Big Data Quality / Governance

David Pejcoch

david.pejcoch@sopprasteria.com
david@pejcoch.com

Data a Znalosti
Oct 2015
... a long time ago

... and meantime somewhere on planet Earth a volume of data was growing exponentially
What is this presentation about?

- What is Data Quality Management / Governance
- What are Big Data
- What is Data Universe
- Different views on Big Data Quality / Governance
- Specifics of “Big Data Governance”
- Big Data Governance case study
- Introduction of some useful tools
What is Data Quality Management

- **J. M. Juran**: “*Data have high quality when it is possible to use them for intended use in operations, decision making and planning*”

- **Data Quality** = a level of some pre-defined (subjective / objective) characteristics data meet

- **Information Quality** = Data Quality + quality of context + presentation quality

- ... and so on ...

- Wrong data => wrong information => wrong knowledge ... (garbage in – garbage out)

- Data Quality in practice: missing data, incorrect data, duplicate data, inconsistent, delivered too late, with different formats, ...

Hierarchy of Knowledge (DIKW Hierarchy)
Zaveri has collected 109 different data characteristics, … still not all imaginable.

Impossible to manage all of them. You have to choose some reasonable subset.
Subjectivity of Information Quality

... a hell for European driver, ... something straightforward for Americans
Levels of Data Quality Management

Strategic Level (Data Governance)

- Definition of politics, key principles and rules

Tactical Level

- Definition of responsibility for data – Data Stewardship
- Definition of DQM components (standardization, validation, matching / merging, data imputation, data enrichment, geocoding, householding, ...)
- Quality Knowledge Base management
- Data Quality Audit (Assessment) as a part of IS Assessment (technical profiling, control testing, process analysis)
- Master Data Management (single version of truth about „master“ data)

Operational Level

- Implementation of controls
- Monitoring data quality
- Ad-hoc („local“) DQ
Data Governance

- Setting up the key strategy, goals and politics for (not just) corporate data
- Data considered as one of key assets governed under IT Governance
- **Principles** *(WHAT)* = key ideas, prerequisites, e.g. „master“ principle, effectiveness of information (availability), „due diligence“, best practices, data stewardship and responsibility / accountability for data
- **Politics** *(WHAT)* = „shared behaviour“ and rules
- **Standards** *(HOW)* = naming conventions, Common (Cannonical) Data Model, SOA, industry-specific standards
- **Drivers**: compliance (Basel, Solvency, HIPAA, …), risk management (COBIT, ISO/IEC 38500), new data sources, data quality requirements, certification, dynamics of business, data-centric systems, …
- **Data Stewards**: domain-based, technology-based, customer-based, process-based, system-based, project-based, hybrid model, … technical vs. business steward …. everybody is steward (L. English) vs. role defined within organisation vs. hybrid model (role + culture)
## Big Data

<table>
<thead>
<tr>
<th>Traditional definition</th>
<th>Gartner's 3V</th>
<th>3 + n*V</th>
<th>Data Universe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data which are impossible or not efficient to be processed using conventional approaches (RDBMs)</td>
<td><strong>Volume</strong> = huge amount of data</td>
<td><strong>Veracity</strong> = these data should be reliable</td>
<td><strong>Blending</strong> different data sources into single information product</td>
</tr>
<tr>
<td>Bad definition: some companies used these technologies for ages + relative definition</td>
<td><strong>Velocity</strong> = availability (but also changing very quickly) ... reference to stream data</td>
<td><strong>Value</strong> = reference to monetised data ... it doesn't make sense to store and manage data which cannot bring some current or potential value</td>
<td>... also medium and small data</td>
</tr>
<tr>
<td></td>
<td><strong>Variety</strong> = multiple structures and locations ... also multi-structured data</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Journey towards Big Data

![Graph showing the growth of different types of data over time (2000 to 2020), with labels for structured data, unstructured + semi-structured data, and IoT data.]

