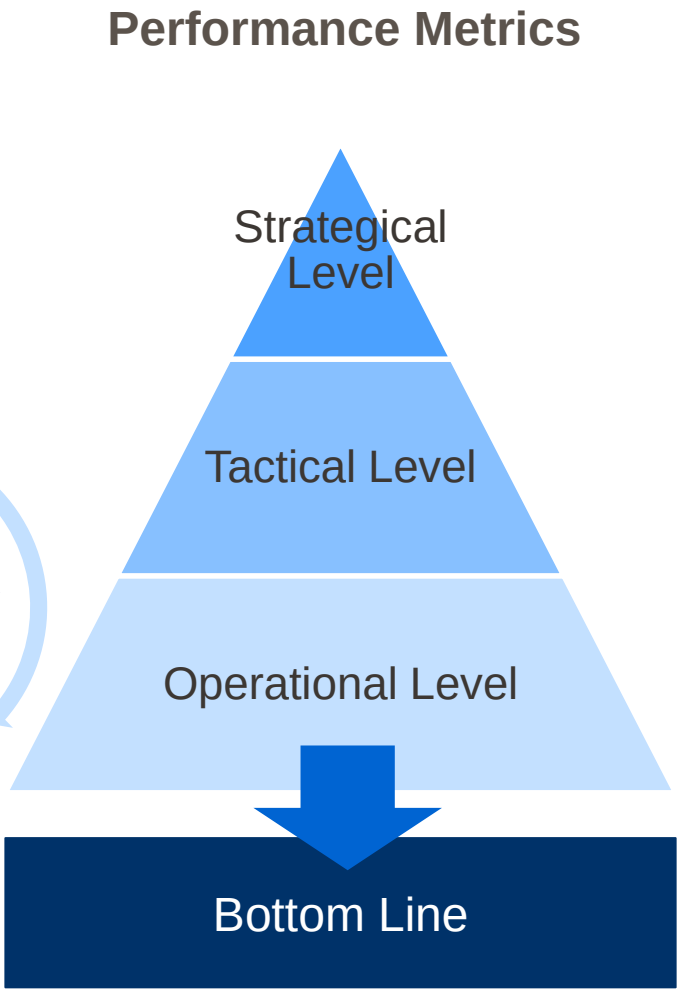


Entity Resolution workflow definition - Selection of the Blocking key generation function

