DATA+AI SUMMIT
BY databricks

DATA QUALITY

The Greatest Challenge for GenAI Enterprise Adoption

Amy Reams and Taly Kanfi, Anomalo
June 12, 2024

Enterprise Data Quality is Broken

91%

Think their company's data quality needs improvement

82%

IT decision-makers reworked data projects due to poor data quality

Amy Reams
VP, Business Development
amy@anomalo.com

Taly Kanfi
Director, Data Solutions Architect
taly@anomalo.com

Come Visit Us at Booth 46

Anomalo

We Are Rethinking Data Quality

And We're Rethinking Data Quality With Databricks

Winner

2024 Databricks
Partner Awards

Emerging Partner of the Year

databricks

Anomalo

Impact Award

1. High Quality Data is the New Gold
2. Anomalo and Databricks
3. Anomalo's AI-Powered Data Quality Monitoring
4. Data Quality for GenAI

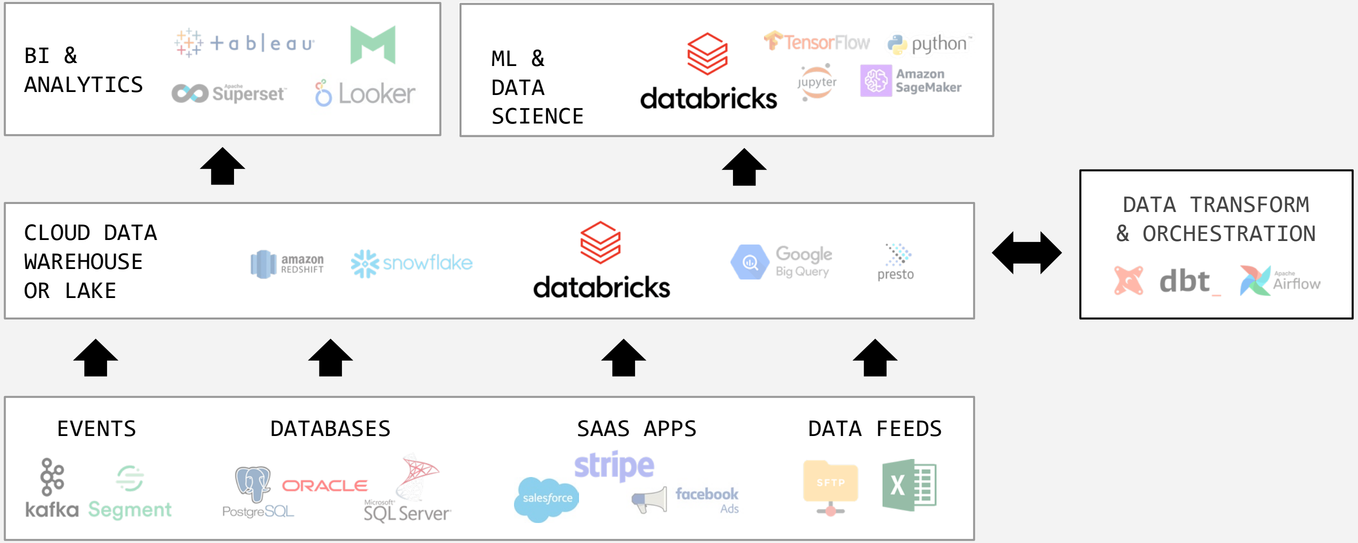
Data is the New Gold

Source: Salesforce.com

High Quality Data is the New Gold

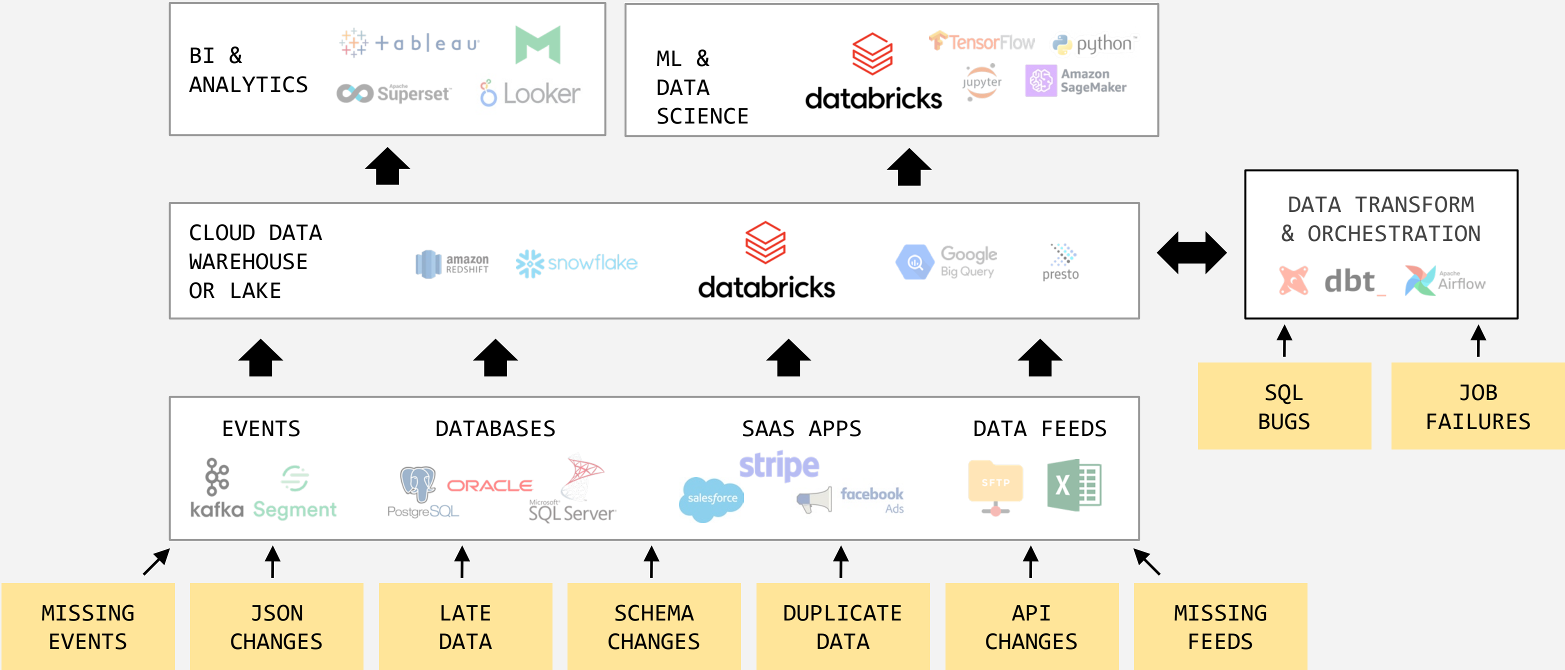
The Gold Mining Challenge

Your Modern Data Stack Is Only As Good As The Quality Of Your Data



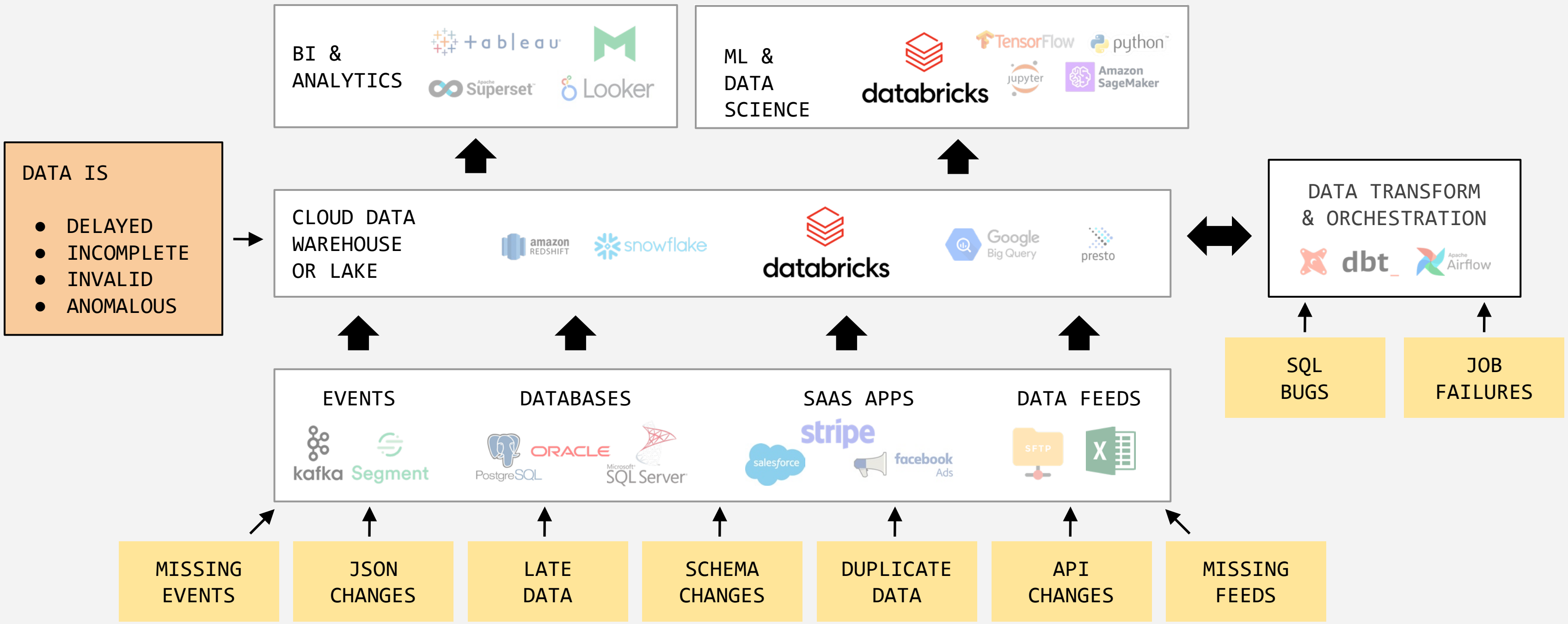