* IDC 2011 Digital Universe Study
** Internet of Things
Journey towards Big Data: Key Drivers
Data Universe

**Internal data**: relatively under control

**External data**: mostly out of control
Big Data Quality / Governance: Different Views

Quality of data within Big Data ecosystem

- Retrospective DQM: DQM at the level of respective tools
- Proactive DQM: DQ Management of Big Data sources – solving DQ issues as much close to original root-cause as it is possible

Big Data Ecosystem as a reference source

- For Data Enrichment – using derived information as a new attributes
- For Data Validation – using derived information as a reference source for validation

Big Data Ecosystem as DQM platform

- Platform for Master Data Management – using computing capacity of cluster for online / offline data deduplication and entity recognition – rainessance of distance-based similarity metrics
- Platform for retrospective data validation and metadata governance
Potential DQ issues of Data Universe

- **Consistency** across different sources
- **Compliance** - „Data in Hadoop can be less accurate“? - but this is not excuse for some regulations
- **Availability** – sometimes dependent on specific technical skills
- **Security and Privacy** – unified security rules across whole infrastructure to avoid data leaks
- **Quality of architecture** – requires technical skills from different areas
- **Interoperability** (availability of metadata) – big issue in Big Data world, for external data usually not available or inaccurate
- **Boundedness** (only useful data are stored) – some consultants say „store everything“ but then you need to manage these data and their quality (otherwise you just collect garbage)
- **Uniqueness** – a lot of duplicities across whole Data Universe (the same data in Hadoop and DWH)
Central Metadata Management Gap

What are metadata:

- Structured info, description, explanation, context, semantical layer, ...  
- **Technical metadata**: data types, lengths, formats, ...  
- **Business metadata**: labels, business context, ...

**Trends in metadata**: description of files with metadata instead of location in self-descriptive named folders (organisation by what data are about, not where they are) => definition of abstract layers on the top of different heterogeneous data sources

**Another trend**: metadata can be also result of structuralization

DCMI (Dublin Core Metadata Initiative) - Metadata Management Maturity Curve: the highest level = **Integrated Metadata Architecture** based on: 1) full metadata exchange, 2) enterprise wide reuse, 3) full automation ... and that's exactly what is missing
## How to get metadata from different data sources

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Method</th>
<th>Available Standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDBMS</td>
<td>Metadata synchronization + Extraction + Data Profiling</td>
<td>Industry specific canonical data models / reference data models, CWM (Common Warehouse Metamodel)</td>
</tr>
<tr>
<td>CSV, TXT</td>
<td>Metadata derivation from sample</td>
<td></td>
</tr>
<tr>
<td>Audio files</td>
<td>Metadata extraction, Annotation / Tagging</td>
<td>XMP, EXIF, IPTC</td>
</tr>
<tr>
<td>Video files</td>
<td>Metadata extraction, Annotation / Tagging</td>
<td>XMP, EXIF, IPTC</td>
</tr>
<tr>
<td>Images</td>
<td>Metadata extraction, Annotation / Tagging</td>
<td>XMP, EXIF, IPTC</td>
</tr>
<tr>
<td>PDF</td>
<td>Metadata extraction</td>
<td>XMP, EXIF, IPTC</td>
</tr>
<tr>
<td>Geospatial</td>
<td></td>
<td>CSDGM, ISO 19115</td>
</tr>
<tr>
<td>XML</td>
<td>Schema processing</td>
<td>WSDL, XSD, FOAF, ...</td>
</tr>
<tr>
<td>Web</td>
<td>Annotation</td>
<td>RDF, Dublin Core, ...</td>
</tr>
</tbody>
</table>
Disconnected Data Quality Management

- **What does it mean**: solving DQM separately at the level of different systems
- **Symptoms (1)**: Lack of central management of definitions
- **Symptoms (2)**: Respective rules are usually implemented as a spaghetti code (not re-usable)
- **Result**: Disciplines / buzzwords which don't deserve to survive = Metadata Quality, Linked Data Quality and Big Data Quality itself ... there is only single Data Quality topic with huge scope, single set of definitions, practices, rules which must reflect purpose of data
Data Maturity Model

