

Data Validation In A Data Centric Era

Validation and testing techniques through the different phases of a Data Science projects

By Aviram Berg

Definitions

Data Quality

Characterizes how reliable the information is to serve some intended purpose.

Definitions

Data Cleansing

Is the 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.

Definitions

Data Validation

The practice of checking the integrity, accuracy and structure of data before it is used for a business operation

Definitions

Data Verification

Is the process of checking a copy of data to make sure that it is exactly equal to the original copy of the data.

Why Data Quality Is Important?

Impact on organisational decisions

- missing or incorrect data can result in wrong decision making

Legal obligations in certain business scenarios

- Medtech, consumer facing products, etc'...

Impact on machine learning models

- Cleaner data can greatly improve model performance

Potential for causing biased decisions in ML-based systems

- Not well understood, area of active research

Operational stability: missing and inconsistent data can cause havoc in production systems

- Crashes (e.g., due to "NullPointerExceptions" for missing attributes)
- Wrong predictions (e.g., change of scale in attributes)

Data: Academia vs the Real-World

Academic datasets

- Static
- Often down-sampled, cleaned and aggregated before publication
- Attributes typically well understood
- Most of time: size convenient for processing on desktop machines
- Example: UCI ML datasets

Real-world data

- Constantly changing
- Often hundreds of attributes
- Data originates from multiple sources / people / teams / systems
- Several potentially inconsistent copies
- Often too large to conveniently handle on a desktop machine
- Often difficult to access (e.g., data compressed and partitioned in a distributed filesystem)

Sources of Error in Data

Data entry errors

- Typos in forms
- Different spellings for the same real-world entity (e.g., addresses, names)

Measurement errors

- Outside interference in measurement process
- Placement of sensors

Distillation errors

- Editorial bias in data summaries
- Domain-specific statistical analyses not understood by database manager

Data integration errors

- Resolution of inconsistencies w.r.t. duplicate entries
- Unification of units, measurement periods