The Gold Mining Challenge

Your Modern Data Stack Is Only As Good As The Quality Of Your Data



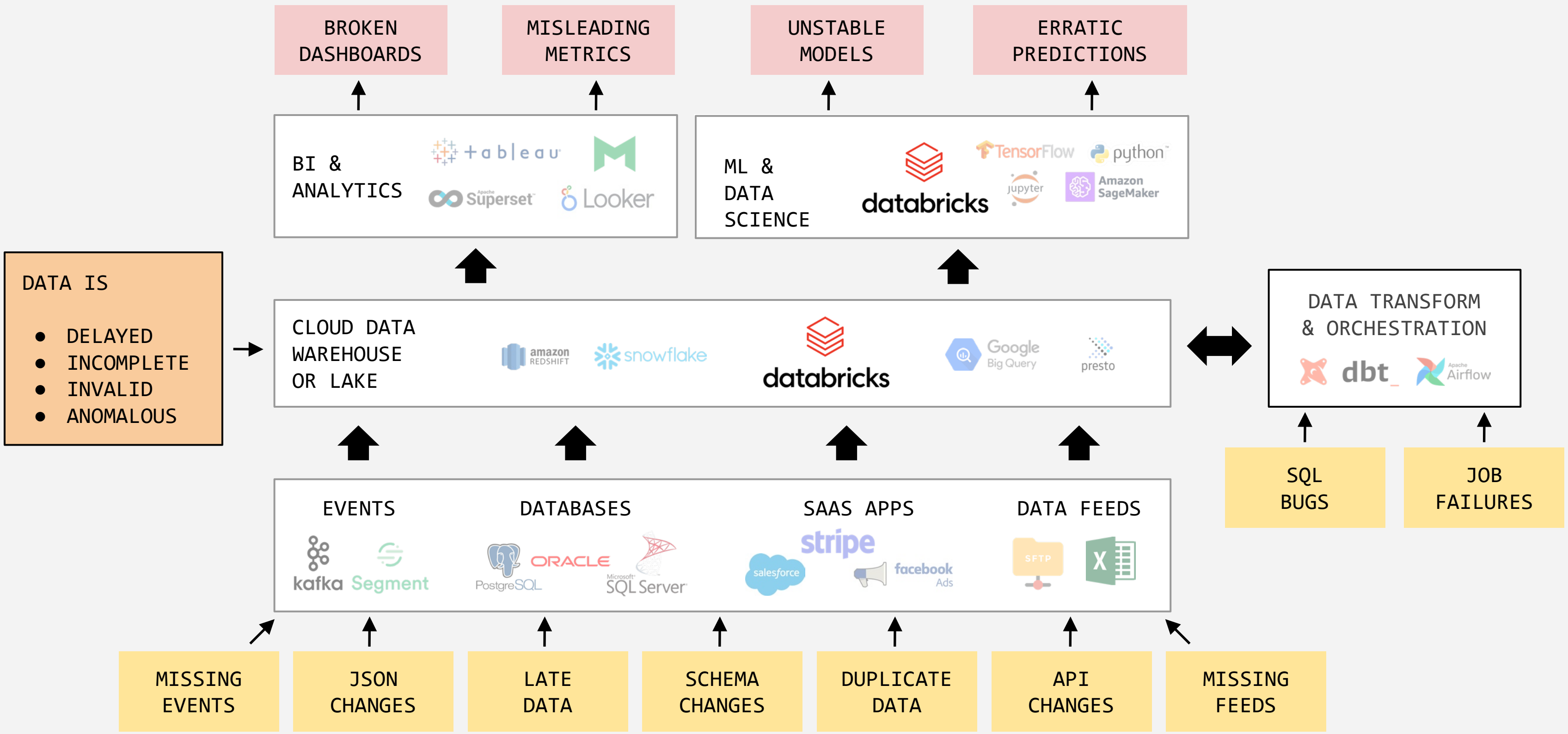
The Gold Mining Challenge

Your Modern Data Stack Is Only As Good As The Quality Of Your Data



The Gold Mining Challenge

Your Modern Data Stack Is Only As Good As The Quality Of Your Data



Gen AI will only make these problems worse

Formatting

Does not comply with traditional standard formats

Retrieval **Search** **Storage**

Sensitive Information

Affect specific value or cells of data

PII **Length** **Abusive Language**

Inconsistences

Full of errors and duplicated content.

Incomplete **Duplicates** **Temporal Inconsistency**

Sentiment

How do you evaluate the tone of a document?

Accuracy **Document Characteristics** **Tone**

Anomalo and Databricks

Winner

2024 Databricks
Partner Awards

Emerging Partner of the Year

databricks

Anomalo

Impact Award

Data Observability

Is data moving through my warehouse in a timely manner?

- Catalog metadata
- Job monitoring
- Data lineage

