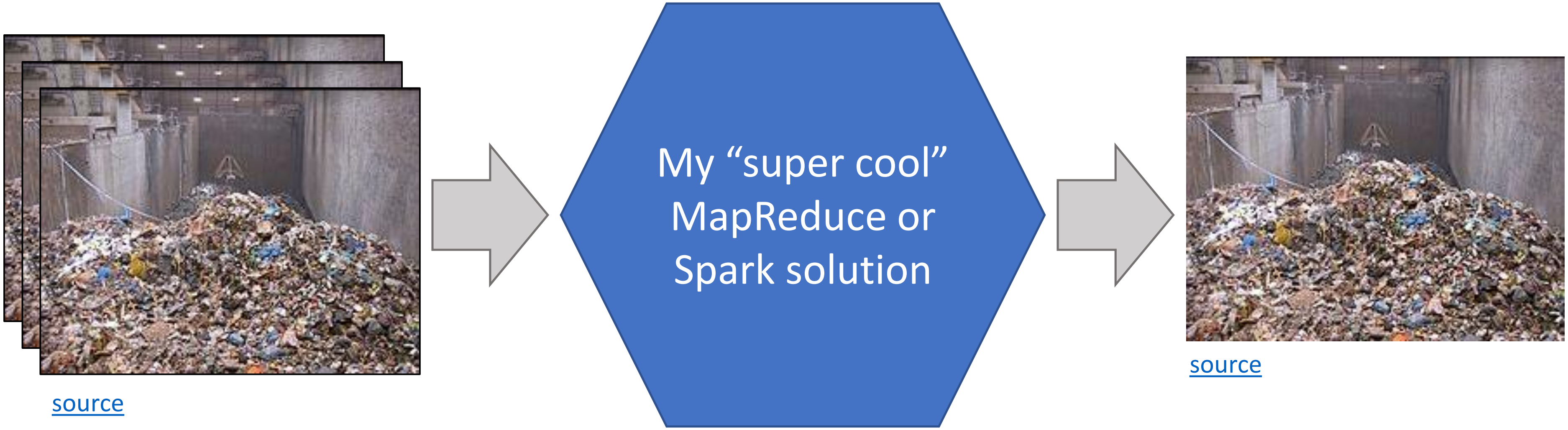
# Data Curation

COMP9313: Big Data Management

# Garbage in…



source

source

"around 20% of records in any data source are garbage." *M. Stonebraker*

**Data Curation is a must in any Big Data project!**

# What is data curation?

“Data curation is the process of identifying which data sources are needed, putting that data in the context of the business so that business users can interact with it, understand it, and use it to create their analysis.”

https://www.tableau.com/reports/business-intelligence-trends/data-curation

# Data Curation & Big Data

- We have too much data...
  - ➢ Problem -> **V**olume
- Data is coming too fast to our infrastructures...
  - ➢ Problem -> **V**elocity
- Data is coming from so many places and in different formats...
  - ➢ Problem -> **V**ariety

source:
Data Tamer: A Scalable Data Curation System. Michael Stonebraker, Database Group, MIT CS and AI Lab, 2013