- **The goal**: to enable decision whether some data source meets business expectations / requirements from DQ perspective and whether it can be used for intended purpose

- **Methodology**: evaluation from the perspective of key data characteristics. For each characteristic a set of criteria is defined

- **Evaluation**: weighted summary of answers to questions (for different questions / characteristics it is possible to set up different severity)

- **Decision**: how far is current level of characteristics from what is intended

- **Visualisation**: star plot
## Data Maturity Model - example

<table>
<thead>
<tr>
<th>Data Characteristic</th>
<th>Description</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interoperability</strong></td>
<td>Technical steward is defined for this data source</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Canonical data model or known ontology was used when creating this data source</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data model is known for this data source</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Description of columns / attributes is known and documented</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data lengths / types are known and documented</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Integrity constraints are known and documented</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sample data are available</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Description of a logic for batch integration is known and documented</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Information about used delimiter is known and documented</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Information about used formats is known and documented</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Information about charset is known and documented</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Description of a logic for online integration is known and documented</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WSDL or any alternative service description is available</td>
<td></td>
</tr>
</tbody>
</table>
Case Study: Why is important to have central governance

- Let's imagine **hypothetical internet business company with hundreds PB DWH** based on Teradata with several mil. records increment per day

- Let's imagine **huge table** with listing data described relatively well in Mediawiki-based metadata repository

- **Queries** on the top of this table are **generally skewed**. Especially for statistics related to # of active listings per site (join with site dimension) / week (join with calendar table) for some subset of products (join with product catalogue) and period longer than single quarter. That's the reason this company will decide to replicate some tables to Hive using Sqoop

- Original tables were available using views (security reasons and complex logic behind some columns not described in metadata). As a source for Sqoop used tables behind views without complex logic.

- The logic of Hive queues definition disconnected from DWH security policy.

- Without deep background knowledge it is not possible to identify from MR results whether everything is ok or not, thanks to not similar functionality of Teradata SQL and Hive QL in combination to Hive batch query it is not possible to know completeness of results.
Solutions: Big Data Capability within traditional DQM tools

- **SAS**: SAS/ACCESS Interface to „Whatever“ including Hadoop (MR, Pig, HDFS commands, Hive QL), possible to use SAS tools (EMiner, DI Studio, DataFlux,...), unified data management platform, PROC HADOOP, federated queries, central sophisticated QKB

- **Talend family** (TOS for Big Data / Integration / MDM): enormous amount of connectors to almost everything including Hadoop, extensible by Java/Groovy, however no central QKB, insufficient data profiling capability

- Ataccama
- Informatica PowerCenter Big Data Edition
- IBM InfoSpere Quality Stage
- Oracle Enterprise Data Quality
- Trillium Software BQuality
- SAP Business Objects Data Quality Management
Hortonworks Big Data Ecosystem (HDP)
2015: Data Governance Initiative (DGI)

- **Members:** Hortonworks, Aetna (Insur.), Merck (Pharm.), Target (Retail), SAS, SAP, Schlumberger (Oilfields serv. & equip.)

- **Guiding principles:**
  - Hadoop must snap into the existing frameworks and openly exchange metadata
  - Hadoop must address governance within its own stack of technologies

- **Key areas:** auditability (lineage), transparency of standards and protocols, reproducibility, consistency

- **Planned activities** (corresponds to Apache Atlas vision):
  - Flexible knowledge store – Titan Graph DB
  - Advanced policy rules engine – TBD
  - Agile auditing – YARN ATS
  - Support for Falcon work flows
  - Extension for Ranger to add real-time attribute-base access control to already implemented role-based access control
HCatalog

