Data Validation with Great Expectations

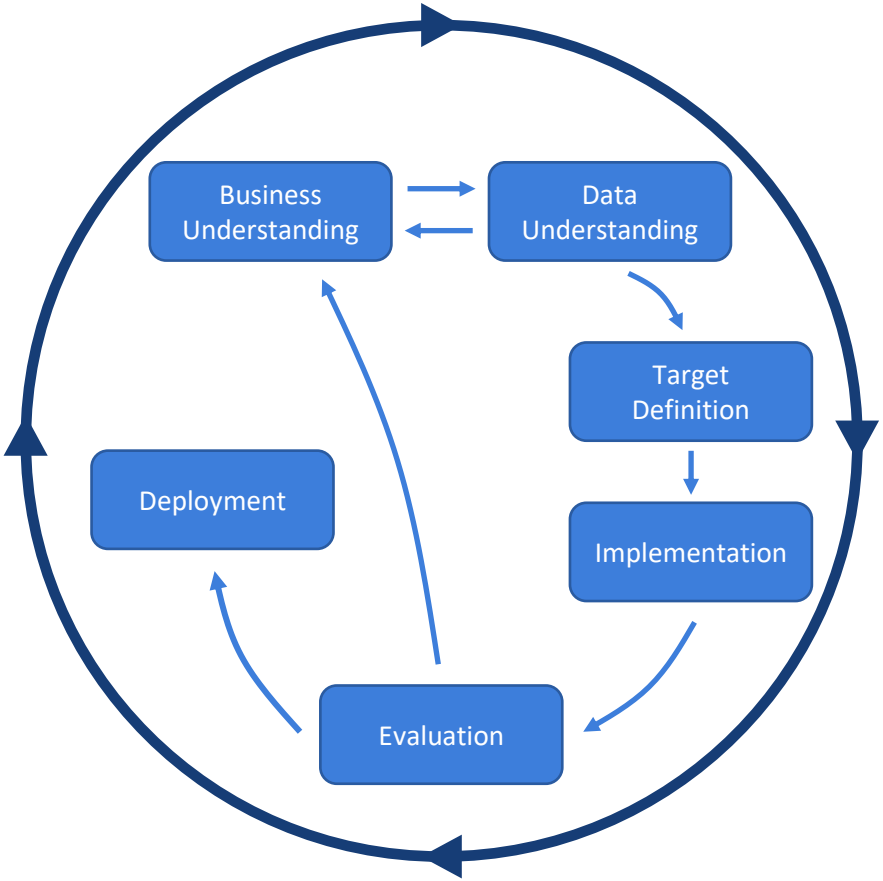
AI Monday Leipzig

Frank Stumpf | 08.06.2020

IT SONIX

The begin of a data story

IT SONIX



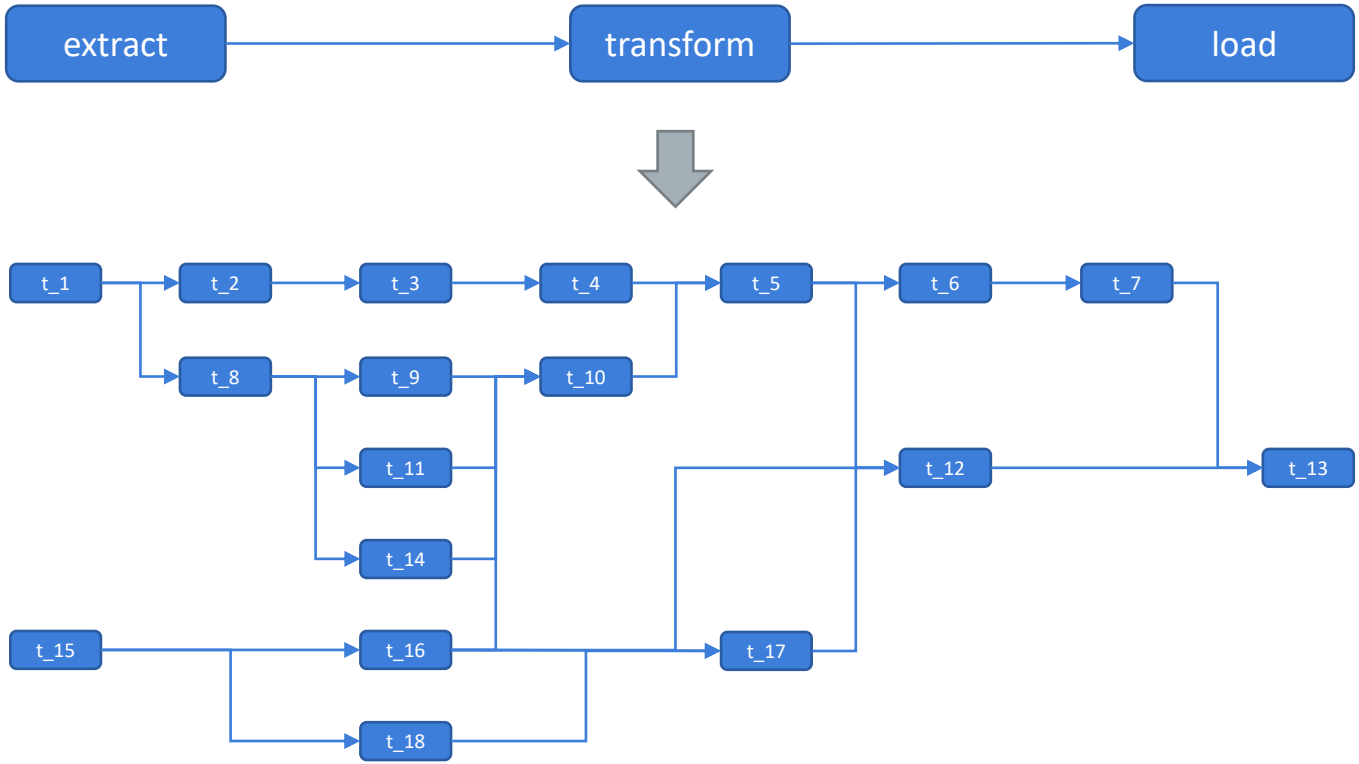
Two months later (or so) ...