✅ Catches data *movement* failures

⚠️ Ignores data *contents*

Data Quality Monitoring

Is the data my warehouse produces of high quality?

- Queries the data
- Requires experts or ML
- Explainability is key

✅ Deep monitoring of data

⚠️ More work to scale

Data Observability

Is data moving through my warehouse in a timely manner?

- Catalog metadata
- Job monitoring
- Data lineage

✅ Catches data *movement* failures

⚠️ Ignores data *contents*

Data Quality Monitoring

Is the data my warehouse produces of high quality?

- Queries the data
- Requires experts or ML
- Explainability is key

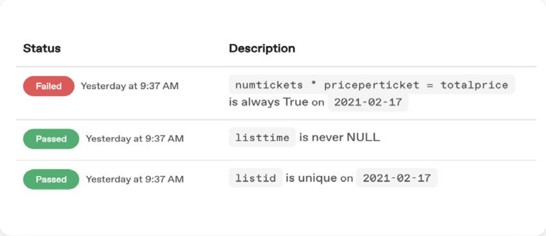
✅ Deep monitoring of data

⚠️ More work to scale

Anomalo Offers a Unique, AI-First Approach

Validation Rules

Experts specify hard and fast rules about the data



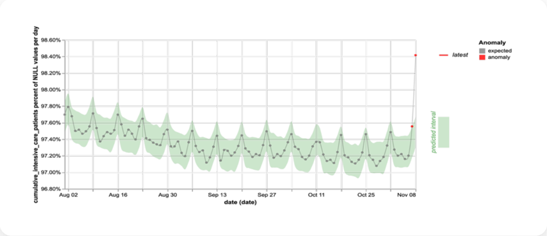
Status		Description
Failed	Yesterday at 9:37 AM	`numtickets * priceperticket = totalprice` is always True on `2021-02-17`
Passed	Yesterday at 9:37 AM	`listtime` is never NULL
Passed	Yesterday at 9:37 AM	`listid` is unique on `2021-02-17`