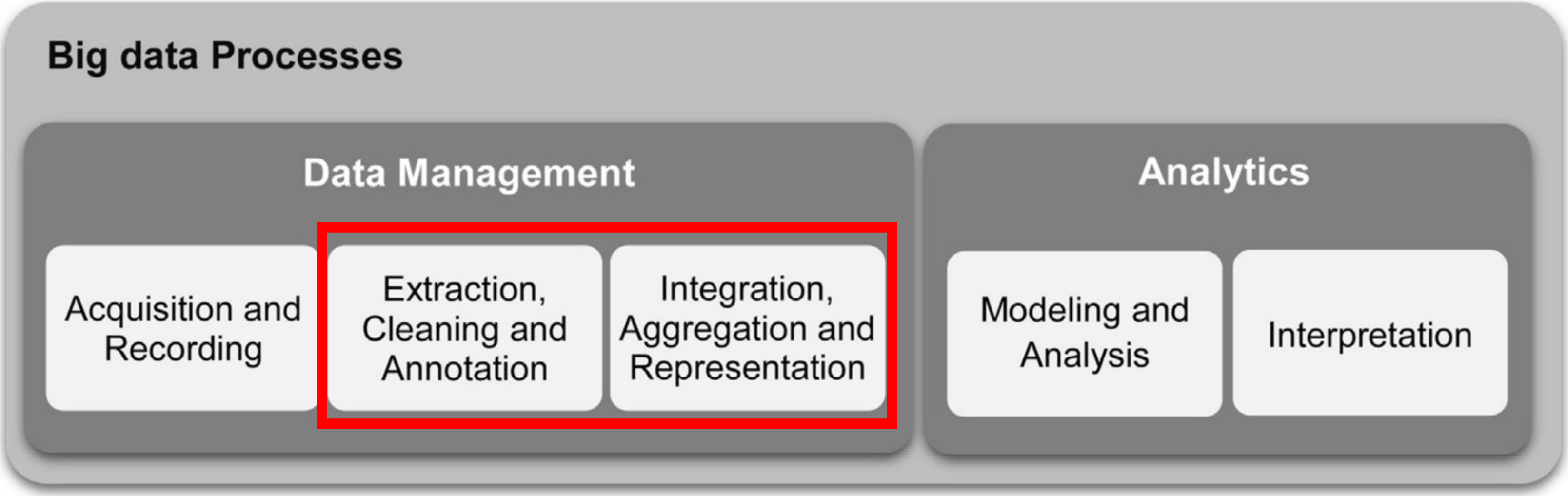
# Data Curation & Big Data

- We are uncertain as to whether this data reflects the true facts...
  - ➢ Problem -> **V**eracity

- We can’t access this piece of information...
  - ➢ Problem -> **V**isibility

- Storing/Processing this data is too costly, does it really contribute to obtaining insights?...
  - ➢ Problem -> **V**alue

# Big Data Processes


Big data Processes
Data Management
Analytics
Acquisition and Recording
Extraction, Cleaning and Annotation
Integration, Aggregation and Representation
Modeling and Analysis
Interpretation

https://www.sciencedirect.com/science/article/pii/S0268401214001066

# Data Quality Dimensions

| Dimension | Definition |
|---|---|
| Accessibility | "the extent to which data is available, or easily and quickly retrievable" |
| Appropriate amount of data | "the extent to which the volume of data is appropriate for the task at hand" |
| Believability | "the extend to which data is regarded true and credible" |
| Completeness | "the extent to which data is not missing and is of sufficient breadth and depth for the task at hand" |

https://dl.acm.org/citation.cfm?id=506010

# Data Quality Dimensions

| Dimension | Definition |
|---|---|
| Concise representation | “the extent to which data is compactly represented” |
| Consistent representation | “the extend to which data is presented in the same format” |
| Ease of manipulation | “the extent to which data is easy to manipulate and apply to different tasks” |
| Free-of-error | “the extent to which data is correct and reliable” |

https://dl.acm.org/citation.cfm?id=506010

# Data Quality Dimensions

| Dimension | Definition |
|---|---|
| Interpretability | “the extent to which data is in appropriate languages, symbols, and units, and the definitions are clear” |
| Objectivity | “the extent to which data is unbiased, unprejudiced, and impartial” |
| Relevancy | “the extent to which data is applicable and helpful for the task at hand” |
| Reputation | “the extent to which data is highly regarded in terms of its source or content” |

https://dl.acm.org/citation.cfm?id=506010

# Data Quality Dimensions

| Dimension | Definition |
|---|---|
| Security | "the extent to which access to data is restricted appropriately to maintain its security" |
| Timeliness | "the extent to which the data is sufficiently up-to-date for the task at hand" |
| Understandability | "the extent to which data is easily comprehended" |
| Value-added | "the extent to which data is beneficial and provides advantages from its use" |

https://dl.acm.org/citation.cfm?id=506010

# Data Curation

- Ingestion
- Validation
- Transformation
- Correction
- Consolidation
- Visualization

Data Tamer: A Scalable Data Curation System. Michael Stonebraker, Database Group, MIT CS and AI Lab, 2013

# Ingestion

- Obtaining / Importing data from potentially large number of sources

- Streaming / Storing data
  - Streaming -> E.g. online ML
  - Storing -> E.g. HDFS, Amazon S3, HBase, etc.

- Main "V" challenges
  - **V**olume, **V**elocity, **V**isibility

# Validation

- Is this data valid and does it represent true facts?

- Is this data valuable for the goals of my big data project?

- Main "V" challenges
  - **V**eracity
  - **V**alue

# Transformation

- Schema mapping
  - Global schema creation
  - Mapping of global-to-local schema
- Record linkage
  - Same logical entities, different data sources
  - Traditional Record Linkage -> static/structured records, same schema
  - Record Linkage in Big Data -> heterogenous sources, dynamic and continuously evolving
- Data fusion
  - Resolving conflicts
  - Finding truth about real-world -> veracity of data
- Main “V” challenges
  - **V**ariety, **V**isibility

# Correction

- Any good big data project needs to satisfy certain quality criteria (garbage in -> garbage out)

- Main quality dimensions
  - Free-of-error, believability, objectivity

- Main “V” Challenges
  - **V**eracity

# Consolidation

- Schema Integration
  - Partial or complete global schema? (or nothing at all)
- Consolidating data sources
  - Use of synonyms, templates and authoritative tables
  - Use of crowdsourcing?
  - Incremental improvement of consolidation techniques
- Consolidating entities

Data Tamer: A Scalable Data Curation System. Michael Stonebraker, Database Group, MIT CS and AI Lab, 2013

# Visualization

- Data/results visualization
  - Spatial data (e.g. bacteria spread over tissue)
  - Geospatial data (e.g. heat map forecast for weather temperatures)
  - Time-oriented data (e.g. interest rates forecast)
  - Multivariate Data (e.g. individual income vs. literacy)
  - Trees, Graphs, Networks (e.g. social network visualization)
  - Text / Document visualizations (e.g. sentiment analysis)