IT SONIX

Reason #1

Reason #1

IT SONIX



Reason #2

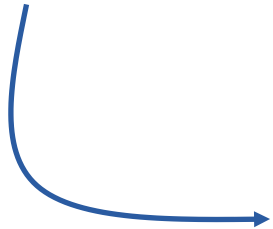
IT SONIX

id	account_status	birth_date	postal_code	active
e57j3yiq	APPROVED	01.03.1972	null	Y
t19k7qdm	APPROVED	18.05.1983	04103	N
c82u9awb	SUSPENDED	10.02.1981	null	N
p25i1gbs	APPROVED	23.09.1987	null	Y

Reason #2

IT SONIX

id	account_status	birth_date	postal_code	active
e57j3yiq	APPROVED	01.03.1972	null	Y
t19k7qdm	APPROVED	18.05.1983	04103	N
c82u9awb	SUSPENDED	10.02.1981	null	N
p25i1gbs	APPROVED	23.09.1987	null	Y



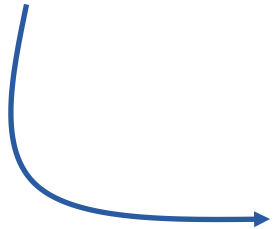
id	account_status	birthDate	postal_code	active	wallet
2020_384	VERIFIED	null	None	true	35.87
2019_147	APPROVED	null	04103	false	2.54
2020_502	SUSPENDED	null	None	false	0.0

Reason #2

IT SONIX

id	account_status	birth_date	postal_code	active
e57j3yiq	APPROVED	01.03.1972	null	Y
t19k7qdm	APPROVED	18.05.1983	04103	N
c82u9awb	SUSPENDED	10.02.1981	null	N
p25i1gbs	APPROVED	23.09.1987	null	Y

- new columns



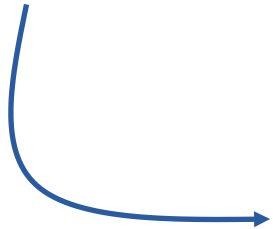
id	account_status	birthDate	postal_code	active	wallet
2020_384	VERIFIED	null	None	true	35.87
2019_147	APPROVED	null	04103	false	2.54
2020_502	SUSPENDED	null	None	false	0.0

Reason #2

IT SONIX

id	account_status	birth_date	postal_code	active
e57j3yiq	APPROVED	01.03.1972	null	Y
t19k7qdm	APPROVED	18.05.1983	04103	N
c82u9awb	SUSPENDED	10.02.1981	null	N
p25i1gbs	APPROVED	23.09.1987	null	Y

- new columns
- column names changed



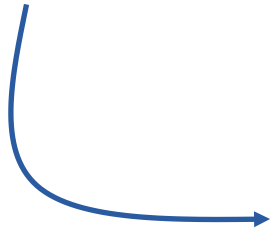
id	account_status	birthDate	postal_code	active	wallet
2020_384	VERIFIED	null	None	true	35.87
2019_147	APPROVED	null	04103	false	2.54
2020_502	SUSPENDED	null	None	false	0.0

Reason #2

IT SONIX

id	account_status	birth_date	postal_code	active
e57j3yiq	APPROVED	01.03.1972	null	Y
t19k7qdm	APPROVED	18.05.1983	04103	N
c82u9awb	SUSPENDED	10.02.1981	null	N
p25i1gbs	APPROVED	23.09.1987	null	Y

