Data Validation at Scale

Managing Data Quality in Complex Data Pipelines

35th Web Perf Meetup, Hamburg, Germany

September 07, 2023

Wolle

Material Available at https://wolle.science









• ...

- Stream Processing
- Real-Time Databases
- NoSQL & Cloud Systems



Carl von Ossietzky Universität Oldenburg



Practice:

- Web Caching •
- Big Data Analytics •
- Anger Management •

• • •







The Importance of Data Validation

Where is data validation integrated into data science pipelines and what is its impact?



Data Quality & Constraints

What dimensions of data quality are there and how can they be ensured?



Scalability-Related Challenges

Why is data validation difficult in data-intensive domains?









Wolfram Wingerath (Hamburg, September 07, 2023)

Slides: wolle.science/data-validation



- **Goal**: verify that data in the pipeline is in an acceptable state for downstream processing, e.g.
 - External reporting (statistics, visualizations & dashboarding)
 - Internal reporting (debugging, product optimizations)
 - Decision-making (analytics, machine learning)
- Data validation can be integrated in and between all stages

Dimensions of Data Quality: Completeness



- There are different **dimensions of data quality**, especially:
 - Completeness: Do we have all the data we need to assess page load performance?
 - o Consistency: Does data have a valid format and does it comply with business semantics?
 - Accuracy: Do data items represent their corresponding real-world entities well?
 - Uniqueness: Are duplicate records known and are all unique attributes actually distinct?

Dimensions of Data Quality: Completeness

Timestamp	Pageload ID	Browser	LCP (Performance)	Session ID	URL
09:05:04.578	37ab08	Edge	dge "670ms" 123		null
13:26:48.139	9cddf7	Firefox	692.654	456	abc.de/red
13:28:23.857	0b577a	Firefox	0.256	456	abc.de/blue
13:29:17.468	faf55e	Edge	1.598	456	abc.de/sold
20:45:38.941	faf55e	null	null	null	abc.de/sold

• There are different **dimensions of data quality**, especially:

• Completeness: Do we have all the data we need to assess page load performance?

- o Consistency: Does data have a valid format and does it comply with business semantics?
- Accuracy: Do data items represent their corresponding real-world entities well?
- o Uniqueness: Are duplicate records known and are all unique attributes actually distinct?

Unload Beacon Reliability as an Example Challenge



- Is beacon loss a problem at all?
- When is beacon loss a problem?

 \rightarrow For which beacon types? For which beacon strategies?

Where is beacon loss a problem?

 \rightarrow For which browsers? For which device types?

Wolfram Wingerath (Hamburg, September 07, 2023)

Unload Beacon Reliability by Strategy



 Erik Witt. Unload Beacon Reliability: Benchmarking

 Strategies for Minimal Data Loss, Speed Kit Tech Blog (2023)

Unload Beacon Reliability by Strategy



Unload Beacon Reliability by Strategy & Device



Erik Witt. <u>Unload Beacon Reliability: Benchmarking</u> Strategies for Minimal Data Loss, Speed Kit Tech Blog (2023)

Unload Beacon Reliability by Strategy & Browser



Based on 52 million page views on a globally operating e-commerce site measured by Speed Kits real user-monitoring | August 2023

Erik Witt. <u>Unload Beacon Reliability: Benchmarking</u> <u>Strategies for Minimal Data Loss</u>, Speed Kit Tech Blog (2023)

Unload Beacon Reliability : The Ideal Combo Strategy

Can be used with Backward-Forward Cache?



Based on 52 million page views on a globally operating e-commerce site measured by Speed Kits real user-monitoring | August 2023

Erik Witt. <u>Unload Beacon Reliability: Benchmarking</u> Strategies for Minimal Data Loss, Speed Kit Tech Blog (2023)

Timestamp	Pageload ID	Browser	LCP (Performance)	Session ID	URL
09:05:04.578	37ab08	Edge	"670ms"	123	null
13:26:48.139	9cddf7	Firefox	692.654	Same	abc.de/red
13:28:23.857	0b577a	Firefox	0.256	456 456	abc.de/blue
13:29:17.468	faf55e	Edge	1.598	456 ⁹	abc.de/sold
20:45:38.941	faf55e	null	null	null	abc.de/sold

<u>Note</u>: Browser values should be unified for all records in the same session!