Proactive DQM: My Causal Simulative Model



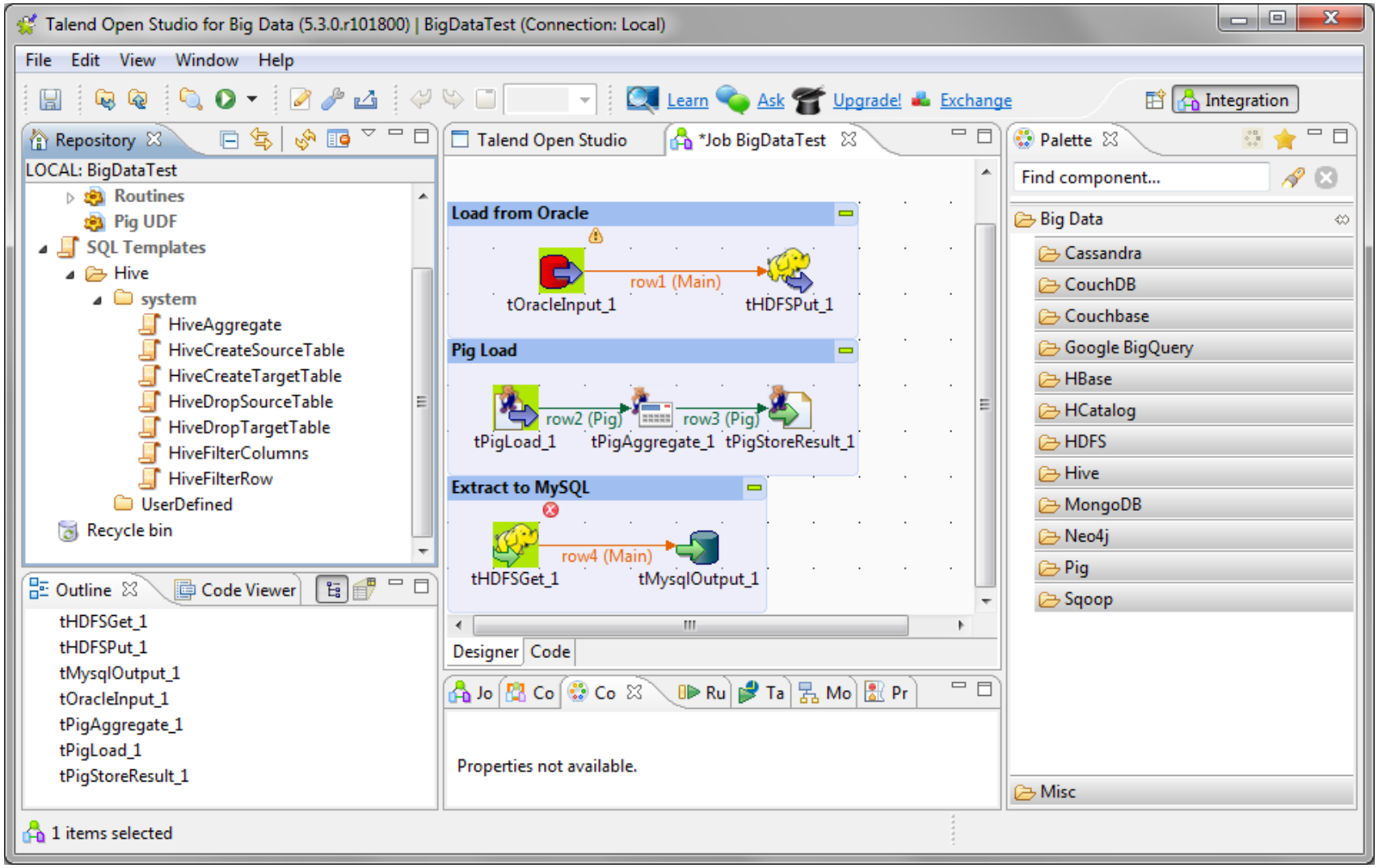




Support for Big Data Quality in DQM Tools

DataConference 2013

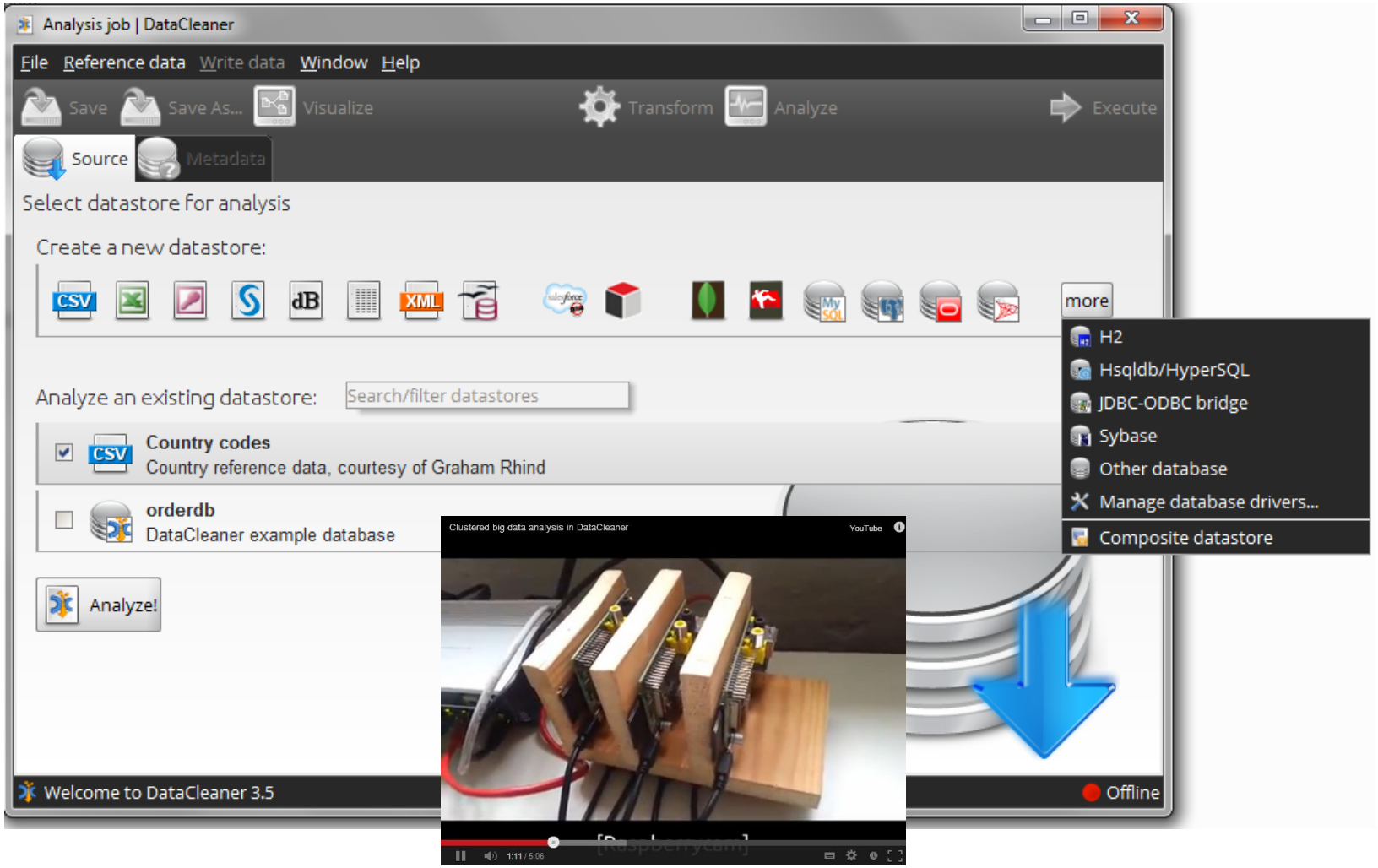
Talend Open Studio for Big Data



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 simmilar to ESB model based on industry standard

