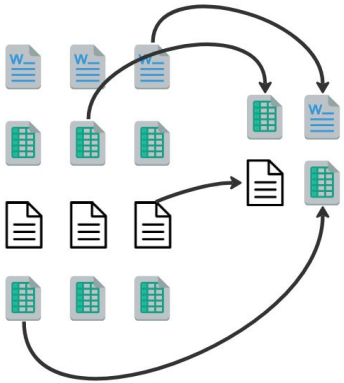
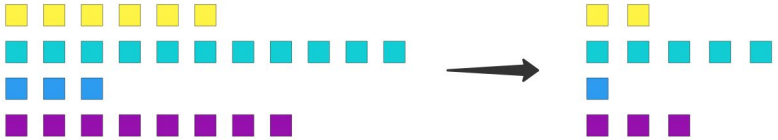
DQ Management in ETL Process under Resource Constraints

Kashosi Aser, Ternopil National Technical University

aser_kashosi2107@tntu.edu.ua

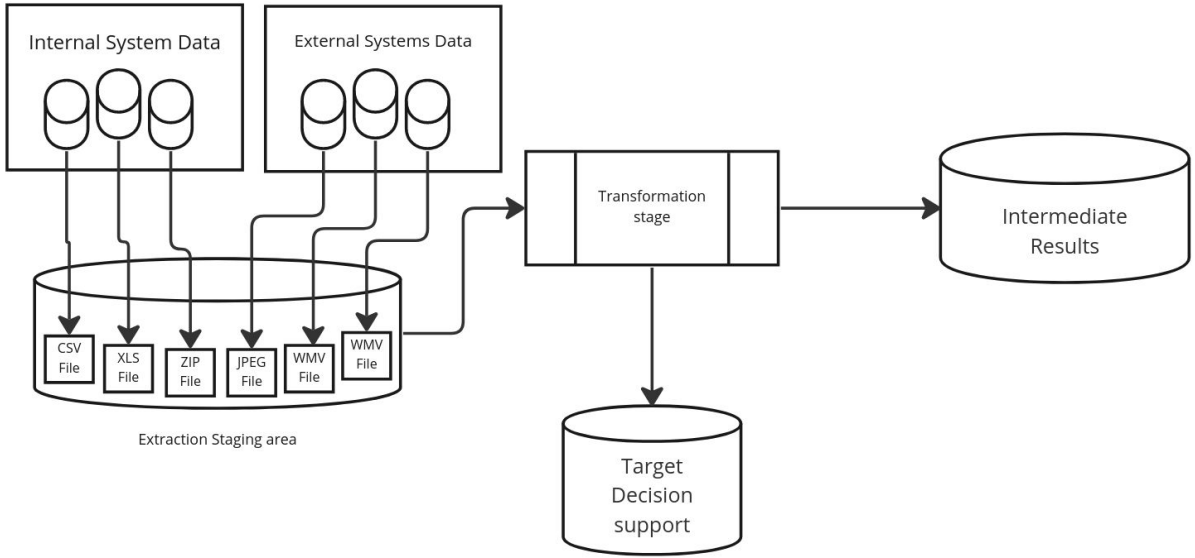




ETL (Extract, Transform, Load)

- Data collection is fraught with difficulty:
  - Heterogeneous sources: They will arrange information in entirely distinct schemas
  - Quality issues: ranging from simple spelling errors in textual attributes to inconsistencies in values, database constraint violations, and conflicting or missing information
  - Up-to-date information: information that populate the warehouse is continually being updated
- What ETL brings to the table:
  - Transform incoming source data into a common "global" data warehouse schema
  - Remove data “noise”
  - Routinely refresh the contents of the data warehouse

Data Quality Management in ETL Process under Resource Constraints



DQ (Data Quality)

- DQ dimensions considered in the context of ETL:
  - Completeness: Record count validation, integrity constraint checking
  - Accuracy: field-to-field comparison
  - Timeliness: Data are stored for the required period
  - Validity: Checking the data type of a field, and checking the field length.
  - Consistency: All values must be constant across all datasets
  - Uniqueness: stored data are free of duplicates.