- There are different **dimensions of data quality**, especially:
 - Completeness: Do we have all the data we need to assess page load performance?
 - Consistency: Does data have a valid format and does it comply with business semantics?
 - Accuracy: Do data items represent their corresponding real-world entities well?
 - Uniqueness: Are duplicate records known and are all unique attributes actually distinct?

Timestamp	Pageload ID	Browser	LCP (Performance)	Session ID	URL
09:05:04.578	37ab08	Edge	"670ms"	123	null
13:26:48.139	9cddf7	Firefox	692.654	456	abc.de/red
13:28:23.857	0b577a	Firefox	0.256	456	abc.de/blue
13:29:17.468	faf55e	Firefox	1.598	456	abc.de/sold
20:45:38.941	faf55e	null	null	null	abc.de/sold

Note: Browser values should be unified for all records in the same session!
→ value may be replaced with majority vote (another reasonable option: replace old values with latest one)

- There are different **dimensions of data quality**, especially:
 - o Completeness: Do we have all the data we need to assess page load performance?
 - Consistency: Does data have a valid format and does it comply with business semantics?
 - Accuracy: Do data items represent their corresponding real-world entities well?
 - Uniqueness: Are duplicate records known and are all unique attributes actually distinct?

Timestamp	Pageload ID	Browser	LCP (Performance)	Session ID	URL
09:05:04.578	37ab08	Edge	"670ms"	123	null
13:26:48.139	9cddf7	Firefox	692.654	456	abc.de/red
13:28:23.857	0b577a	Firefox	0.256	456	abc.de/blue
13:29:17.468	faf55e	Firefox	1.598	456	abc.de/sold
20:45:38.941	faf55e	null	null	null	abc.de/sold

<u>Note</u>: All values represent milliseconds, but formats differ depending on browser.

- There are different **dimensions of data quality**, especially:
 - Completeness: Do we have all the data we need to assess page load performance?
 - $\circ~$ Consistency: Does data have a valid format and does it comply with business semantics?
 - Accuracy: Do data items represent their corresponding real-world entities well?
 - Uniqueness: Are duplicate records known and are all unique attributes actually distinct?

Timestamp	Pageload ID	Browser	LCP (Performance)	Session ID	URL
09:05:04.578	37ab08	Edge	670	123	null
13:26:48.139	9cddf7	Firefox	692654	456	abc.de/red
13:28:23.857	0b577a	Firefox	256	456	abc.de/blue
13:29:17.468	faf55e	Firefox	1598	456	abc.de/sold
20:45:38.941	faf55e	null	null	null	abc.de/sold

<u>Note</u>: All values represent milliseconds, but formats differ depending on browser. → values may be converted to integer

- There are different **dimensions of data quality**, especially:
 - Completeness: Do we have all the data we need to assess page load performance?
 - Consistency: Does data have a valid format and does it comply with business semantics?
 - Accuracy: Do data items represent their corresponding real-world entities well?
 - Uniqueness: Are duplicate records known and are all unique attributes actually distinct?

Dimensions of Data Quality: Accuracy

Timestamp	Pageload ID	Browser	LCP (Performance)	Session ID	URL
09:05:04.578	37ab08	Edge	670 123		null
13:26:48.139	9cddf7	Firefox	692654	456	abc.de/red
13:28:23.857	0b577a	Firefox	256	456	abc.de/blue
13:29:17.468	faf55e	Firefox	1598	456	abc.de/sold
20:45:38.941	faf55e	null	null	null	abc.de/sold

<u>Note</u>: Despite being in the right format, one value does not represent a reasonable timer value.

- There are different **dimensions of data quality**, especially:
 - Completeness: Do we have all the data we need to assess page load performance?
 - Consistency: Does data have a valid format and does it comply with business semantics?
 - Accuracy: Do data items represent their corresponding real-world entities well?
 - Uniqueness: Are duplicate records known and are all unique attributes actually distinct?