Completeness

Uniqueness

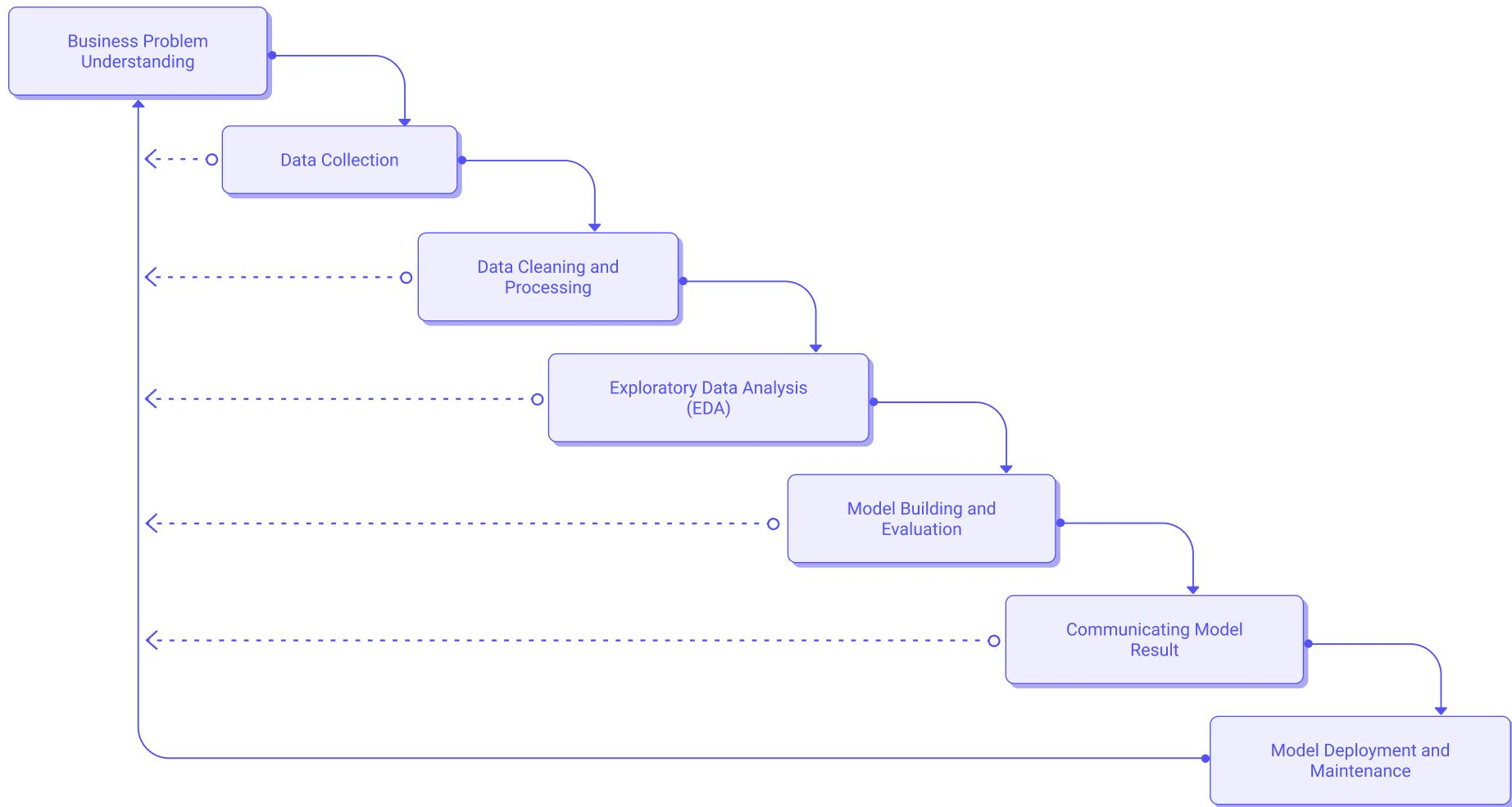
Consistency

Data Quality & Data Integrity

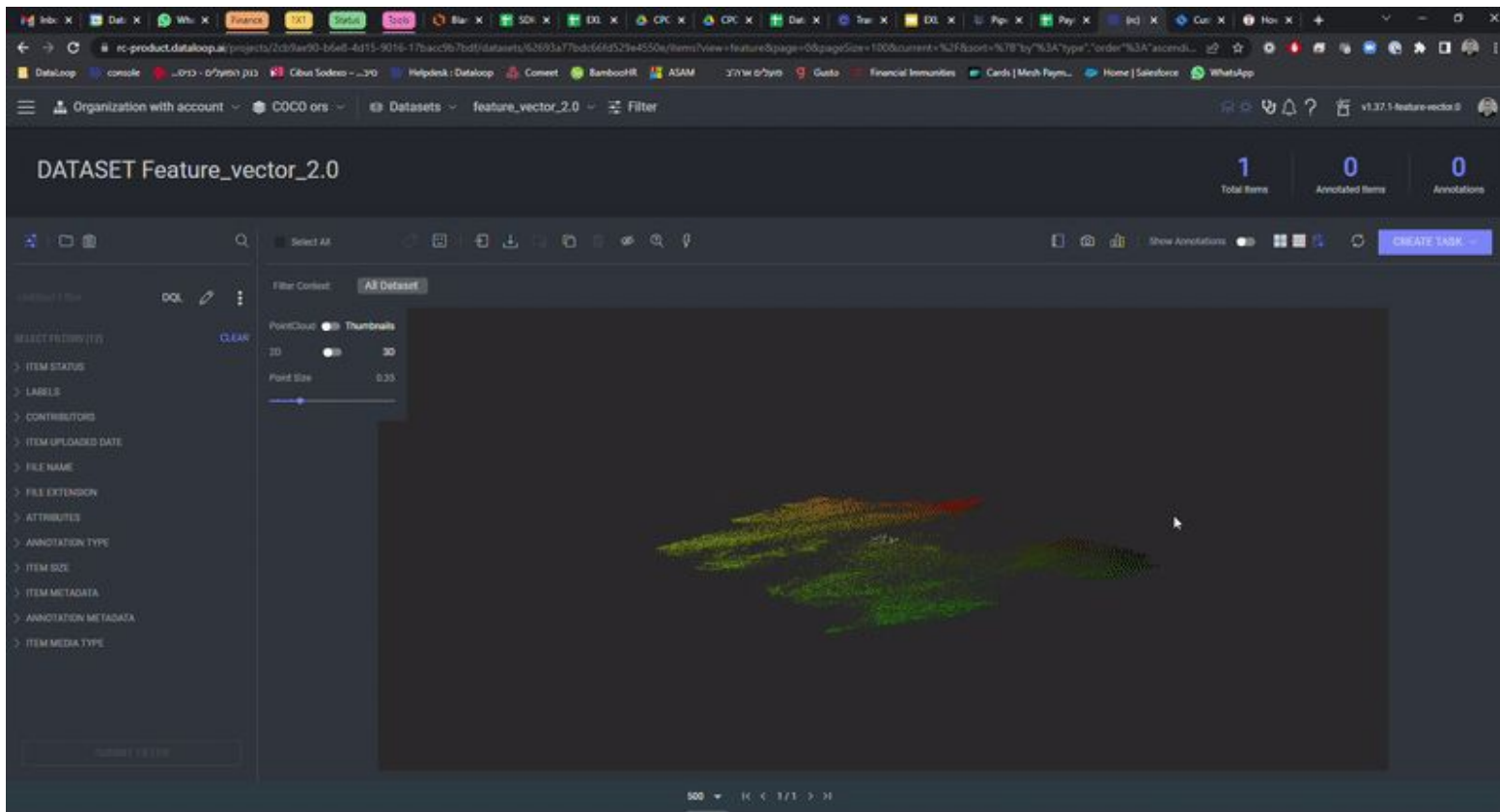
Timeliness

Accuracy

Validity



Data Collection -





Data Collection -

Filters

Data Distribution

Tags

☐ GOLD DATA

☐ PENDING

☐ VERIFIED

☐ FAILED (2)

☐ HAS NOTE

> Custom

> Advanced

☒ Confidence Level

0%

25%

Is the text below the image written in English, Blank or Written only in emojis?

each

Yes

Does the post show a commercial to a lifestyle product (i.e. fashion, beauty, home & garden products)?

each

Yes or Possibly yes

product_bb

each

Is the product in the box identical/similar/different to the one on the right?

each

Identical

Are these the same images (or part of the same image)?

each

Yes

No

Similar

Different

Are the products from the same category?

each

Yes

No

clean_data

Output: class

Output: class

Is_commercial_1

Output: class

Output: class

product_bb

Output: class

Similarity_test

Output: class

Output: class

Same_image

Output: class

Output: class

Output: class

Output: class

Same_category

Output: class