BD (Big Data)

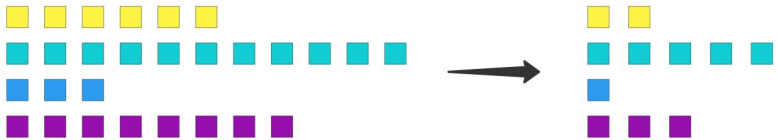
- "Big" data arises in many forms:
  - Activity data: GPS location, social network activity
  - Business data: customer behavior tracking at fine detail
- Common themes:
  - Data is large, and growing
  - There are important patterns and trends in the data

Why reduce BD for DQ assessment?

- Although “big” data is about more than just the volume…
  …most big data is big!
- It is not always possible to store the data in full
  - Many applications (telecoms, ISPs, search engines) can’t keep everything
- It is inconvenient to work with data in full
  - Just because we can, doesn’t mean we should
- It is faster to work with a compact summary
  - Better to explore data quality on a laptop than a cluster

Why Sample?

- Sampling has an intuitive semantics
  - We obtain a smaller data set with the same structure
- Estimating on a sample is often straightforward
  - Run the analysis on the sample that you would on the full data
  - Some rescaling/reweighting may be necessary
- Sampling is general and agnostic to the analysis to be done
  - Other summary methods only work for certain computations
  - Though sampling can be tuned to optimize some criteria
- Sampling is (usually) easy to understand
  - So prevalent that we have an intuition about sampling



Alternatives to Sampling

- Sampling is not the only game in town
  - Many other data reduction techniques by many names
- Dimensionality reduction methods
  - PCA, SVD, eigenvalue/eigenvector decompositions
  - Costly and slow to perform on big data
- “Sketching” techniques for streams of data
  - Hash based summaries via random projections
  - Complex to understand and limited in function
- Other transform/dictionary based summarization methods
  - Wavelets, Fourier Transform, DCT, Histograms
  - Not incrementally updatable, high overhead

Outline

- Motivating application: sampling in ETL data
- Stratified sampling: concepts and estimation
- Stratified sampling: Introduction of the weight evaluator parameter to apply the concept to text data DQ assessment

Sampling and Resource Constraints

Resource Constraints (Bandwidth, Storage, CPU)

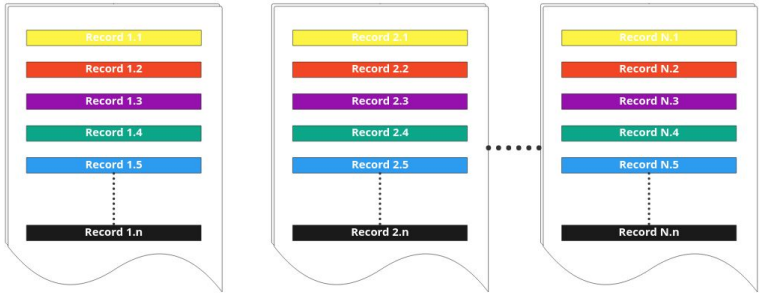
Sampling

Why Summarize (ETL) Big Data for DQ?

- Typically raw accumulation is not feasible
  - High volume batch data
  - Maintain historical summaries for time series analysis
- To facilitate fast queries
  - When infeasible to run data quality queries over full data

We prove that Sampling is a flexible method to accomplish this

Massive Dataset: File Records



- DQ tasks
  - Integrity constraint checking
  - Checking the data type of a field
  - Consistency
  - etc

Records and Sampling

- We use stratified sampling
  - Gets better data representation

Abstraction

- The population U consisting of N units :
  - Example : total number of combined records of all files stored in the data warehouse before DQ assessment
  - The population mean (yij is a DQ value assessment of the record y for the jth unit of the ith stratum)

$$\bar{Y} = \sum_{i=1}^{K} \sum_{j=1}^{N_i} y_{ij}/N = \sum_{i=1}^{K} W_{i\bar{Y}_i} \text{ With } W_i = N_i/N$$