Wolfram Wingerath (Hamburg, September 07, 2023)

Dimensions of Data Quality: Accuracy

Timestamp	Pageload ID	Browser	LCP (Performance) Session ID		URL
09:05:04.578	37ab08	Edge	670	70 123 r	
13:26:48.139	9cddf7	Firefox	null	456	abc.de/red
13:28:23.857	0b577a	Firefox	256	456	abc.de/blue
13:29:17.468	faf55e	Firefox	1598	456	abc.de/sold
20:45:38.941	faf55e	null	null	null	abc.de/sold

<u>Note</u>: Despite being in the right format, one value does not represent a reasonable timer value.

 \rightarrow broken value may be removed

- There are different **dimensions of data quality**, especially:
 - Completeness: Do we have all the data we need to assess page load performance?
 - Consistency: Does data have a valid format and does it comply with business semantics?
 - Accuracy: Do data items represent their corresponding real-world entities well?
 - Uniqueness: Are duplicate records known and are all unique attributes actually distinct?

Dimensions of Data Quality: Uniqueness

Timestamp	Pageload ID	Browser	LCP (Performance) Session ID		URL	
09:05:04.578	37ab08	Edge	670 123		null	
13:26:48.139	9cddf7	Firefox	null	456	abc.de/red	
13:28:23.857	0b577a	Firefox	256	456	abc.de/blue	
13:29:17.468	faf55e	Firefox	1598	456	abc.de/sold	
20:45:38.941	faf55e	null	null	null	abc.de/sold	

<u>Note</u>: The ID field should be unique, but two different records share the same value!

• There are different **dimensions of data quality**, especially:

- Completeness: Do we have all the data we need to assess page load performance?
- Consistency: Does data have a valid format and does it comply with business semantics?
- Accuracy: Do data items represent their corresponding real-world entities well?

• Uniqueness: Are duplicate records known and are all unique attributes actually distinct?

Dimensions of Data Quality: Uniqueness



- There are different **dimensions of data quality**, especially:
 - Completeness: Do we have all the data we need to assess page load performance?
 - o Consistency: Does data have a valid format and does it comply with business semantics?
 - Accuracy: Do data items represent their corresponding real-world entities well?

• Uniqueness: Are duplicate records known and are all unique attributes actually distinct?

Dimensions of Data Quality: Uniqueness

Timestamp.	Pageload ID	Browser	LCP (Performance)	Session ID	URL
09:05:04.578	37ab08	Edge	670	123	null
13:26:48.139	9cddf7	Firefox	null	456	abc.de/red
13:28:23.857	0b577a	Firefox	256	456	abc.de/blue
13:29:17.468	faf55e	Firefox	1598	456	abc.de/sold
-20:45:38.941	faf55e	null	null	null	abc.de/sold

<u>Note</u>: The ID field should be unique, but two different records share the same value! → merge duplicates into a single record

• There are different **dimensions of data quality**, especially:

- Completeness: Do we have all the data we need to assess page load performance?
- Consistency: Does data have a valid format and does it comply with business semantics?
- Accuracy: Do data items represent their corresponding real-world entities well?

• Uniqueness: Are duplicate records known and are all unique attributes actually distinct?

Maintaining Data Quality With Constraints

- **Constraints** are rules, conditions, or limits that data must adhere to
- **Type Checks** represent expectations on the data format, e.g.
 - Value range for numerical data (e.g. [0,MAX_INTEGER) for load timers)
 - *Format or pattern* for string-valued data (e.g. ISO 8601 for timestamps)
 - *Structure* for complex attributes (e.g. required keys for JSON objects)
- Complex conditions can further describe complex semantics such as
 - Cross-field or cross-record relationships (e.g. same browser within sessions)
 - $\circ~$ Referential integrity between records in different collections
 - $\circ~$ Custom constraints for domain semantics

Example : Declarative Constraints With Pandera (1/3)

```
import pandas as pd
import pandera as pa
from pandera import Column, DataFrameSchema, Check
# Define the schema
schema = DataFrameSchema(
        "Timestamp": Column(pa.DateTime),
        "Pageload ID": Column(pa.String,
            Check(lambda x: x.str.len() == 8)),
        "Browser": Column(pa.String,
            Check(lambda x: x.isin(["Chrome", "Edge", "Firefox"]))),
        "LCP": Column(pa.Int,
            Check(lambda x: (x \ge 0) \& (x \le 60000))),
        "Session ID": Column(pa.Int),
    }
```