✅ Easy to understand

⚠️ Hard to maintain at scale

Metric Anomalies

Monitor changes in key business or data quality metrics

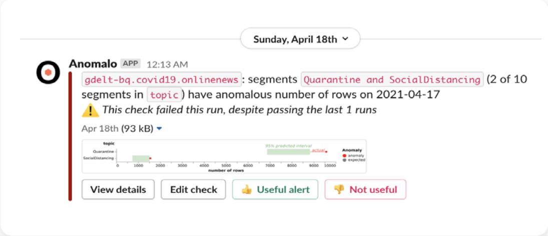


✅ Great for key KPIs

⚠️ Alert fatigue at scale

AI-Powered Monitoring

Automatically find significant changes in data contents



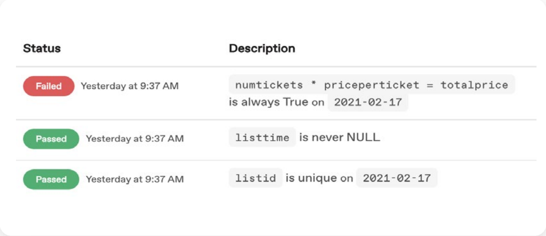
✅ Exhaustive validation with no setup

⚠️ Not as targeted as other methods

Anomalo Offers a Unique, AI-First Approach

Validation Rules

Experts specify hard and fast rules about the data

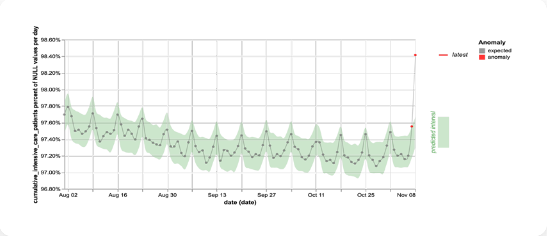


✅ Easy to understand

⚠️ Hard to maintain at scale

Metric Anomalies

Monitor changes in key business or data quality metrics

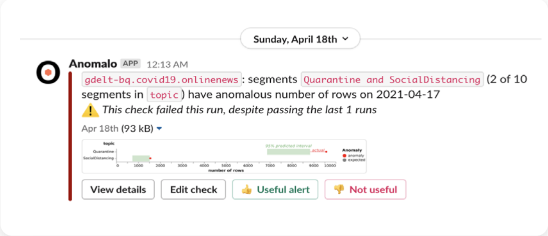


✅ Great for key KPIs

⚠️ Alert fatigue at scale

AI-Powered Monitoring

Automatically find significant changes in data contents



✅ Exhaustive validation with no setup

⚠️ Not as targeted as other methods

Natively Integrated With Databricks

Monitor any Databricks table across the entire data and AI lifecycle

DELTA LIVE TABLES

DELTA LAKE

DATABRICKS SQL

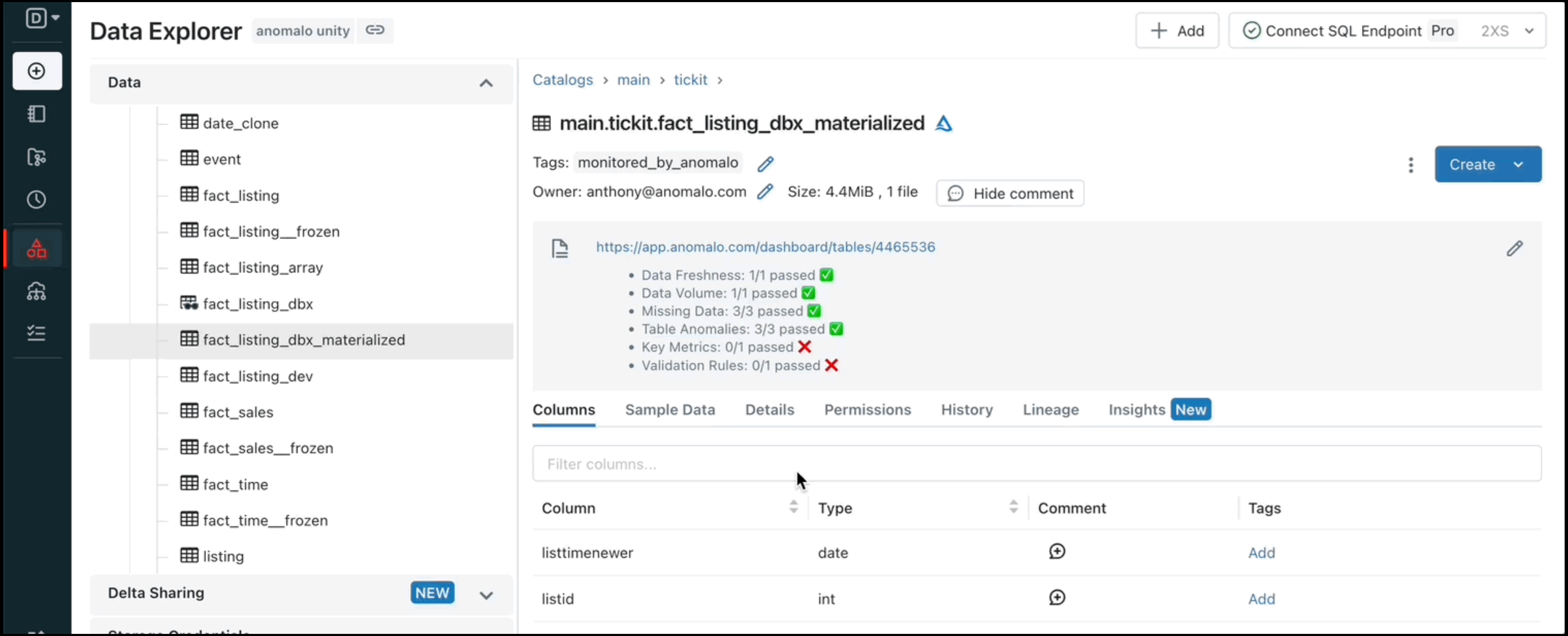
Extend the value of Unity Catalog with a native bidirectional integration

UNITY CATALOG

Access an exclusive free trial via Partner Connect

PARTNER CONNECT

Databricks > Anomalo Demo



Using Anomalo + Databricks to Promote Data Trust

Anomalo has been the silver bullet in helping us promote trust in data across our organization. Since migrating to Anomalo, it is easy to detect false positives and has removed dependencies on data engineering. This makes both my data engineers and data consumers happy as it means less time fire-fighting issues, and more time using data to build products our customers love.

—

TIM NG
DATA PRODUCTS ENG LEAD, BLOCK

BLOCK

Using Anomalo + Databricks to Save Millions

→ **Dozens of teams onboarded**

→ **Thousands of automated checks**

→ **Over 100,000 hours saved**

→ **More than $6MM dollars saved in < 1 year**

We have not only been able to replace manually created data quality checks with Anomalo's automated checks, but Anomalo's **unsupervised machine learning has also found data quality issues that are hard to predict—thus savings us millions of dollars in a short period of time.**

— DATA GOVERNANCE PROGRAM LEAD

Deeper Dive

Easy
How easy is it for an end user to configure and execute the evaluation?

Sensitive
How small of a change can it find?

Interpretable
When differences are detected, how much work is required to understand them?

Comprehensive
Can it cover all of the columns, segments, metrics, relationships, values?

Scalable
Can it scale to billions of rows and/or thousands of columns?

01

02

03

Validation Rules

Is the data my warehouse produces of high quality?

User input required: Table and column(s), Rule type, Constraint

Status		Description
Failed	Yesterday at 9:37 AM	`numtickets * priceperticket = totalprice` is always True on `2021-02-17`
Passed	Yesterday at 9:37 AM	`listtime` is never NULL
Passed	Yesterday at 9:37 AM	`listid` is unique on `2021-02-17`

✅ Easy to understand

⚠️ Hard to maintain at scale

02

03

Validation Rules

Is the data my warehouse produces of high quality?