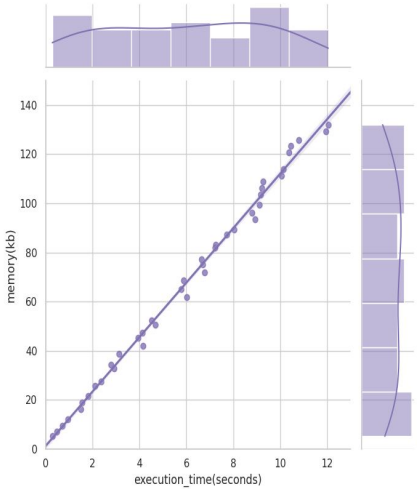
  - Estimation of the Population Mean (A sample si of size ni drawn from the stratum Ui with probability p(si))

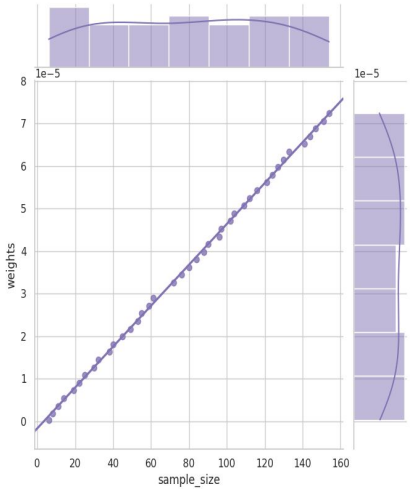
$$\widehat{\bar{Y}}_i = \sum_{j \in s_i} b_j(s_i) y_{ij}$$

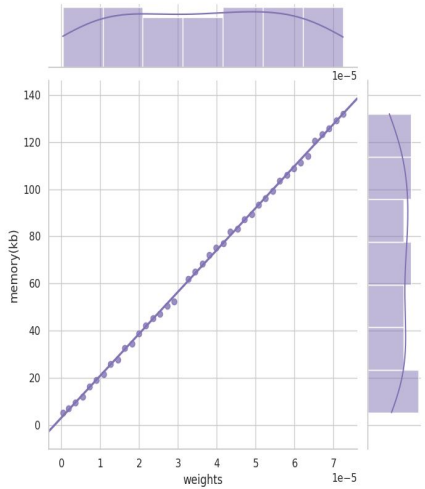
- Number of strata K, and the ith stratum Ui consists of Ni units:
  - Example: total number of files and each file contains Ni records

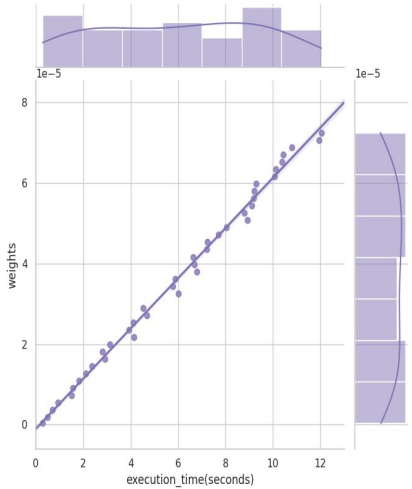
Test results

- Weight evaluator parameter:
  - Value: We defined the weighting parameter F for a number of files N with total size S(N) such that the weight
    w = F /S(N)
    This is to ensure a better distribution of data across all samples.



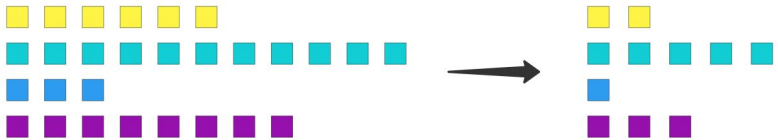






Observations

- Finding the most diverse records in the final sample is directly proportional to the chosen weighting.
- Choice of weight trades volume against variability
- In a resource-constrained environment
  - It is necessary to choose a minimum weight
  - It is of utmost importance to test and determine the applicable thresholds for weight, memory, and execution time on a small scale before determining applicability in a production environment.



Summary

- Extract, Transform and Load general challenges
- Data Quality management in ETL
- Big data and data quality challenges
- Stratified Sampling as a solution in the case study
  - The cost for applied stratified sampling
  - Practical weighting parameter for stratified sampling in the context of text data.

Data Quality Management in ETL Process under Resource Constraints