Example : Declarative Constraints With Pandera (2/3)

```
# valid data item
qood = \{
    "Timestamp": [pd.Timestamp("2023-05-10 13:26:48.139")],
    "Pageload ID": ["9cddf7"],
    "Browser": ["Firefox"],
    "LCP": [256],
    "Session ID": [456],
}
# invalid data item
bad = \{
    "Timestamp": [pd.Timestamp("2023-05-10 09:05:04.578")],
    "Pageload ID": ["37ab08"],
    "Browser": ["Edge"],
    "LCP": [692654], # timer value out of bounds
    "Session ID": [123],
}
```

Example : Declarative Constraints With Pandera (3/3)

```
for record in [good, bad]:
    record_id = record['Pageload ID'][0]
    try:
        validated = schema(pd.DataFrame(record))
        print(f"\nValidation passed for record {record_id}!")
    except pa.errors.SchemaError as e:
        print(f"\nValidation FAILED for record {record_id}:")
        print(e)
```

Validation passed for record 9cddf7!

Fundamental Challenge: Scalability

One Month in Data Errors at Bagend: April 2023



Challenging at Scale: Complexity

Timestamp	Pageload ID	Browser	LCP (Performance)	Session ID	URL		Session Length	Conversion
09:05:04.578	37ab08	Edge	670	123	null	•••	1	0
13:26:48.139	9cddf7	Firefox	null	456	abc.de/red			
13:28:23.857	0b577a	Firefox	256	456	abc.de/blue		3	1
13:29:17.468	faf55e	Firefox	1598	456	abc.de/sold			
-20:45:38.941	faf55e	null	null	null	abc.de/sold			

- Manual constraint definition is often infeasible, because of ...
 - o ... inherent data complexity (often <u>hundreds</u> of attributes)
 - ... aggregation, derived storage, and evolving schemas
 - ... a plethora of other data stores to integrate!

→ Automation is necessary!

Wolfram Wingerath (Hamburg, September 07, 2023)

Frequency

http archive

SimilarWeb

US DOL

2

3

Session Length

Challenging at Scale: Continuity



- Computing validation metrics from scratch periodically can be infeasible, because of ...
 - ... strict timing requirements
- ... efficiency or cost reasons
 ... data privacy reasons
 Details in Felix' talk!
 W. Wingerath, B. Wollmer, M. Bestehorn, S. Succo, F. Bücklers, J. Donnik, F. Panse, E. Witt, A. Sener, F. Gessert, N. Ritter. Beaconnect: Continuous Web Performance A/B-Testing at Scale, VLDB (2022)
 Incremental computation can be the only option!

Challenging at Scale: Volatility



• Specifying generalized constraints can be difficult in large deployments, because of ...

o ... temporal fluctuations (e.g. throughout the day, on black Friday, or during holydays)

o ... multi-tenancy (e.g. different data patterns by customer timezone or domain)

→ Elasticity & Multi-Tenancy requirements can be challenging!

Wolfram Wingerath (Hamburg, September 07, 2023)

Slides: <u>wolle.science/data-validation</u>

Challenging at Scale: "Continility" (Continuity + Volatility)



So How Do You Handle All This?

Advanced Techniques

- Inferring constraints
- $\circ~$ Adapting to schema changes
- \circ $\,$ Incremental computation of complex measures

• Tooling & Frameworks

- Validation libraries such as Great Expectations, Pandera, TFDV, or Deequ
- Preprocessing and validation with Apache Spark and Apache Flink
- Further Challenges
 - Handling distribution (validation per partitioning, avoiding skew, ...)
 - Efficiency and performance (load distribution, approximation, ...)
 - Operational challenges (anomaly detection, fixing, load shedding, ...)

Data Validation at Scale: Summary

- **Data Quality** can be measured along dimensions such as completeness, consistency, accuracy, and uniqueness
- Constraints specify expectations about the data and can be used to enforce them
- **Data Validation** is the process of ensuring high data quality for processes like analysis, modeling, and decision-making
- Data Validation Challenges at Scale include
 - *Complexity*: schemas are often too complex to define constraints manually
 - Volatility: data varies throughout the day, by season, or by customer
 - *Continuity*: incremental processing is required when computation from scratch is infeasible



Thanks! Questions ?



Material Available at https://wolle.science

Wolfram Wingerath (Hamburg, September 07, 2023)

Slides: wolle.science/data-validation