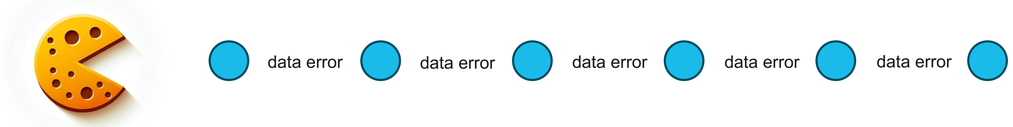
CheDDaR: Checking Data – Data Quality Report

Indra Diestelkämper[1], Ralf Diestelkämper[2], Valerie Restat[1]
[1]FernUniversität in Hagen, [2]Flinkback GmbH

Motivation

- With growing significance of data across various domains data quality becomes more important
- Data quality is imperial for the safety, reliability, and effectiveness of various high-stakes applications
- Many datasets suffer poor quality due to data errors

CheDDaR identifies data errors leveraging multiple verification methods to minimize manual efforts



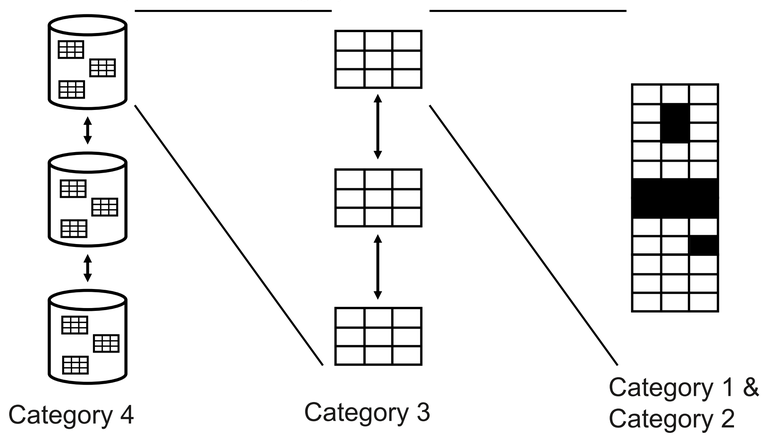
Foundations and Scope

Verification Methods

CheDDaR supports five verification methods

Error Categories

CheDDaR currently supports category 1 and 2 data errors



Data Quality Metrics

CheDDaR supports ten data quality metrics to identify data quality errors

Category Type	Error Type	Error Description
Category 1: An Attribute Value of a Single Tuple	Missing Values	This error indicates a missing attribute value in a tuple
	Syntax Violation	A data entry's format differs from the required one
	Interval Violation	A numeric value exceeds the predefined range
	Set Violation	An attribute value is not within the set of allowed values
	Wrong Data Type	The attribute's expected data type is not met
Category 2: The Values of a Single Attribute	Uniqueness Value Violation	Two or more attributes in a tuple redundantly represent the same entity
	Outlier	An outlier is a value that deviates significantly from others in the attribute
	Missing Attribute	This error indicates a missing attribute in all table tuples
	Additional Attribute	A dataset includes an unexpected or undefined attribute
	Zero Variance	An attribute holds the same unchanging value

Each metric is formally defined in relational algebra, e.g., the Missing Value (MV) metric:

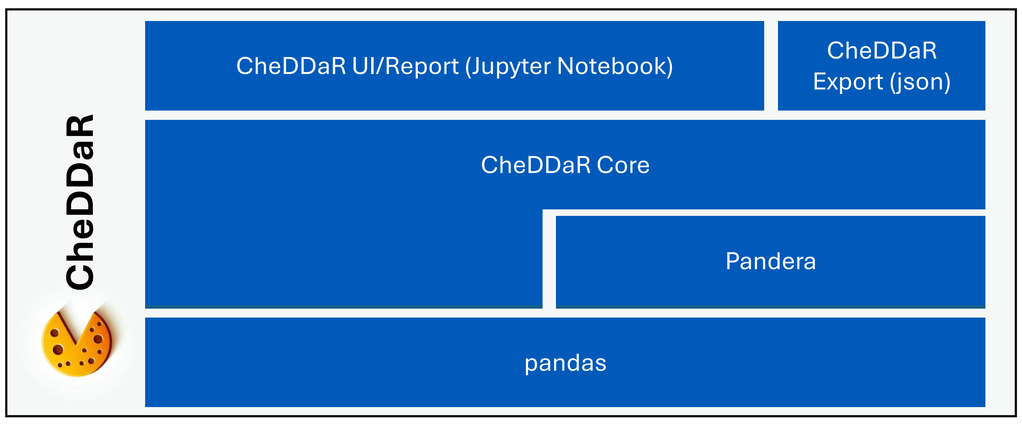
$$MV(R, A_i) := \frac{|\sigma_{A_i \ IS \ NULL}(R)|}{|R|}$$