Output: class

Output: class

Polygon Task

Settings

Quality

Examples

Qualification Requirement (IoU)

95%

Accuracy measure for PolygonTask

*10% - 30% - Low. (Not recommended)

Judgments Per Resource

The number of different people you want to tag each image

3

Verify

Verify task submission with another Tasqer

☒

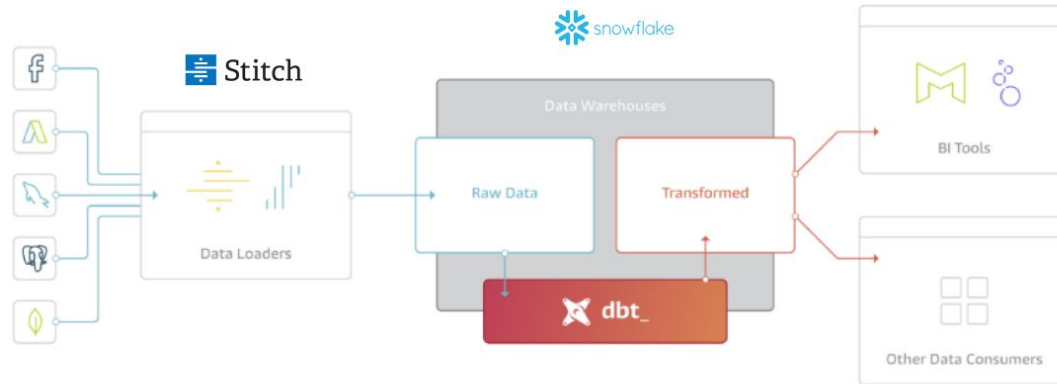
Verification instructions

Is the PLAYER outlined perfectly?

Quality Assurance

Task will be double checked

Data Cleaning & Processing



Dbt's key functions

Testing

- Dbt tests data quality, integration, and code performance
- Create test programs that check for missing/incomplete entries, unique constraints, and accepted values within specific columns
- Manually run scripts that will then run automated tests and deploy changes after passing said tests

Deployment

- Publish both public and private repositories that can then be referenced by other users

Documentation

- Automatically creates a visual representation of data flows throughout an organization
- Create documentation through schema files, reate documentation through schema files.