User input required: Table and column(s), Rule type, Constraint

Status		Description
Failed	Yesterday at 9:37 AM	`numtickets * priceperticket = totalprice` is always True on `2021-02-17`
Passed	Yesterday at 9:37 AM	`listtime` is never NULL
Passed	Yesterday at 9:37 AM	`listid` is unique on `2021-02-17`

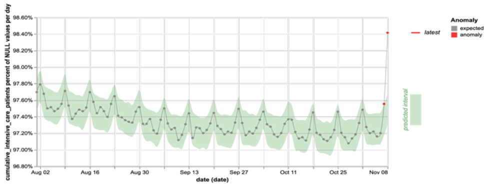
✅ Easy to understand

⚠️ Hard to maintain at scale

Metric Anomalies

Monitor changes in key business or data quality metrics

User input required: Table and column(s), Metric definition



✅ Great for key KPIs

⚠️ Alert fatigue at scale

03

	Validation Rules	Metric Anomalies
Easy	✗	✗
Interpretable	✓	✓
Scalable	✗	✗
Sensitive	✓	⚠
Comprehensive	✗	✗

Validation Rules

Is the data my warehouse produces of high quality?

User input required: Table and column(s), Rule type, Constraint

Status		Description
Failed	Yesterday at 9:37 AM	`numtickets * priceperticket = totalprice` is always True on `2021-02-17`
Passed	Yesterday at 9:37 AM	`listtime` is never NULL
Passed	Yesterday at 9:37 AM	`listid` is unique on `2021-02-17`

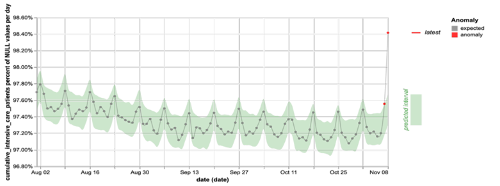
✅ Easy to understand

⚠️ Hard to maintain at scale

Metric Anomalies

Monitor changes in key business or data quality metrics

User input required: Table and column(s), Metric definition



✅ Great for key KPIs

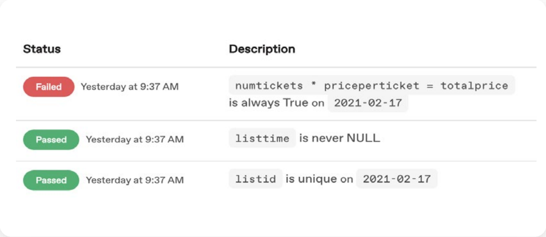
⚠️ Alert fatigue at scale

03

Validation Rules

Is the data my warehouse produces of high quality?

User input required: Table and column(s), Rule type, Constraint



Status	Description
Failed Yesterday at 9:37 AM	`numtickets * priceperticket = totalprice` is always True on `2021-02-17`
Passed Yesterday at 9:37 AM	`listtime` is never NULL
Passed Yesterday at 9:37 AM	`listid` is unique on `2021-02-17`

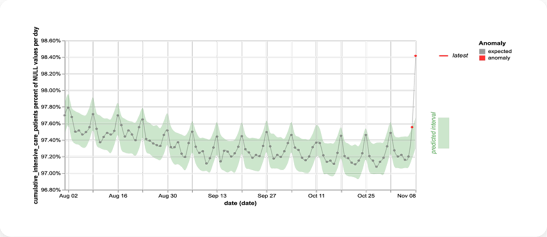
✅ Easy to understand

⚠️ Hard to maintain at scale

Metric Anomalies

Monitor changes in key business or data quality metrics

User input required: Table and column(s), Metric definition



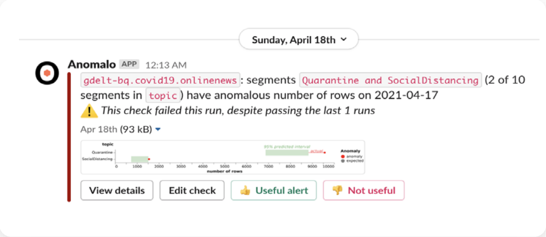
✅ Great for key KPIs

⚠️ Alert fatigue at scale

AI-Powered Monitoring

Automatically find significant changes inside the raw data

User input required: Table



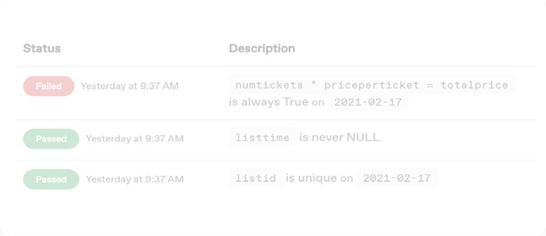
✅ Exhaustive validation with no setup

⚠️ Not as targeted as other methods

Validation Rules

Is the data my warehouse produces of high quality?