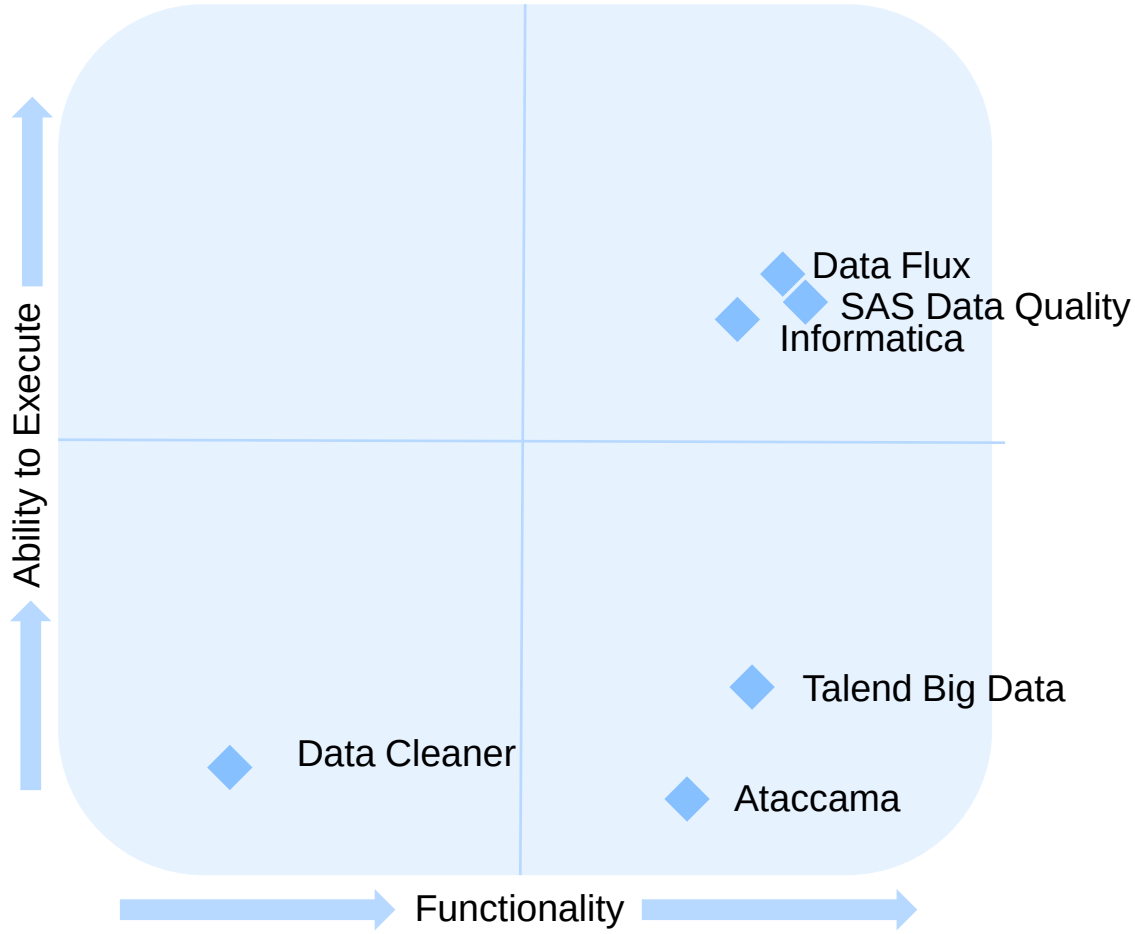
Data Cleaner



Ahother 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:

- Comparision 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 apliacated 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?