MV is the quotient of the number of missing (null) values in an attribute A_i and the total number of tuples in relation R

Architecture

Implementation

- Combines and extends robust libraries like Pandas and Pandera
- Uses an interactive, user-friendly reporting interface (Jupyter Notebooks)
- Ensures effective analysis and presentation of data quality metrics
- Provides multiple interfaces for increased interoperability (i.e., Notebooks and JSON)



Process

- Starts with an Analysis Dataset and ends with a Data Quality Report
- Users select verification methods based on available reference data or predefined rules
- Default path applies automatic checks when no specific rules or reference data are available
- Aggregates results from all verification methods into a single report

Application Example

Fiktiva Dataset Results

Category	Metric	Gold Standard	Manual	Rule-based	Automatic
Category 1	Missing Values	x	x	x	x
	Syntax Violation	–	x	x	–
	Interval Violation	x	x	x	–
	Set Violation	x	x	x	–
	Wrong Data Type	x	x	x	–
Category 2	Uniqueness Value Violation	x	x	x	–
	Outlier	–	x	x	x
	Missing Attribute	x	x	x	–
	Additional Attribute	x	x	x	–
	Zero Variance	–	x	x	x

Data Quality Report

General Info | Ground Truth Checks | Gold Standard Check | Domain Expert Check | Rules Checks | Auto Checks

Attribute	Metric	Category	Amount	Percent	Technique	Failure Cases
Dept	Set Violation	Category 1	4	0.8	isin(['HR', 'Technology', 'Sales', 'Purchasing', 'Marketing'])	['Engineering', 'IT']
age	Interval Violation	Category 1	2	0.4	greater_than_or_equal_to(18.0)	[5, 15]
	Interval Violation 2	Category 1	7	1.4	less_than_or_equal_to(67.0)	[69, 73, 78, 87, 89, 'etc.']
awards	Interval Violation	Category 1	1	0.2	less_than_or_equal_to(25.0)	[57]
certifications	Interval Violation	Category 1	1	0.2	less_than_or_equal_to(1.0)	[9]
education	Set Violation	Category 1	2	0.4	isin(['PG', 'UG'])	['PhD']
emp_id	Uniqueness Value Violation	Category 2	4	0.8		['MKT7287', 'TECH7949']
entry_date	Interval Violation	Category 1	1	0.2	greater_than_or_equal_to(2004-01-05 00:00:00)	[Timestamp('1900-01-01 00:00:00')]
gender	Missing Values	Category 1	3	0.6		
	Set Violation	Category 1	394	78.8	isin(['Male', 'Female', 'Diverse'])	['d', 'f', 'm']
job_level	Interval Violation	Category 1	1	0.2	less_than_or_equal_to(5.0)	[99]
last_raise	Interval Violation	Category 1	2	0.4	greater_than_or_equal_to(0.01)	[-0.03, -0.01]
rating	Missing Values	Category 1	29	5.8		
	Wrong Data Type	Category 1	500	100.0	dtype('Int64')	['Float64']
recruitment_type	Set Violation	Category 1	3	0.6	isin(['Referral', 'Recruitment Agency', 'On-Campus', 'Walk-in'])	['Home', 'n.a']
salary	Interval Violation	Category 1	2	0.4	greater_than_or_equal_to(24076.0)	[-86000, -25000]
satisfied	Interval Violation	Category 1	1	0.2	less_than_or_equal_to(1.0)	[9]

Summary

- Manual validation performs best, but requires lots of hands-on analysis
- Multiple verification methods allow for a high quality data error report with little effort
- Additional evaluation on two real world datasets strengthen these findings

FernUniversität in Hagen