# Data Curation in Practice

# Generations of Data Curation Approaches

- 1st Generation (90s) -> Traditional Extraction-Transform-Load (ETL)
- 2nd Generation (2000s) -> ETL enhanced
- 3rd Generation (now) -> Scalable Data Curation

Scalable Data Curation
Michael Stonebraker, Strata + Hadoop Word, 2015, San Jose (CA)

# Generations of Data Curation Approaches

- 1st Generation (90s) -> Traditional Extraction-Transform-Load (ETL)
  - ➢ Started with sales data integration in the retail sector
  - ➢ Goal -> Make better stock decisions
  - ➢ Smarter buying decisions helped paid the warehouse within months
  - ➢ Methodology:
    - Human defines global schema upfront
    - Programmer works on each datasource
    - Scales to approx. 25 data sources

Scalable Data Curation
Michael Stonebraker, Strata + Hadoop Word, 2015, San Jose (CA)

# Generations of Data Curation Approaches

- 2nd Generation (2000s) -> ETL enhanced
  - Incorporated deduplication systems
  - Outlier detections for cleaning data
  - Standard domains for cleaning data
  - Enhanced generation 1, but still scales to only approx. 25 data sources.

Scalable Data Curation
Michael Stonebraker, Strata + Hadoop Word, 2015, San Jose (CA)

# Generations of Data Curation Approaches

- 3rd Generation (now) -> Scalable Data Curation
  - Focuses on scalability and automation
  - Scalability
    - 1,000s or 10,000s data sources
  - Automation
    - Use of ML and statistics for “low hanging fruits”
    - Parallelization is a must (big data)

Scalable Data Curation
Michael Stonebraker, Strata + Hadoop Word, 2015, San Jose (CA)

# Data Curation at Scale

- Data curation is an ongoing task
- Use of expert sourcing **is a must**
- Fitting into the organization’s ecosystem **is a must**
- A scheme for finding data sources **is a must**

Scalable Data Curation
Michael Stonebraker, Strata + Hadoop Word, 2015, San Jose (CA)

# Data curation is an Ongoing Task

- Organizations add new data on a regular basis
- Streams of data keeps arriving all the time
- Data properties and characteristics keep changing
- Mergers -> Integration, transformation, data fusion may totally change
- Recommendation: Global schema and curation algorithms must be incremental

Scalable Data Curation
Michael Stonebraker, Strata + Hadoop Word, 2015, San Jose (CA)

# Use of Expert Sourcing is a Must

- Cannot always rely on purely automated curation
- Domain expertise required (e.g. genomics, medicine, material engineering)
- Expertise is hierarchical
- Use expertise in a smart way (e.g. load balance, consult with the right expert, etc.)

Scalable Data Curation
Michael Stonebraker, Strata + Hadoop Word, 2015, San Jose (CA)

# Fitting into the Organization's Ecosystem is a Must

- Ingest from all kinds of data sources (**Variety**)
- Export to a variety of data sinks (**data sharing**)
- Keep original data sources in situ (**data governance**)
- Access control is a must (**privacy and security**)
- Support for data partners (**data sharing**)

Scalable Data Curation
Michael Stonebraker, Strata + Hadoop Word, 2015, San Jose (CA)

# A Scheme for Finding Data Sources is a Must

- CIOs typically have no idea of how many data sources they have

- CIOs typically have no idea of how many duplicates they have in their data sources

- Use of templates for common integration problems can be useful here:
  - ➢ Procurement optimization
  - ➢ Customer data integration
  - ➢ Etc.

Scalable Data Curation
Michael Stonebraker, Strata + Hadoop Word, 2015, San Jose (CA)

# Data Curation Tools

# Data Tamer

- Automated data integration
- Automatic schema mapping
- Entity de-duplication
- Leverages human experts and crowd for integration verification

https://link.springer.com/chapter/10.1007/978-3-319-21569-3_6

# ZenCrowd

- Named-entity to Knowledge Base (KB) linking

- Main goal:
  - Bridge gap between automated/manual linking

  - with automated linking with humans

https://link.springer.com/chapter/10.1007/978-3-319-21569-3_6

# CrowdDB

- Answers queries that cannot be answer with traditional DB systems or search engines
- Uses fuzzy operations with help of humans
- Ranks items based on relevancy

https://link.springer.com/chapter/10.1007/978-3-319-21569-3_6

# Talend

- Data integration & cleaning
- Provides Master Data Management (MDM) functionality
- Data governance (data catalog, data quality, data stewardship, etc.)

https://www.talend.com

# Pentaho Data Integration (Kettle)

- Data integration
- Extraction-Transform-Load (ETL)
- Uses dataflow programming
- Integrates with various storages and web services

https://community.hitachivantara.com/docs/DOC-1009855-data-integration-kettle

Thanks