- new columns
- column names changed
- new data format



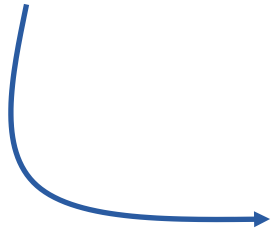
id	account_status	birthDate	postal_code	active	wallet
2020_384	VERIFIED	null	None	true	35.87
2019_147	APPROVED	null	04103	false	2.54
2020_502	SUSPENDED	null	None	false	0.0

Reason #2

IT SONIX

id	account_status	birth_date	postal_code	active
e57j3yiq	APPROVED	01.03.1972	null	Y
t19k7qdm	APPROVED	18.05.1983	04103	N
c82u9awb	SUSPENDED	10.02.1981	null	N
p25i1gbs	APPROVED	23.09.1987	null	Y

- new columns
- column names changed
- new data format
- data type changed



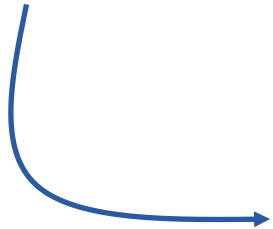
id	account_status	birthDate	postal_code	active	wallet
2020_384	VERIFIED	null	None	true	35.87
2019_147	APPROVED	null	04103	false	2.54
2020_502	SUSPENDED	null	None	false	0.0

Reason #2

IT SONIX

id	account_status	birth_date	postal_code	active
e57j3yiq	APPROVED	01.03.1972	null	Y
t19k7qdm	APPROVED	18.05.1983	04103	N
c82u9awb	SUSPENDED	10.02.1981	null	N
p25i1gbs	APPROVED	23.09.1987	null	Y

- new columns
- column names changed
- new data format
- data type changed
- new possible field values



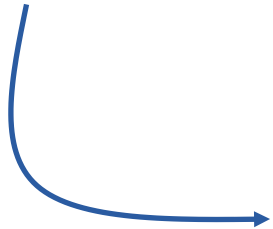
id	account_status	birthDate	postal_code	active	wallet
2020_384	VERIFIED	null	None	true	35.87
2019_147	APPROVED	null	04103	false	2.54
2020_502	SUSPENDED	null	None	false	0.0

Reason #2

IT SONIX

id	account_status	birth_date	postal_code	active
e57j3yiq	APPROVED	01.03.1972	null	Y
t19k7qdm	APPROVED	18.05.1983	04103	N
c82u9awb	SUSPENDED	10.02.1981	null	N
p25i1gbs	APPROVED	23.09.1987	null	Y

- new columns
- column names changed
- new data format
- data type changed
- new possible field values
- missing column values



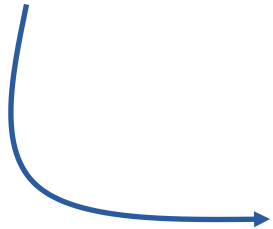
id	account_status	birthDate	postal_code	active	wallet
2020_384	VERIFIED	null	None	true	35.87
2019_147	APPROVED	null	04103	false	2.54
2020_502	SUSPENDED	null	None	false	0.0

Reason #2

IT SONIX

id	account_status	birth_date	postal_code	active
e57j3yiq	APPROVED	01.03.1972	null	Y
t19k7qdm	APPROVED	18.05.1983	04103	N
c82u9awb	SUSPENDED	10.02.1981	null	N
p25i1gbs	APPROVED	23.09.1987	null	Y

- new columns
- column names changed
- new data format
- data type changed
- new possible field values
- missing column values
- stringified null values



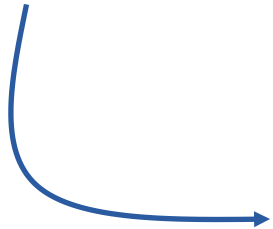
id	account_status	birthDate	postal_code	active	wallet
2020_384	VERIFIED	null	None	true	35.87
2019_147	APPROVED	null	04103	false	2.54
2020_502	SUSPENDED	null	None	false	0.0

