REQUIREMENT ENGINEERING FOR DATABASES AND DATA CLEANING

PRESENTED TO: PROF. DAN BERRY
PRESENTED BY: MARIAN BOKTOR
COURSE CODE: CS846

UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS
David R. Cheriton School
of Computer Science

Outline

- Introduction
- Databases and Data Quality
- Data Cleaning
- Requirements for a good Data Cleaning system
- Summary

How to know if a database is any good?

- **Helps support and ensure the accuracy and integrity of information**
- Divides information into subject-based tables to reduce redundant data
- Provides Access with the information it requires to join the information in the tables together as needed
- Accommodates your data processing and reporting needs
- It is easy to modify and maintain without affecting other fields or tables in the database
- Information is easy to retrieve, and user applications are easy to develop and build

How to know if a database is any good?

- The database is scalable, meaning that it can be expanded to meet the changing needs of an organization
- Normalized, to minimize data redundancy, I/O redesign transaction sizes and to enforce referencing integrity
- Semantically same data have same representations in heterogeneous sources
- We have to have a single view and access to accurate and consistent data
- Quality Design: data is stored in a single logical unit instead of multiple files, maintaining data integrity and security, and finally increase the performance of the database

How to know if database software is any good?

- Easy to use
- Customizable
- Mobile-optimized
- Real-time insights
- Cloud-based
- Tailored for any company
- Multiple database formats
- Define constraints to maintain data integrity
- e.g. TeamDesk, Knack, Oracle Database Cloud Service, Microsoft Azure Cloud

Why we need a good database?

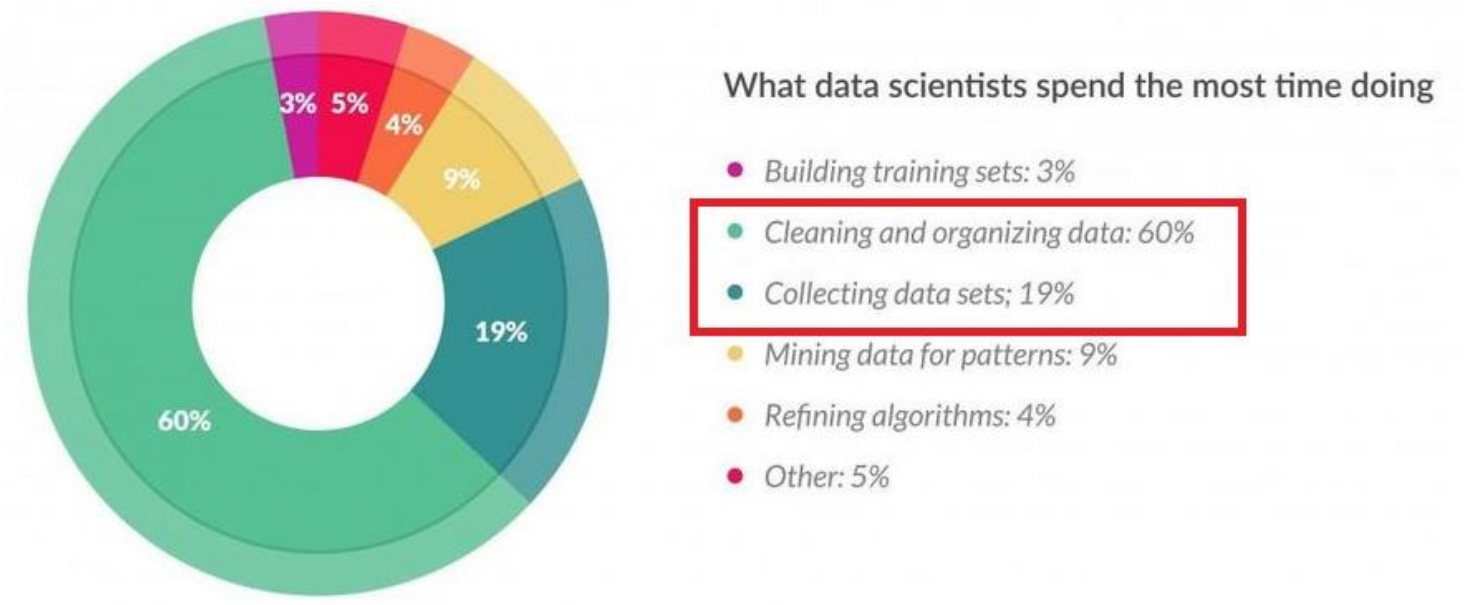
- Centralized systems
- Better management of human resource (HR) matters
- Managing customer data and relationships
- Efficient inventory tracking
- Planning for growth