User input required: Table and column(s), Rule type, Constraint



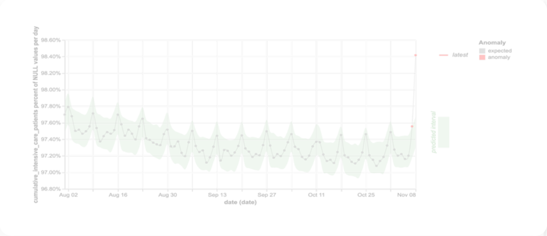
✅ Easy to understand

⚠️ Hard to maintain at scale

Metric Anomalies

Monitor changes in key business or data quality metrics

User input required: Table and column(s), Metric definition



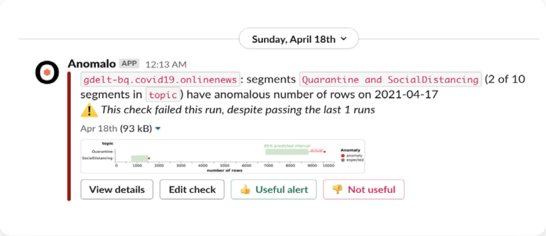
✅ Great for key KPIs

⚠️ Alert fatigue at scale

AI-Powered Monitoring

Automatically find significant changes inside the raw data

User input required: Table



✅ Exhaustive validation with no setup

⚠️ Not as targeted as other methods

Easily Scale Data Quality Monitoring

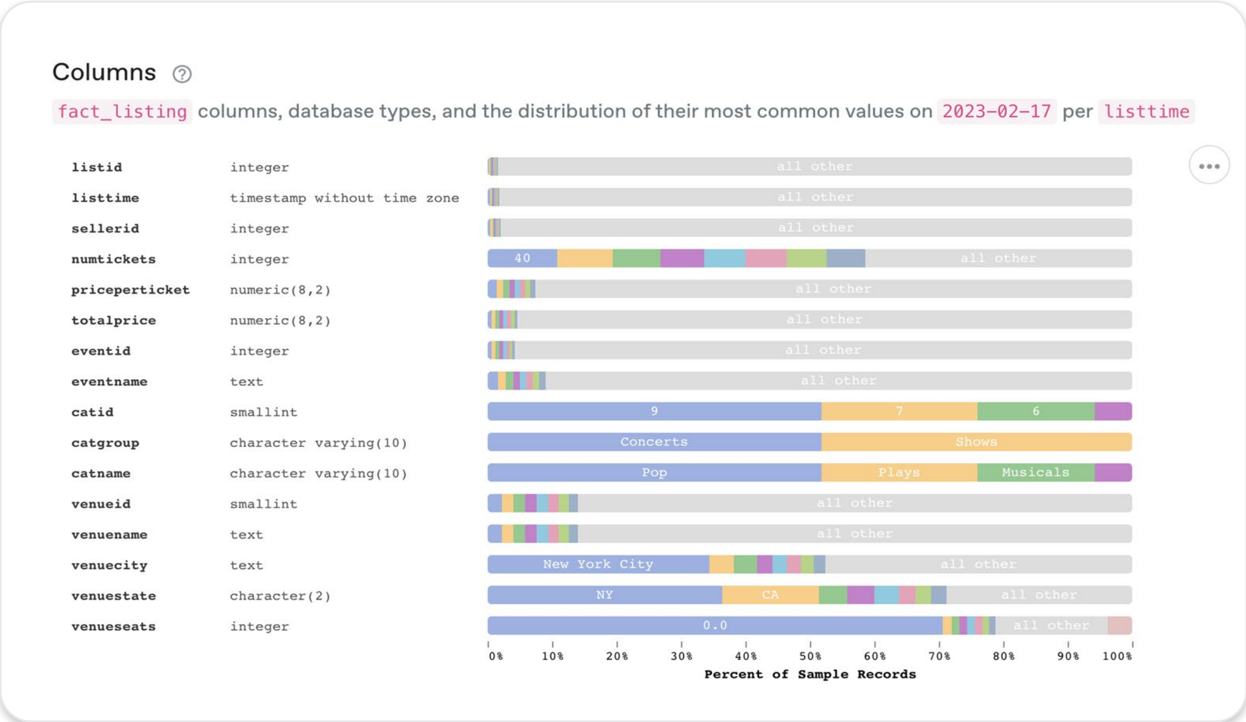
Ticket Sales Data

listid	listtime	numtickets	priceperticket	totalprice	eventname	catgroup	catname	venuename	venuecity	venuestate	venueseats
219247	2022-05-03 11:04:54	4	1077	4308	Hairspray	Shows	Musicals	Ambassador Theatre	New York City	NY	0
17151	2022-05-03 15:27:17	3	106	318	Cypress Hill	Concerts	Pop	Verizon Center	Washington	DC	0
157318	2022-05-03 9:16:35	9	370	3330	Seven Brides for Seve	Shows	Musicals	Cort Theatre	New York City	NY	0
75570	2022-05-03 7:05:29	4	298	1192	Peter Frampton	Concerts	Pop	Fox Theatre	Redwood City	CA	0
230609	2022-05-03 8:08:54	4	388	1552	South Pacific	Shows	Musicals	Paramount Theatre	Seattle	WA	0
217754	2022-05-03 18:19:40	6	1166	6996	A Man For All Seasons	Shows	Plays	Richard Rodgers Theatre	New York City	NY	0
7743	2022-05-03 19:36:02	24	34	816	Hall and Oates	Concerts	Pop	Lambeau Field	Green Bay	WI	72922
39587	2022-05-03 14:13:25	22	124	2728	Rock the Bells	Concerts	Pop	Oriole Park at Camden Y	Baltimore	MD	48876
61926	2022-05-03 12:11:27	12	368	4416	Gnarls Barkley	Concerts	Pop	Hubert H. Humphrey Metr	Minneapolis	MN	64035
43729	2022-05-03 10:11:43	4	131	524	The Who	Concerts	Pop	Reliant Stadium	Houston	TX	72000

