Data Governance to Manage Variety in Big Data

Ontology Summit, March 27, 2014

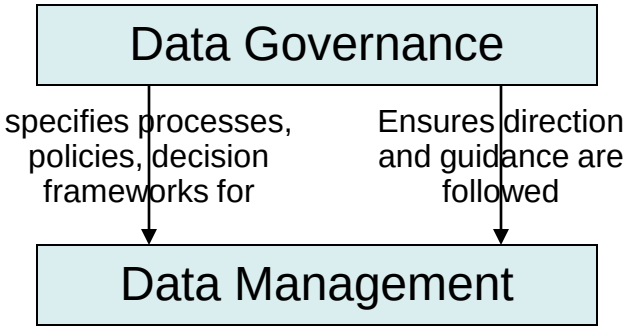
Presented by Malcolm Chisholm Ph.D.
Telephone +1-732-687-9283
MasterDataConsulting@gmail.com
www.referencedatadigest.com

Agenda

- What is Data Governance?
- Governing the Structure of Columnar Databases
- Governing Source On-boarding and Ingestion
- Information Requirements are Critical
- Response to Information Requirements
- Scope of Semantics for Governance

What is Data Governance?

a collection of disciplines that ensure data is managed adequately in an enterprise

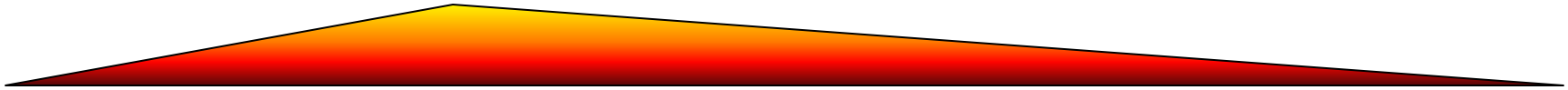


Examples of Data Governance Disciplines
Information Knowledge Management
Legal, Privacy, and Compliance
Principles, Policies, Practices
Issue Management

Metadata	the data that describes all aspects of an enterprise's information assets, and enables the enterprise to effectively use and manage these assets

Governing the Structure of Columnar Databases

rowID	Column Family	Column Qualifier	"Timestamp"	Payload



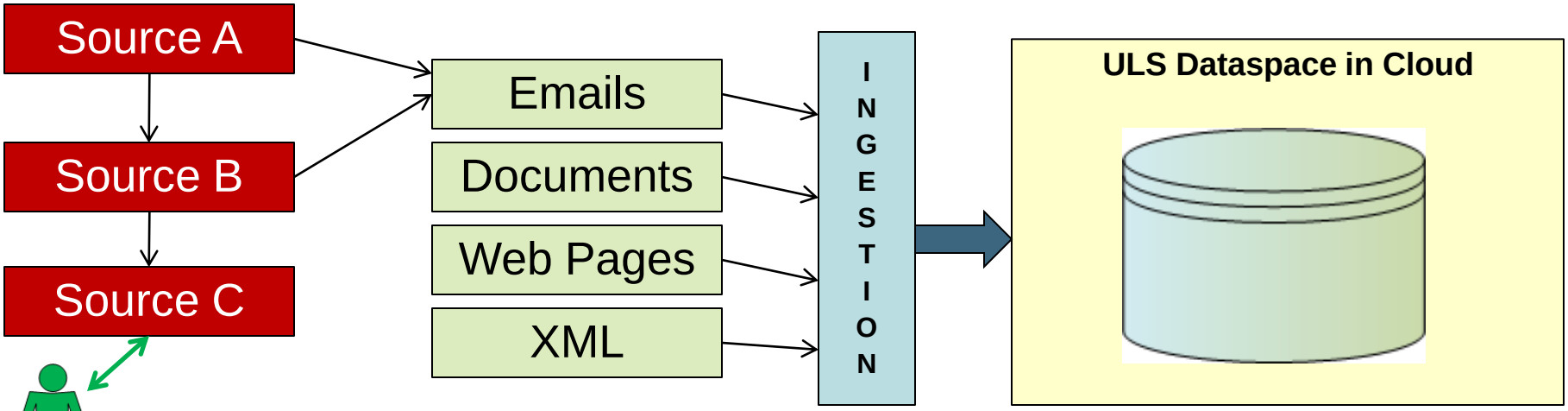
Doe|1968-11-04|John	"CUSTOMER"
Doe|1968-11-04|John	"EMPLOYEE"
Doe|1968-11-04|John	"PURCHASER"
	…and hundreds more…

What does "CUSTOMER" Mean?