High Quality Data | Criteria

- Validity
  e.g. Mandatory, Data-type, Range, Foreign-key and Unique constraints
- Accuracy
- Completeness
- Consistency
- Uniformity
- Traceability
- Timeliness

Data Cleaning

- According to an article on Forbes, by the year 2020, about **1.7 megabytes** of new information will be created every second for every human being on the planet. This means that different types and sources of data will be at play, and that could prove to be messy



Data Cleaning
CYCLE

Data Cleaning

- **What is Data Cleaning?**
  - It is basically the act of identifying and correcting false, incomplete or irrelevant parts of data
  - Process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data

Data Cleaning

Example

	City	Country	Population	Area	Density
r_1	New York	USA	8734520	6400	26403
r_2	Philadelphia	United States	“1,204,542”	“3,231”	NaN
r_3	New York City	USA	8734520	6400	26403

Data Cleaning

- **Why is it important?**
  - It will make everyone more efficient since they'll be able to quickly get what they need from the data
  - Reduce operational and maintenance cost as a result of enhancing data quality
  - Marketing (Increased productivity), Sales (faster cycle), Compliance and Operations (better decisions)
  - Your project will live or die by the quality of your data; garbage in, garbage out!!!

Data Cleaning

Why is it important?

It’s kind of like getting ready for a vacation. You might not like the preparation part, but tightening down the details in advance can save you from one nightmare of a trip.

You just have to do it..
or you can’t start having fun.

Data Cleaning

- **What are the challenges?**
  - Large-scale data, hard to detect and repair missing, erroneous and duplicate data
  - Expensive and time-consuming in terms of analyst effort
    Almost all data cleaning software requires some level of analyst supervision, on a spectrum from defining data quality rules to actually manually identifying and fixing errors.
  - Error correction and loss of information
  - Maintenance of cleansed data (require efficient data collection and management techniques)

Data Cleaning

- **What are the challenges? (cont'd)**
  - Data cleansing in virtually integrated environments
    e.g. IBM, requires cleansing of data every time the data is accessed; lower efficiency
  - Data-cleansing framework is iterative; requires integration and maintenance stages)
  - As a result of human-in-the-loop cleaning systems; novel problems at the intersection of human factors and database research are presented

Therefore; we need to know how data analysts use data cleaning tools, and what changes must be made to make data cleaning faster and more reliable.

Data Cleaning

Existing Systems Categories

- **Extract-Transform-Load (ETL) systems**
  - **Requirements:**
    - manually written data cleaning rules
    - constraint-driven tools to define data quality rules
  - **Drawbacks:**
    - do not provide the opportunity for analyst iteration or user feedback– inhibiting the user's ability to rapidly prototype different data cleaning solutions

Data Cleaning

Existing Systems Categories

- **Interactive frameworks like Wrangler and OpenRefine**
  - **Requirements:**
    - user direct manipulation of a data sample
  - **Drawbacks**
    - limited to specific cleaning tasks (extraction)

Things you should know about data cleaning task!

According to a survey done by Amplab, UC Berkeley and Columbia University

- **Data Cleaning Methodology**
  - **Data cleaning is iterative***

    It needs to be an **interactive** and **iterative** process.
    "*It's an iterative process, where I assess biggest problem, devise a fix, re-evaluate. It is dirty work.*"

    >> PROBLEM! Overfitting! <<

    Needs to be performed not in isolation but together with schema-related data transformations based on comprehensive metadata

*analysts alternate between cleaning and analysis, and use the analysis to guide future cleaning results

Things you should know about data cleaning task!

According to a survey done by Amplab, UC Berkeley and Columbia University

- **Data Cleaning Methodology (cont'd)**
  - **Evaluating data cleaning is ad-hoc**

    "Other than common sense we do not have a procedure to do this."

    "We usually do not do rigorous validation of data cleaning. We typically clean our data until the desired analytics works without error. This is not desirable but practical since in most cases data error is probably overshadowed by errors/inaccuracies in the models themselves."

    "We typically cross-reference data with other published materials to make sure it is in the right ballpark"

    >> Data Cleaning community needs to have a better answer to this problem <<

Things you should know about data cleaning task!

According to a survey done by Amplab, UC Berkeley and Columbia University

- **Analysts vs. Infrastructure**
  - **Infrastructure engineers and analysts address data quality problems differently**

e.g. Analysts GUESS there is an error while looking at the models (highly analysis-dependent, and depends on who knows semantics of data)...