Reason #2

IT SONIX

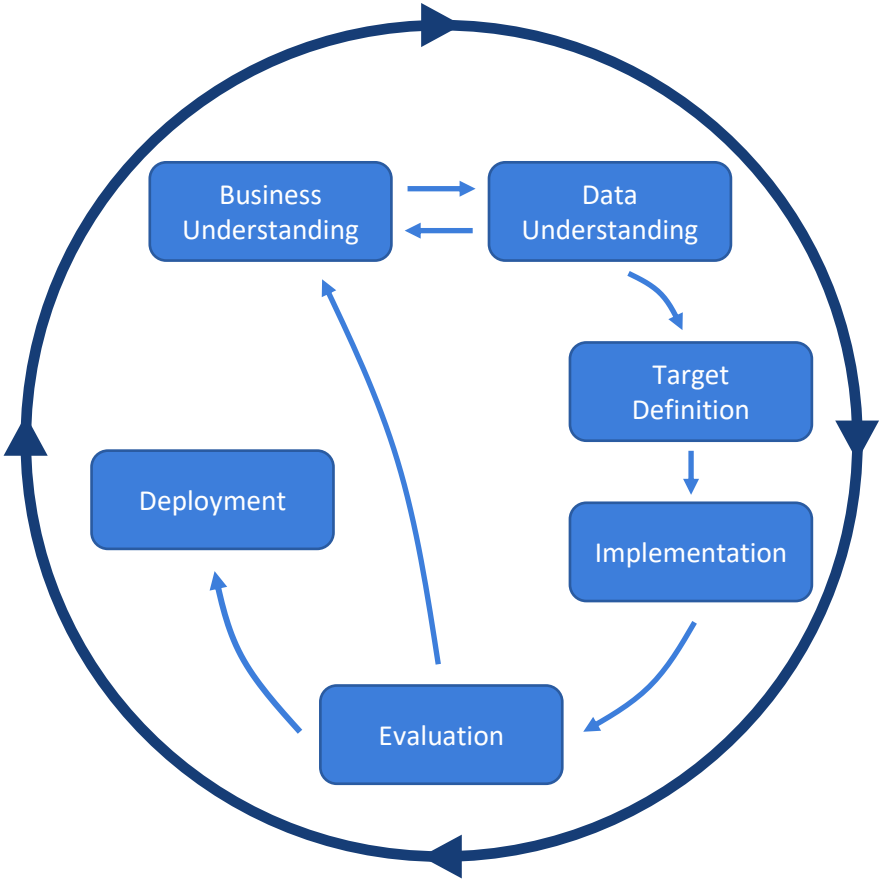
id	account_status	birth_date	postal_code	active
e57j3yiq	APPROVED	01.03.1972	null	Y
t19k7qdm	APPROVED	18.05.1983	04103	N
c82u9awb	SUSPENDED	10.02.1981	null	N
p25i1gbs	APPROVED	23.09.1987	null	Y

- new columns
- column names changed
- new data format
- data type changed
- new possible field values
- missing column values
- stringified null values
- missing data entries