Dbt's Tests

```
version: 2

models:
  - name: orders
    columns:
      - name: order_id
        tests:
          - unique
          - not_null
      - name: status
        tests:
          - accepted_values:
              values: ['placed', 'shipped', 'completed', 'returned']
      - name: customer_id
        tests:
          - relationships:
              to: ref('customers')
              field: id
```

- `unique` : the `order_id` column in the `orders` model should be unique
- `not_null` : the `order_id` column in the `orders` model should not contain null values
- `accepted_values` : the `status` column in the `orders` should be one of `'placed'`, `'shipped'`, `'completed'`, or `'returned'`
- `relationships` : each `customer_id` in the `orders` model exists as an `id` in the `customers` table (also known as referential integrity)

Dbt's Tests

```
dbt test
```

dbt test

dbt_snowflake_workshop

Failed 4 0 1 0 0 02:22:15 10 seconds

RUN STATUS PASS WARN FAIL SKIPPED QUEUED START DURATION

SYSTEM LOGS

> view logs

DETAILS

> not_null_int_fx_rates_currency_exchange_date	2s	✓
> not_null_int_knoema_stock_history_company_symbol_stock_date	2s	✓
> not_null_int_unioned_book_instrument	1s	✓
> unique_int_fx_rates_currency_exchange_date	2s	✓
> unique_int_knoema_stock_history_company_symbol_stock_date	1s	⚠

Summary Details

02:22:24 | 5 of 5 START test unique_int_knoema_stock_history_company_symbol_stock_date [RUN]

02:22:25 | 5 of 5 FAIL 892281 unique_int_knoema_stock_history_company_symbol_stock_date [FAIL 892281 in 1.49s]

Dbt's Documentation



Search for models...

Overview

Project

Database

Tables and Views

- analytics
 - dbt_claire_playbook
 - customer_churn_month
 - customer_revenue_by_month
 - mrr
 - subscription_periods
 - util_months

mrr view

Details Description Columns SQL

Details

TAGS	OWNER	TYPE	PACKAGE	RELATION
untagged	TRANSFORMER	view	acme	analytics.dbt_claire_playbook.mrr

Description

This model represents one record per month, per account (months have been filled in to include any periods of inactivity).

This model classifies each month as one of: new, reactivation, upgrade, downgrade, or churn.