While Infrastructure Engineers see dirty data as SYMPTOMATIC of an error in the processing pipeline (i.e., a software bug or incorrect schema)

Requirements

- **Interactive data cleaning and analysis process**
  - Infrastructure engineers, data analysts and domain experts can design, evaluate, and modify (with automated support) any stage of the data cleaning workflow

- **Developing High-level Language for Domain Experts**
  - Algorithms such as machine learning, clustering, or rule-based procedure too specialized for expert
  - Should describe the data cleaning goals at a logical level (e.g., providing de-duplication examples, descriptions of outliers), using a visual interface; guides implementation too

Requirements

- **Application-oriented Cleaning**
  - Systems tailored to specialized use cases (individual aggregation queries and convex models), and support for other more complex operations as well as multi-stage sequences of analyses is needed

- **Human-Computer Symbiosis**
  - Let experts do what they do best, while machines do the rest

  e.g. active learning is used to optimize an operation such as deduplication; automatic hyper-parameter tuning for machine learning models

Requirements

- **Testing and Debugging**
  - In order to develop automated optimization and tuning procedures, there must be a *metric* to optimize (not just compile and run)
  - For example, one might use performance on gold examples of known clean data to evaluate a cleaning operator
  - To accelerate data cleaning research, there is a need for a *cleaning benchmark* analogous to industry standard analytical data processing benchmarks
  - Provide tools to summarize *and* explain the intermediate results

Requirements

- **Combating Over-fitting**
  - Potential to favor certain outcomes, into the analysis process
  - Design data cleaning tools that:
    - Allow analysts to communicate assumptions (e.g., which records have been removed) when presenting results,
    - Automatically determine when an assumption is risky (e.g., correlates with the tested hypothesis),
    - Manages a “paper trail” of data transformations.

Commercially Available Tools

- **Data quality functionality**
  offered as a separate product, or integrated, or a suite of functions covering full spectrum of capabilities
  - data profiling,
  - data parsing/ correction,
  - data matching/ de-duplication,
  - enrichment,
  - integration,
  - data monitoring

Features	Athanor 3.0	IQ8	WebSphere QualityStage
Vendor	Informatica	First Logic (Business Objects)	IBM
Data Profiling	Basic	Good	Very good
Data parsing/correction	Basic	Very good	Very good
Matching/de-duplicate	Good	Very good	Very good
Enrichment	-	Very good	-
Integration	Good	Good	-
Real time support	-	Good	Good
Database support	-	Very good	Good
User Interface	-	Intuitive	Easy to use

Commercially Available Tools

- **Target**
  - Customer data
  - Product data
  - Financial data
- **Powerful Technologies**
  - Fast access to and from relational databases like: Oracle, SQL Server, DB2, ...etc
  - Support Windows ODBC connectivity and traditional delimited files
  - Provide a scope for connectivity to databases, integration with enterprise systems like Enterprise Resource Planning (ERP) and data warehousing tools like Extract transform Load (ETL)
  - Support Unicode framework, Geocoding and SOA
    also Web Services standards like SOAP, XML, WSDL, UDDI and HTTP

Commercially Available Tools

- **How to know is a data cleaning tool is any good?**
  - Handy for dealing with messy data, raw data ingestion; clean, match and transform data from one format to another at a fast pace
    e.g. OpenRefine, Google
  - Free or at affordable cost, interactive, with machine learning algorithms that helps in suggestion common transformations and aggregations
    e.g. Trifacta Wrangler

Commercially Available Tools

- **How to know is a data cleaning tool is any good?**
  - Advanced data cleansing and fuzzy matching is included, also multi language edition available (e.g. Winpure)
  - Strong data profiling engine for analysing the quality of data to drive better business decisions, also lets you build your own cleansing rules (e.g. Data Cleaner)
  - Helps in delivering quality data for big data, business intelligence, data warehousing, master data management, ...etc. (e.g. IBM Infosphere Quality Stage)

Commercially Available Tools

- **More needs to be done**
  - Support for control and monitoring data in real-time and streaming
  - Expected to cleanse data before it is saved onto a database or data repository, or used for triggering or actuating some action linked to the status of data
  - Tools should enhance their capabilities from syntactical (structure of the data) to semantic (meaning of data, metadata)
  - Tools based on metadata-driven design will enable enterprises to move beyond defect inspection to solving data defects through **root-cause** analysis

THANK YOU!