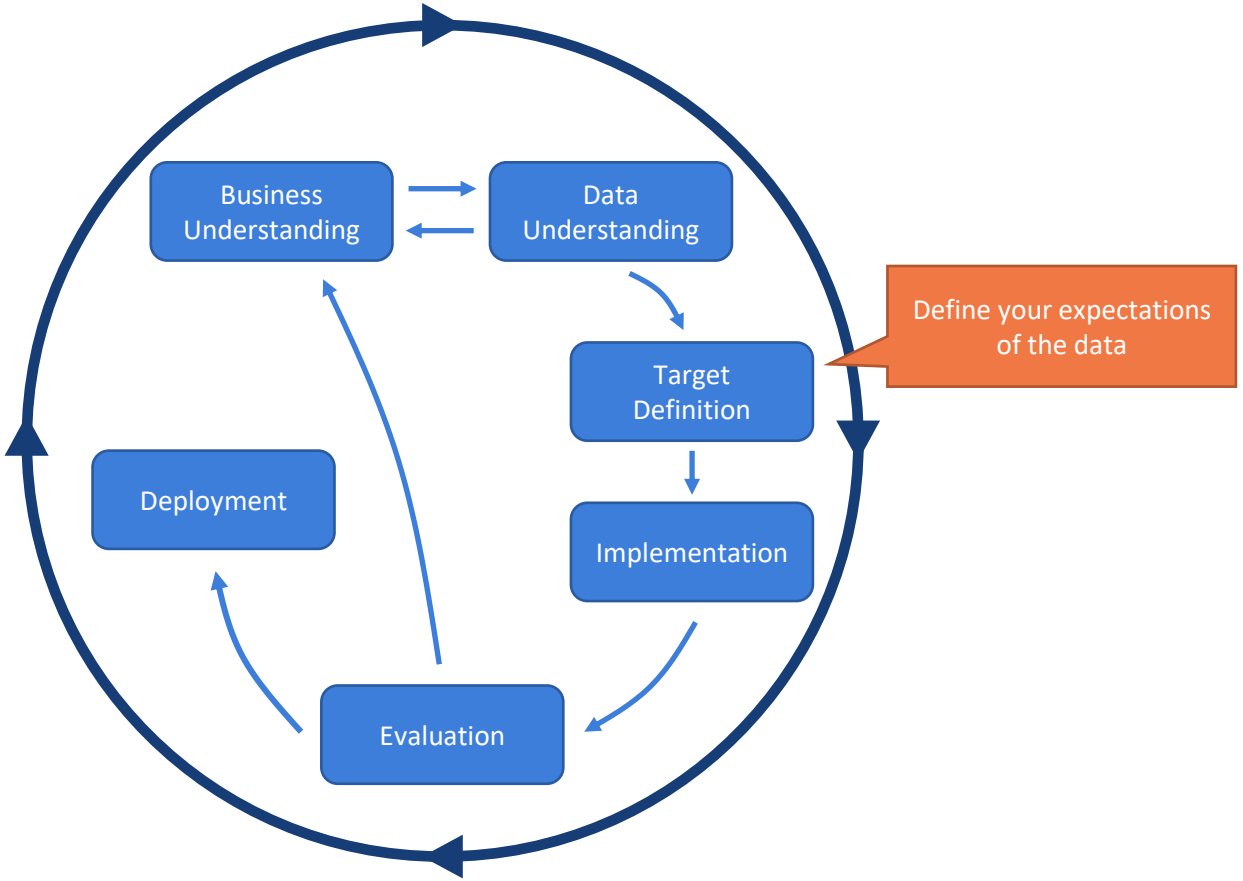


id	account_status	birthDate	postal_code	active	wallet
2020_384	VERIFIED	null	None	true	35.87
2019_147	APPROVED	null	04103	false	2.54
2020_502	SUSPENDED	null	None	false	0.0

How to overcome?



How to overcome?



Great expectations with Great Expectations

IT SONIX

- Python library for making expectations about your data
- open source since December 2017
- under Apache License 2.0
- 2k stars on GitHub
- under active development

[1] https://github.com/great-expectations/great_expectations

More than 50 built-in expectations

IT SONIX

Table shape

- `expect_column_to_exist`
- `expect_table_columns_to_match_ordered_list`
- `expect_table_row_count_to_be_between`
- `expect_table_row_count_to_equal`

Missing values, unique values, and types

- `expect_column_values_to_be_unique`
- `expect_column_values_to_not_be_null`
- `expect_column_values_to_be_null`
- `expect_column_values_to_be_of_type`
- `expect_column_values_to_be_in_type_list`

Sets and ranges

- `expect_column_values_to_be_in_set`
- `expect_column_values_to_not_be_in_set`
- `expect_column_values_to_be_between`
- `expect_column_values_to_be_increasing`
- `expect_column_values_to_be_decreasing`

String matching

- `expect_column_value_lengths_to_be_between`
- `expect_column_value_lengths_to_equal`
- `expect_column_values_to_match_regex`
- `expect_column_values_to_not_match_regex`
- `expect_column_values_to_match_regex_list`
- `expect_column_values_to_not_match_regex_list`

Datetime and JSON parsing

- `expect_column_values_to_match_strftime_format`
- `expect_column_values_to_be_dateutil_parseable`
- `expect_column_values_to_be_json_parseable`
- `expect_column_values_to_match_json_schema`

Aggregate functions

- `expect_column_distinct_values_to_be_in_set`
- `expect_column_distinct_values_to_contain_set`
- `expect_column_distinct_values_to_equal_set`
- `expect_column_mean_to_be_between`
- `expect_column_median_to_be_between`
- `expect_column_quantile_values_to_be_between`
- `expect_column_stdev_to_be_between`
- `expect_column_unique_value_count_to_be_between`
- `expect_column_proportion_of_unique_values_to_be_between`
- `expect_column_most_common_value_to_be_in_set`
- `expect_column_max_to_be_between`
- `expect_column_min_to_be_between`
- `expect_column_sum_to_be_between`

Multi-column

- `expect_column_pair_values_A_to_be_greater_than_B`
- `expect_column_pair_values_to_be_equal`
- `expect_column_pair_values_to_be_in_set`
- `expect_multicolumn_values_to_be_unique`

Distributional functions

- `expect_column_kl_divergence_to_be_less_than`
- `expect_column_bootstrapped_ks_test_p_value_to_be_greater_than`
- `expect_column_chisquare_test_p_value_to_be_greater_than`
- `expect_column_parameterized_distribution_ks_test_p_value_to_be_greater_than`

FileDataAsset

File data assets reason at the file level, and the line level (for text data).

- `expect_file_line_regex_match_count_to_be_between`
- `expect_file_line_regex_match_count_to_equal`
- `expect_file_hash_to_equal`
- `expect_file_size_to_be_between`
- `expect_file_to_exist`
- `expect_file_to_have_valid_table_header`
- `expect_file_to_be_valid_json`

All or nothing?

```
expect_column_values_to_not_be_null("Age", mostly=0.8)
```

- at least 80 % of “Age“ values are not null

Where does the data come from?