How are all these values related? Part of same hierarchy? Independent? Something else?

- For these kinds of databases there is no system catalog and no foreign keys
- Scope for vastly diverse data within one table
- Can easily lose track of what Column Families have been defined per Table – and what they mean.
- Danger is that only the programmers will know (and they will lose track)

Governing Source On-Boarding & Ingestion

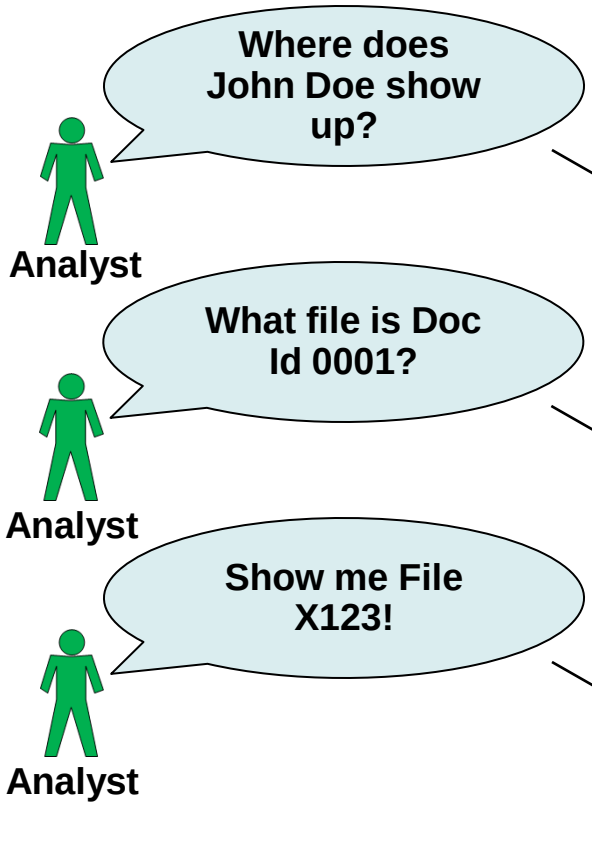


Data Management

What kind of data (ref. Subject Area Model)?
What data formats?
What data is unique vs. duplicated in other sources (profiling)?
Who are contact personnel?
What are SLA's, escalation procedures?
Data quality?

- On-boarding sources is a lot more difficult than hooking up ETL jobs
- Large numbers of sources available (but may copy data from each other)
- On-boarding means understanding sources from semantic and technical viewpoints
- Ingest is for whole data stores