- Table management for Hadoop on the top of Hive Metastore, abstract layer (abstraction from physical format of table)
- By default: JSON, CSV, RCFILE, ... to use custom format you need to define InputFormat, OutputFormat and SerDe
- R/W interfaces for Hive, Pig and MR
- **Data model in Hive**: tables = HDFS directory, partitions = sub-directories, buckets = data split into files based on hash of column
- **Available metadata**: list of columns, types, owners, storage, SerDe information, location of underlying data + any user supplied key/value data
- **Metastore** = critically important information about data structure => needs to be backed up
- Access: Hue, command line interface (hcat.py, e.g. hcat.py -e „CREATE .. „ will execute CREATE statement, useful is also DESCRIBE command)
How HCatalog works

PIG Example:
Rdata = load 'somedata' using HcatLoader();
Store output into 'somedb' using HcatStorer("date = 20150923");
Apache Atlas

- In incubator from 07/2015: atlas.incubator.apache.org
- Central location of all HDP metadata
- Scope:
  - Data classification
    - Export metadata – exchange with other tools
    - Import annotations
    - Define / automate capture relations between data sets
  - Search & lineage
  - Security
- Integration via REST API
- Hive Bridge: importing Hive metadata
- Every HiveQL tracked => complete lineage
Apache Falcon

- In incubator from 2013
- Web UI
- Creates mirrors of data into cloud environment
- Tagging system to track lineage (based on Key-Value pairs)
- Defines, monitor (Data Lineage) and schedule operational and Data Governance policies for Hadoop workflows using XML
- On the top of Oozie job scheduling software
- Defines data pipelines: process + retention rules for data, retry policy for code, alternate workflows for late arrival data, ...
Apache Tika

Detects + extracts metadata and structured text content from files using parsers libraries

**Supported files:** MSO (Excel, Word, PowerPoint, Visio, Outlook), gzip, bzip2, tar, zip, jar, mp3, midi, wave, XML, HTML, Java classes, images, plaintext, OpenDocument, PDF, RTF + your own parsers

**Accessibility:** from command line, from your Java code, GUI

**Examples:**

```
java -jar tika-app-1.10.jar filename.pdf > output
```

... will create XHTML (default) file with extracted metadata and content

```
java -jar tika-app-1.10.jar -g
```

... will start GUI

For Java implementation is important:

```
AutoDetectParser parser = new AutoDetectParser();
```
Example of Tika Output for JPG

```
Apache Tika: justAnotherPictureOfSondraLocke.jpg

Color Transform: YCbCr
Component 1: Y component: Quantization table 0, Sampling factors 1 horiz/1 vert
Component 2: Cb component: Quantization table 1, Sampling factors 1 horiz/1 vert
Component 3: Cr component: Quantization table 1, Sampling factors 1 horiz/1 vert
Compression Type: Baseline
Content-Length: 111738
Content-Type: image/jpeg
DCT Encode Version: 1
Date Precision: 8 bits
File Modified Date: Wed Sep 30 02:02:25 BST 2015
File Name: apache-tika-1984684171200990309.tmp
File Size: 111738 bytes
Flags 0: 192
Flags 1: 0
Image Height: 566 pixels
Image Width: 991 pixels
Number of Components: 3
Resolution Units: none
X Resolution: 100 dots
X-Parsed-By: org.apache.tika.parser.DefaultParser
X-TIKA:digest:MD5: ba6254d9133e0464ac0453a60454c9a3
X-TIKA:digest:SHA256: c1d69e318b542bcbb3a187e490d734e921ad6b5482d5fc7b6e94c6fd06617258c
Y Resolution: 100 dots
resourceName: JustAnotherPictureOfSondraLocke.jpg
tiff:BitsPerSample: 8
tiff:ImageLength: 566
tiff:ImageWidth: 991
```
Concept of Global Metadata Management Framework
Example of Java library

Definition of meta-elements captured for different types of sources

Result table with URI, standardized meta-element and value

https://github.com/dpejcoch/Me
taExtractor
Where to find additional information