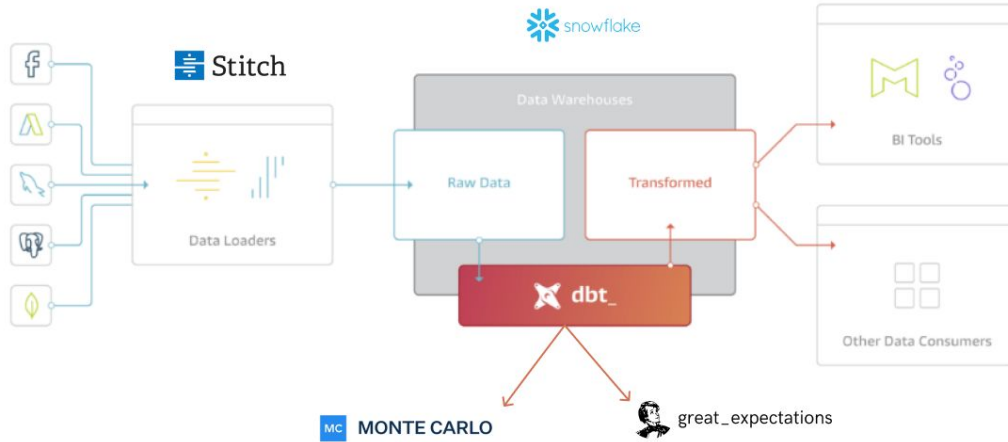
Columns

COLUMN	TYPE	DESCRIPTION
id	TEXT	

Lineage Graph



Data Cleaning & Processing



Great Expectations

Expectations: 270 Total	<input type="text" value="Search here..."/>	<input type="button" value="Filter by"/>	<input type="button" value="Summary"/>	<input type="button" value="Completeness"/>	<input type="button" value="Datasource"/>
	Experimental	Beta	Production		
<input type="radio"/> expect_column_distinct_values_to_equal_set	<input checked="" type="checkbox"/> <input type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> <input type="checkbox"/>		
<input type="radio"/> expect_column_distribution_to_match_benford's_law	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	<input type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>		
<input type="radio"/> expect_column_kurtosis_to_be_between	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	<input type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>		
<input checked="" type="radio"/> expect_column_max_to_be_between	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>		
<input checked="" type="radio"/> expect_column_mean_to_be_between	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>		
<div><div>< 1 2 3 4 5 ></div><div>5 / Page</div><div>Go To</div></div>					
Can't find the Expectation you need? Learn how to create an Expectation here					

Great Expectations

Expectations:

270 Total

Search here...

Filter by

Summary

Completeness

Datasource

<input type="radio"/> expect_column_distinct_values_to_equal_set	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="radio"/> expect_column_distribution_to_match_benford's_law	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="radio"/> expect_column_kurtosis_to_be_between	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="radio"/> expect_column_max_to_be_between	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="radio"/> expect_column_mean_to_be_between	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

<

1

2

3

4

5

>

5 / Page

Go To

Can't find the Expectation you need?

Learn how to create an Expectation here



Completeness



Uniqueness

Consistency



Data Quality & Data Integrity

Timeliness



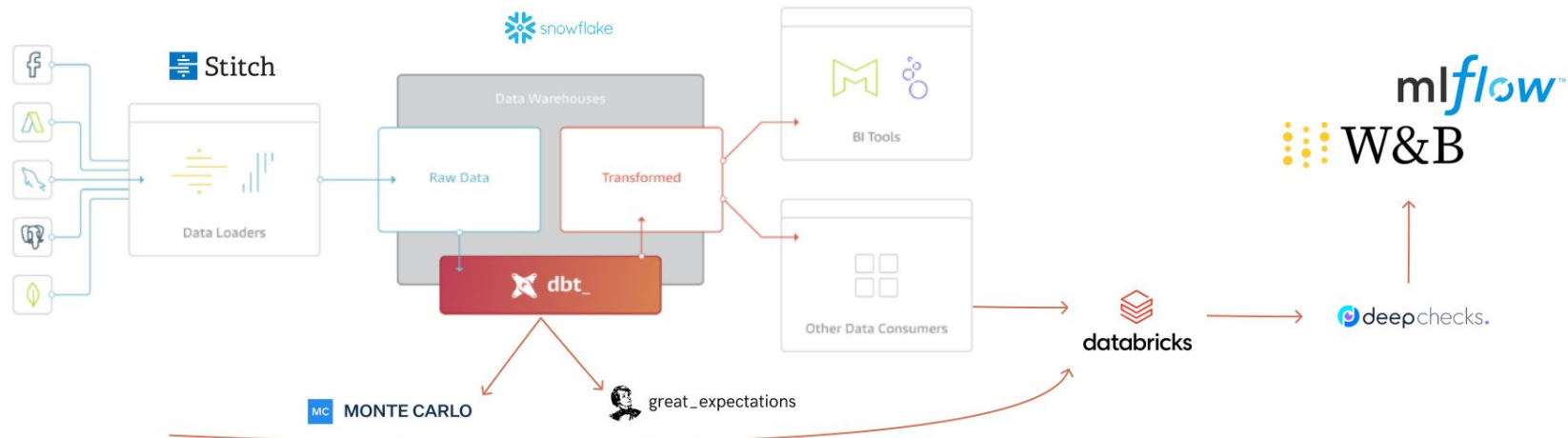
Accuracy



Validity



Unstructured Pipeline



Deepchecks

Property "Brightness"

Total number of outliers: 6

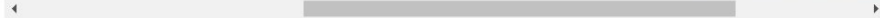
Non-outliers range: 0.24 to 0.69

0.12

0.22

0.71

0.72



Property "RMS Contrast"

Total number of outliers: 3

Non-outliers range: 0.1 to 0.37

Contrast

0.07

0.38

0.4

Image



Deepchecks

Similar Images

Total number of test samples with similar images in train: 5

Samples

Train



Test



```
check = SimilarImageLeakage().add_condition_similar_images_less_or_equal(3)
result = check.run(train_dataset=train_ds, test_dataset=test_ds_modified)
result.show(show_additional_outputs=False)
```