Information Requirements are Critical

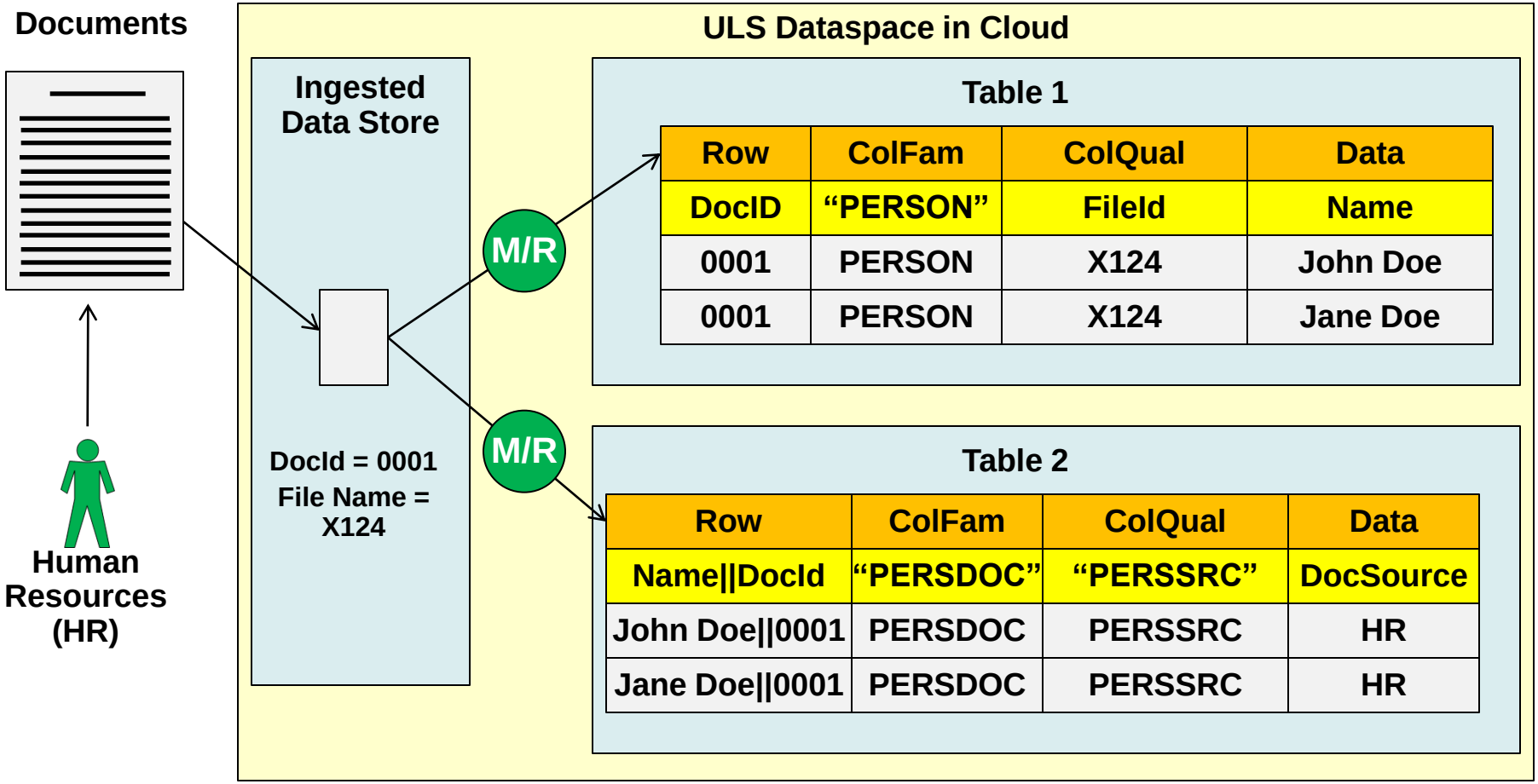


Row	ColFam	ColQual	Data
Name||DocId	"PERSDOC"	"PERSSRC"	DocSource
John Doe||0001	PERSDOC	PERSSRC	HR

Row	ColFam	ColQual	Data
DocID	"PERSON"	FileId	Name
0001	PERSON	X124	John Doe

- Data has to be shaped to answer the information requests
- As in data marts we start with information needs for design, but even more so with columnar databases – no "build it and they will come"
- This dictates tables, inverted indexes, segments
- So need to govern – capture, understand – information requirements

Response to Information Requirement, e.g. Inverted Indexing



- Constraints matter, e.g. rows are stored in lexicographic order
- Have to make the relationships via inverted indexes
- Increases diversity of copies of data in the dataspace
- Governance needed so we do not lose track of what we have and why it is there

Scope of Semantics for Governance

Terms / Concepts	**The concepts used in the business and the terms used to identify them – and their definitions**
Taxonomies	**The relationships of general with specific concepts**
Hierarchies	**How individual things (instances) are associated at multiple levels for specific business needs**
Relationships	**Other business relationships outside taxonomies and hierarchies**
Business Rules	**Atomic units of logic that govern behavior of concepts and relationships**
Ontologies	**Each is a "view" of the business world (and business information) that is required to meet a specific business need**

- Governance must address the above to manage variety in Big Data
- "Semantics" must really be "business semantics"
- Semantics = understanding business information as such – especially without any concern about how it might be stored as data
- There are no "Conceptual Data Models" – only "Conceptual Models".
- These models must be explicit and curated by Data Governance