IT SONIX

[1] https://de.wikipedia.org/wiki/Pandas_(Software)
[2] https://www.sqlalchemy.org/
[3] https://en.wikipedia.org/wiki/Snowflake_Inc.
[4] https://de.wikipedia.org/wiki/MySQL
[5] https://de.wikipedia.org/wiki/Apache_Spark

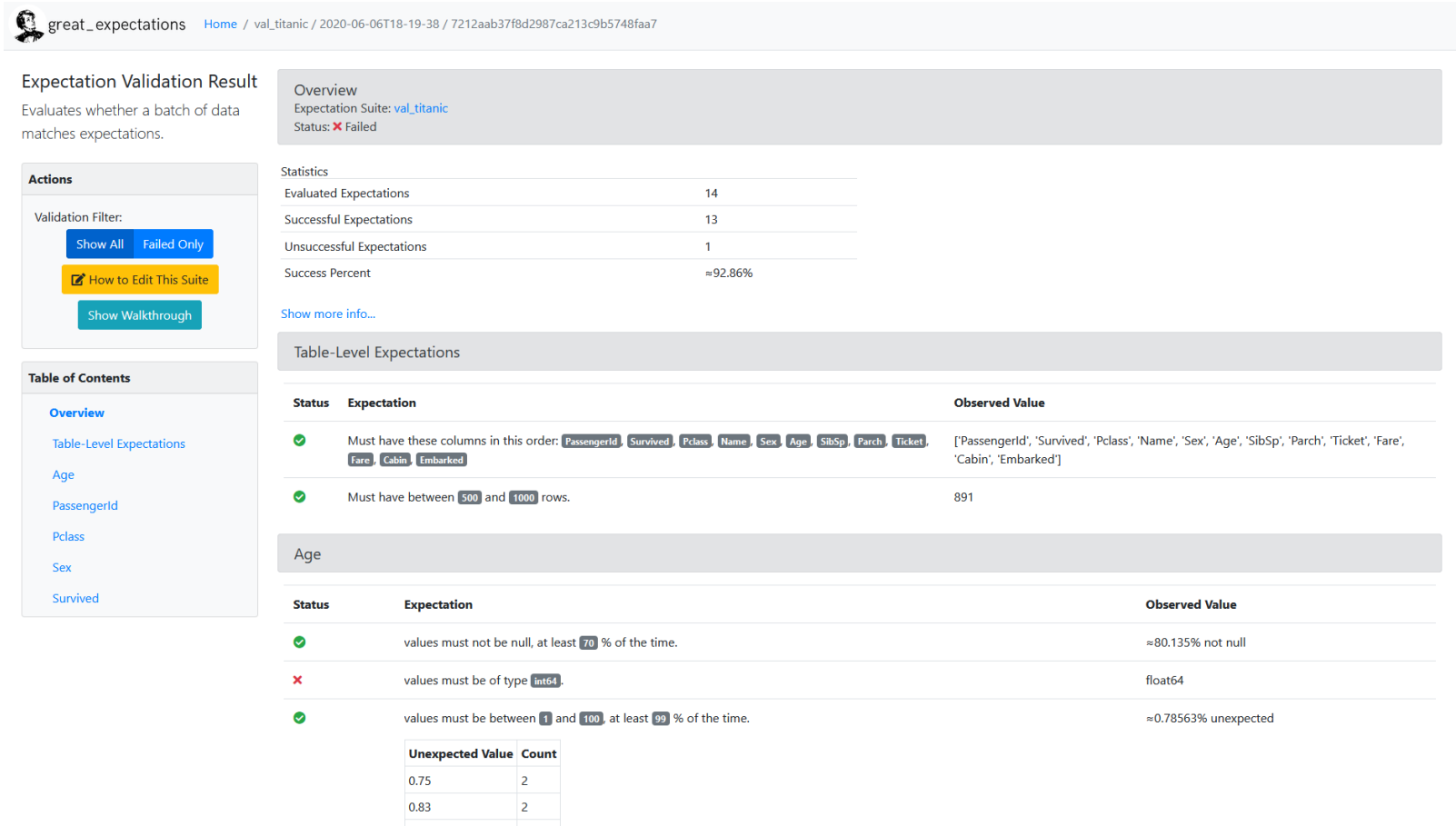
Making expectations is a breeze

IT SONIX

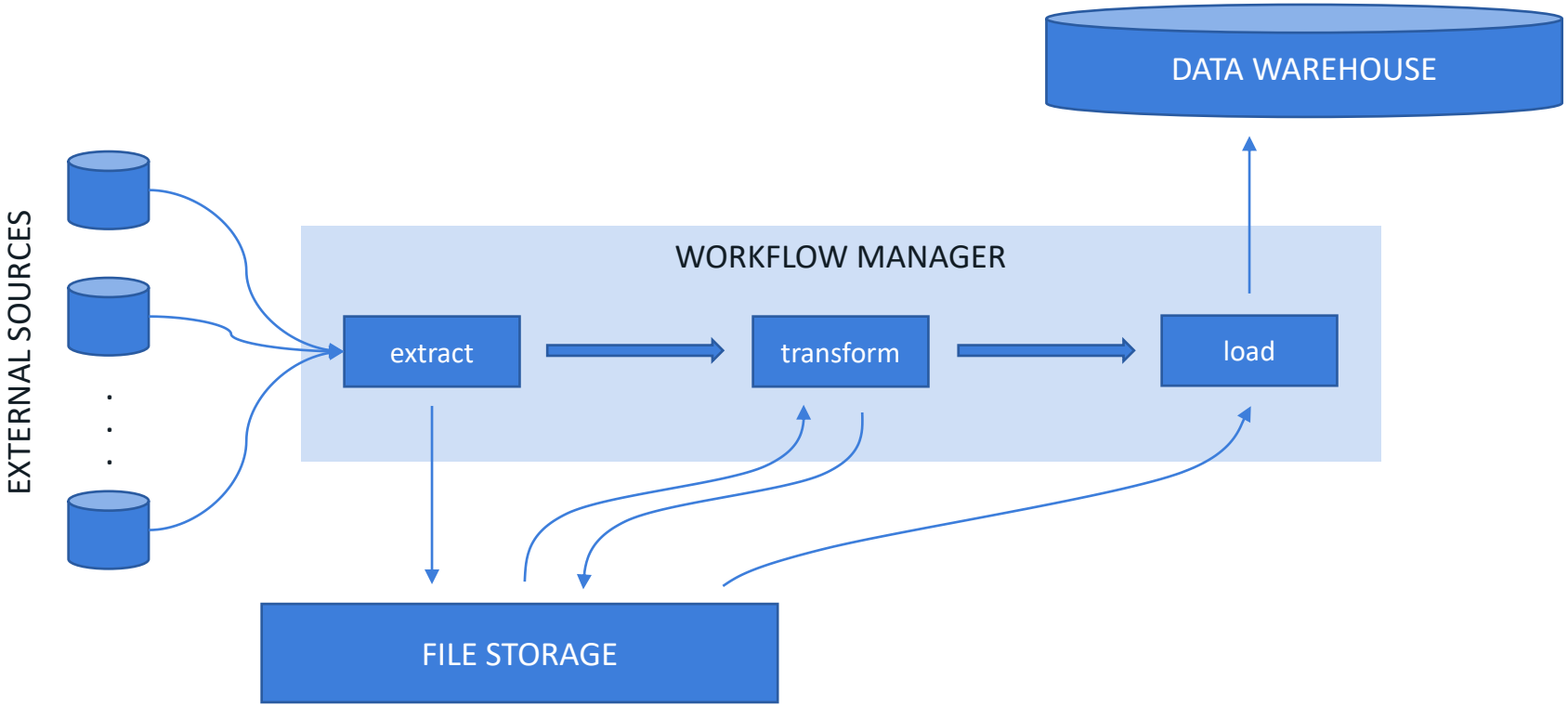
```
context = DataContext()
suite_name = "val_titanic"
context.create_expectation_suite(suite_name, overwrite_existing=True)
batch_kwargs = {
    "path": "titanic.csv",
    "datasource": "files_datasource"
}
batch = context.get_batch(batch_kwargs, suite_name)

batch.expect_table_columns_to_match_ordered_list(["PassengerId", "Survived", "Pclass", "Name", "Sex", "Age",
                                                  "SibSp", "Parch", "Ticket", "Fare", "Cabin", "Embarked"])
batch.expect_table_row_count_to_be_between(min_value=500, max_value=1000)

batch.expect_column_values_to_not_be_null("PassengerId")
batch.expect_column_values_to_be_unique("PassengerId")
batch.expect_column_values_to_be_of_type("PassengerId", "int64")

batch.expect_column_values_to_not_be_null("Survived")
batch.expect_column_values_to_be_in_set("Survived", ["0", "1"])

batch.expect_column_values_to_not_be_null("Pclass")
batch.expect_column_values_to_be_in_set("Pclass", ["1", "2", "3"])

batch.expect_column_values_to_not_be_null("Sex", mostly=0.95)
batch.expect_column_values_to_be_in_set("Sex", ["male", "female"])

batch.expect_column_values_to_not_be_null("Age", mostly=0.7)
batch.expect_column_values_to_be_of_type("Age", "int64")
batch.expect_column_values_to_be_between("Age", min_value=1, max_value=100, mostly=0.99)

context.run_validation_operator("action_list_operator",
                                assets_to_validate=[batch],
                                run_id=arrow.utcnow().format("YYYY-MM-DDTHH-mm-ss"))
```