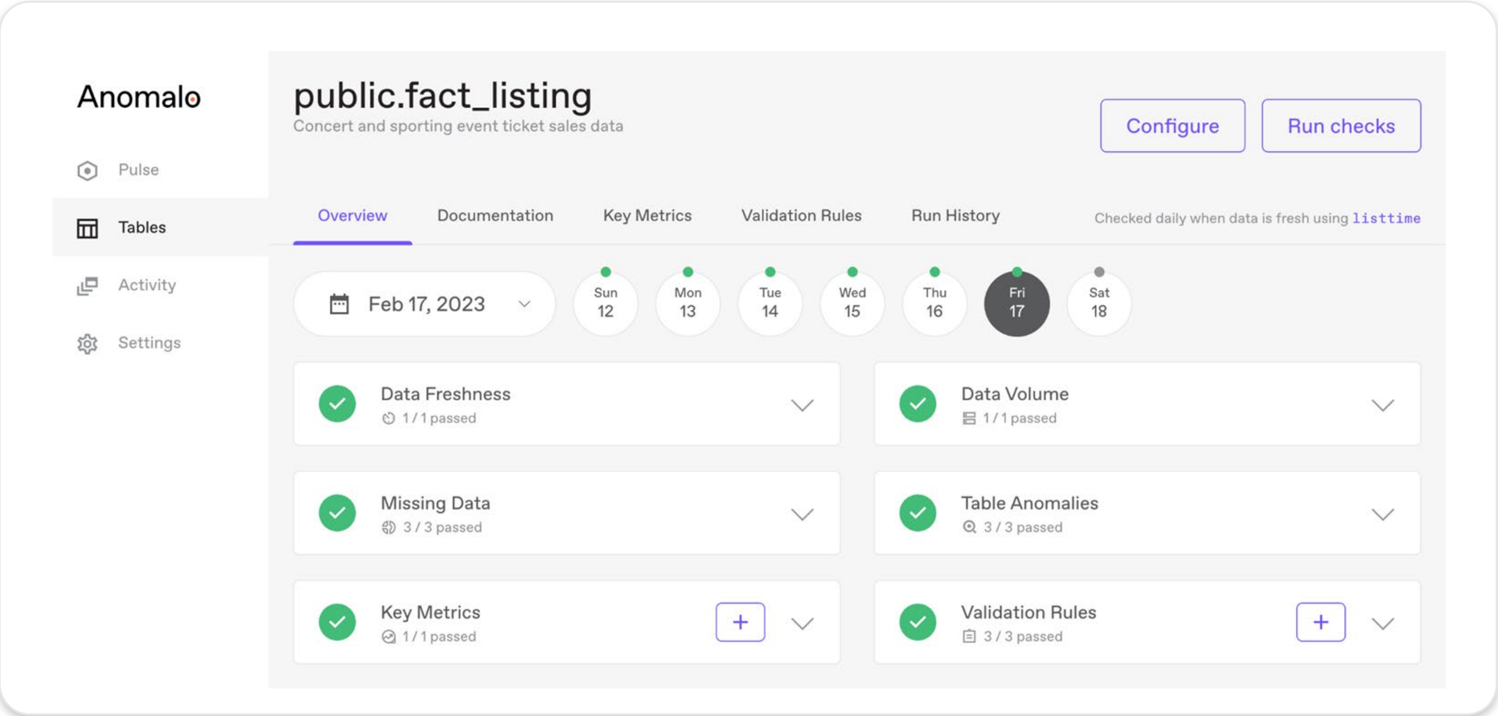
Ticket Sales Data

listid	listtime	numtickets	priceperticket	totalprice	eventname	catgroup	catname	venuename	venuecity	venuestate	venueseats
219247	2022-05-03 11:04:54	4	1077	4308	Hairspray	Shows	Musicals	Ambassador Theatre	New York City	NY	0
17151	2022-05-03 15:27:17	3	106	318	Cypress Hill	Concerts	Pop	Verizon Center	Washington	DC	0
157318	2022-05-03 9:16:35	9	370	3330	Seven Brides for Seve	Shows	Musicals	Cort Theatre	New York City	NY	0
43729	2022-05-03 10:11:43	4	131	524	The Who	Concerts	Pop	Reliant Stadium	Houston	TX	72000
230609	2022-05-03 8:08:54	4	388	1552	South Pacific	Shows	Musicals	Paramount Theatre	Seattle	WA	0
217754	2022-05-03 18:19:40	6	1166	6996	A Man For All Seasons	Shows	Plays	Richard Rodgers Theatre	New York City	NY	0
7743	2022-05-03 19:36:02	24	34	816	Hall and Oates	Concerts	Pop	Lambeau Field	Green Bay	WI	72922
39587	2022-05-03 14:13:25	22	124	2728	Rock the Bells	Concerts	Pop	Oriole Park at Camden Y	Baltimore	MD	48876
61926	2022-05-03 12:11:27	12	368	4416	Gnarls Barkley	Concerts	Pop	Hubert H. Humphrey Metr	Minneapolis	MN	64035
43729	2022-05-03 10:11:43	4	131	524	The Who	Concerts	Pop	Reliant Stadium	Houston	TX	72000

Data Profiling



Anomalo Monitoring



Randomly terminates virtual machine instances

NETFLIX

CHAOS MONKEY

Randomly terminates virtual machine instances

Anomalo

CHAOS LLAMA

Randomly introduces data quality issues

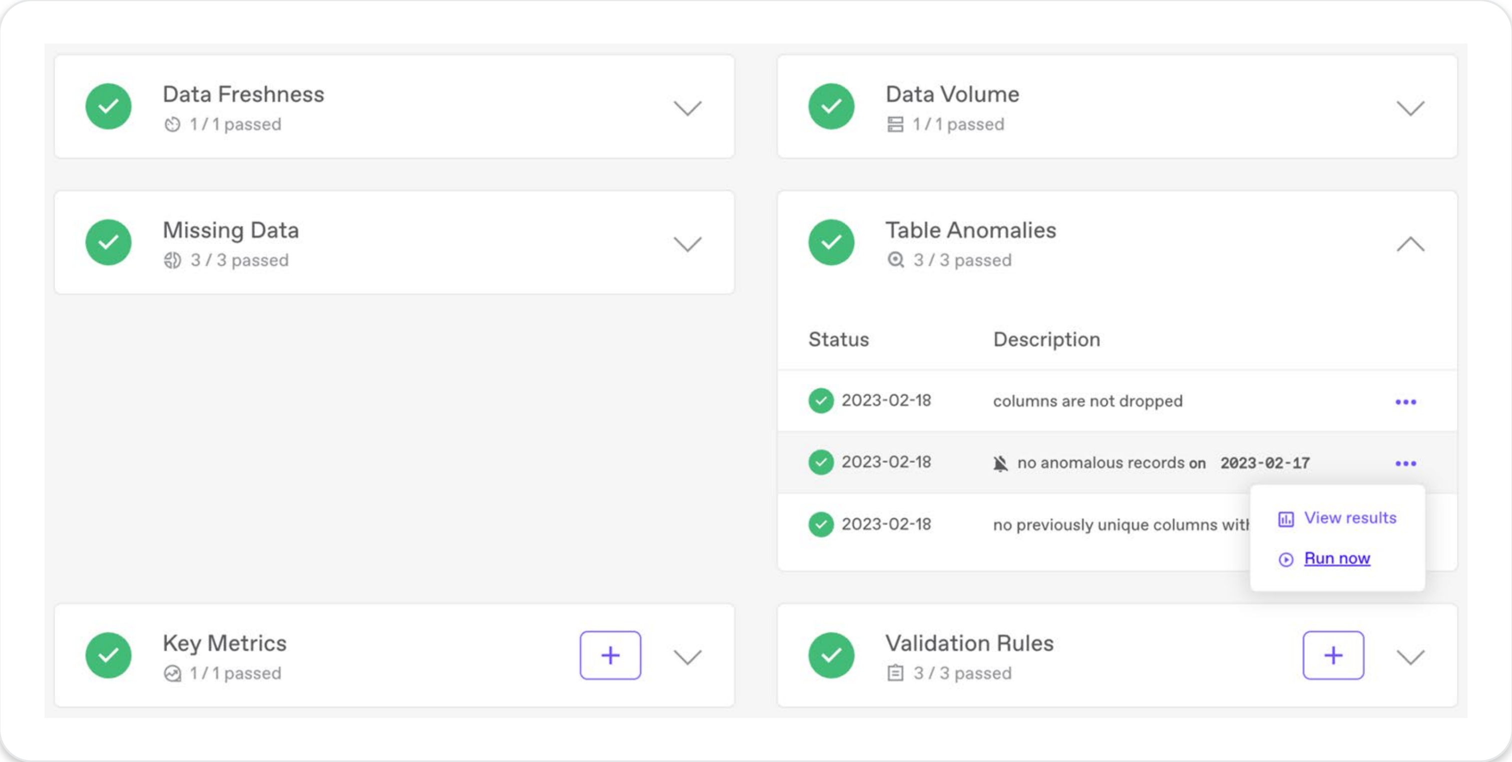
Chaos Library

```
Anomalo >>> chaos help

Available Chaos Commands:

ColumnDropValue†
Drops all rows from table.column with a given value.

ColumnGrow†
Multiplies a column by a random value drawn uniformly from [low, high].  Use grow_sumbol='+' to achieve additive gr
owth.

ColumnIdentity†
Does nothing to a Column (if a tree falls...).

ColumnInfrequentDrop†
Drops rows with values equal to an infrequent randomly chosen value, which must represent between low_threshold and
 high_threshold fraction of records for a given column.  If no such value exists, this check will throw an error.

ColumnModeDrop†
Drops rows with values equal to the mode of a given column.  Requires that the mode represents at least threshold f
raction of the data or else will throw an error. This is designed to prevent chaos where the mode is very rare.

...
```