Finally you get human-readable Data Docs

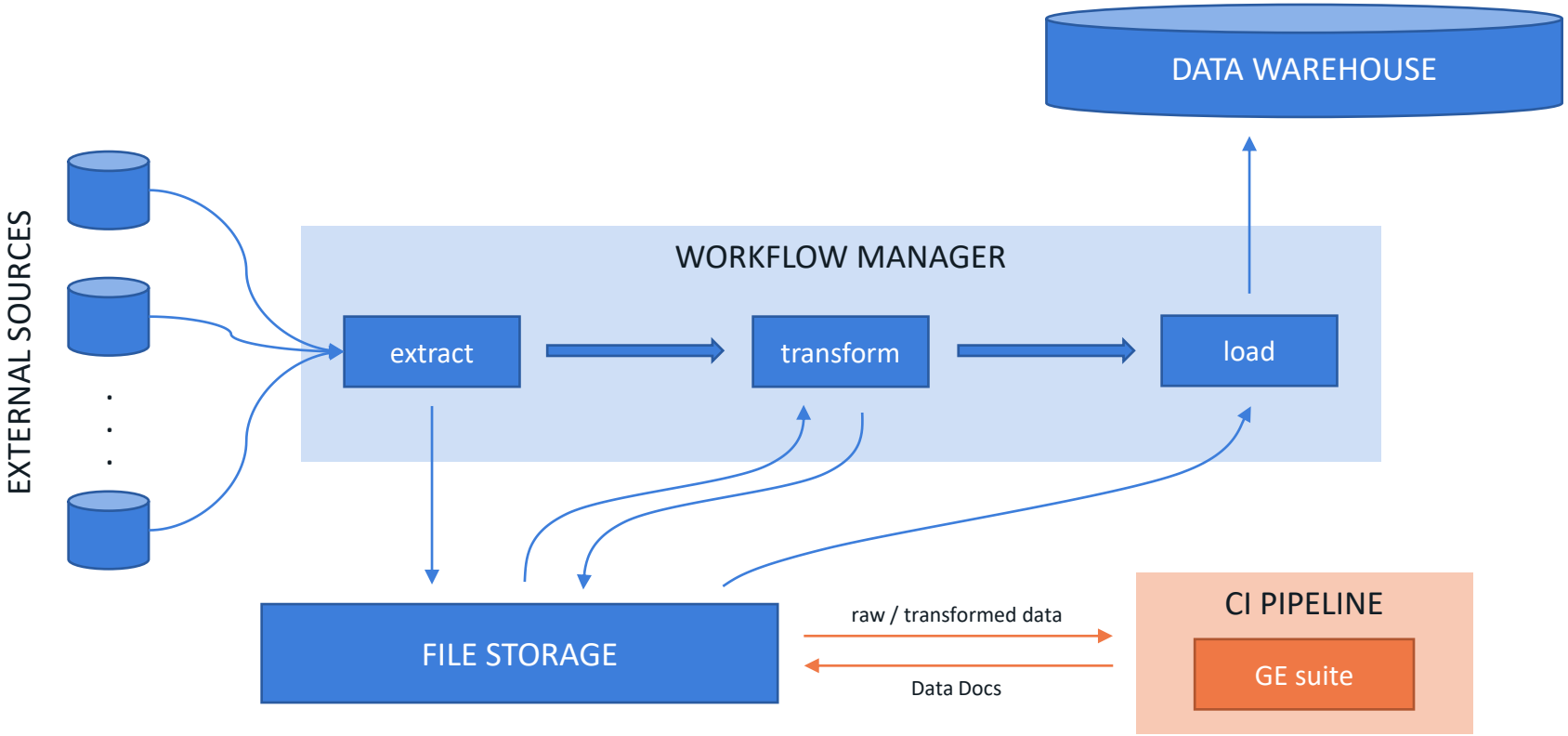
IT SONIX



How to integrate?



How to integrate?



Thumb up or down?

IT SONIX

- ✓ simple to use
- ✓ more than 50 built-in expectations
- ✓ easy to write custom expectations
- ✓ pandas, SQL databases and Spark as data sources
- ✓ Data Docs as HTML
- ✓ under active development

- ✗ still some bugs
- ✗ limited number of expectations for SQL databases and Spark
- ✗ no native support for some common databases
- ✗ problematic integration in Airflow

Thumb up or down?

IT SONIX

- ✓ simple to use
- ✓ more than 50 built-in expectations
- ✓ easy to write custom expectations
- ✓ pandas, SQL databases and Spark as data sources
- ✓ Data Docs as HTML
- ✓ under active development

- ✗ still some bugs
- ✗ limited number of expectations for SQL databases and Spark
- ✗ no native support for some common databases
- ✗ problematic integration in Airflow

➢ *Great Expectations is very promising and worth a closer look*

Contact

Frank Stumpf

Team Lead Data Science

Mail: f.stumpf@itsonix.eu
Phone: +49 341 35576 406
Mobile: +49 172 407 2058

IT SONIX