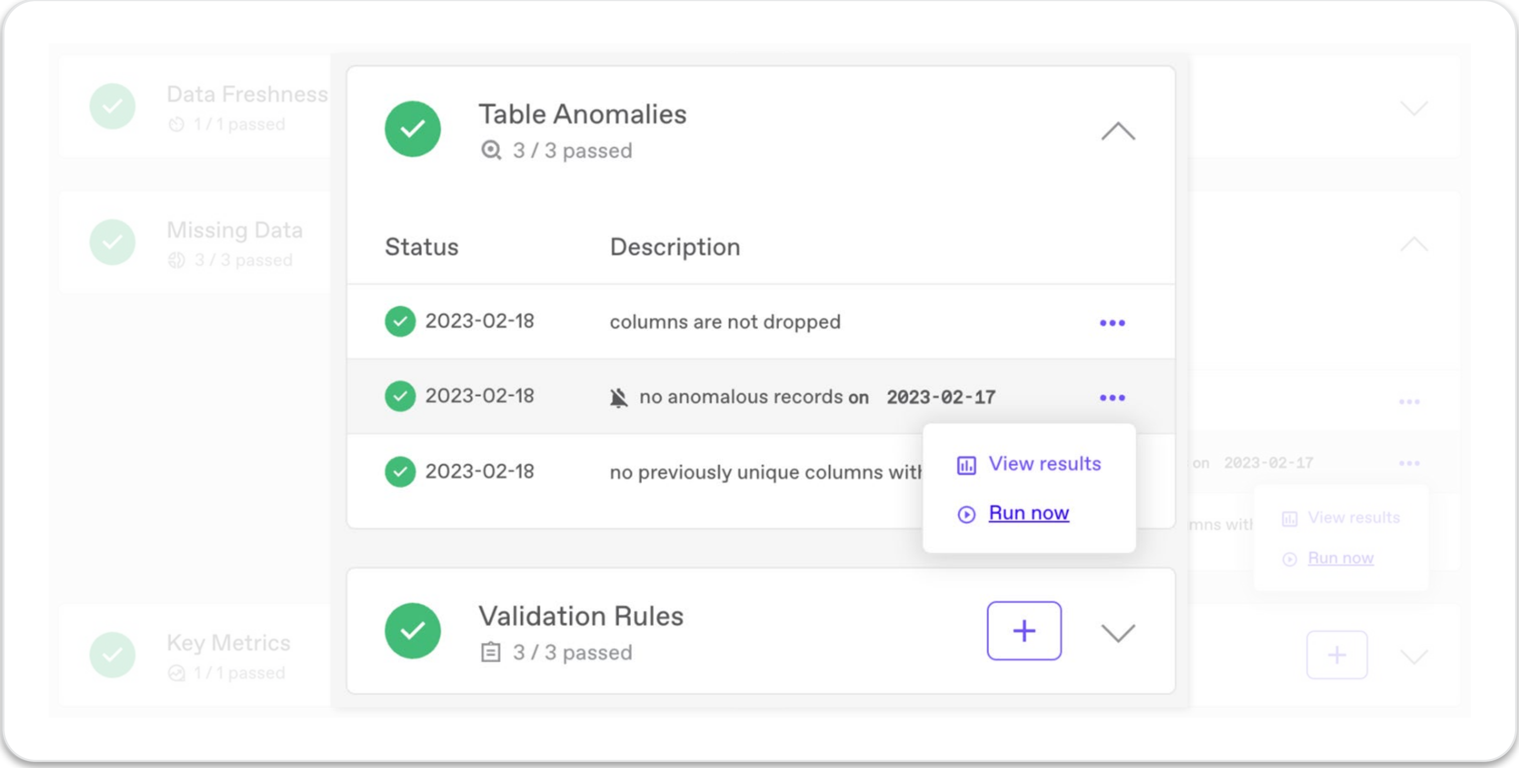
Introducing Chaos

```
Anomalo >>> chaos      TimeColumnValue
table:                                 fact_listing
column:                       priceperticket
value:                            10
frac:                             .3
where_sql:                    venuestate = ‘NY’
time_col:                     listtime
date:                             2023-02-17
```

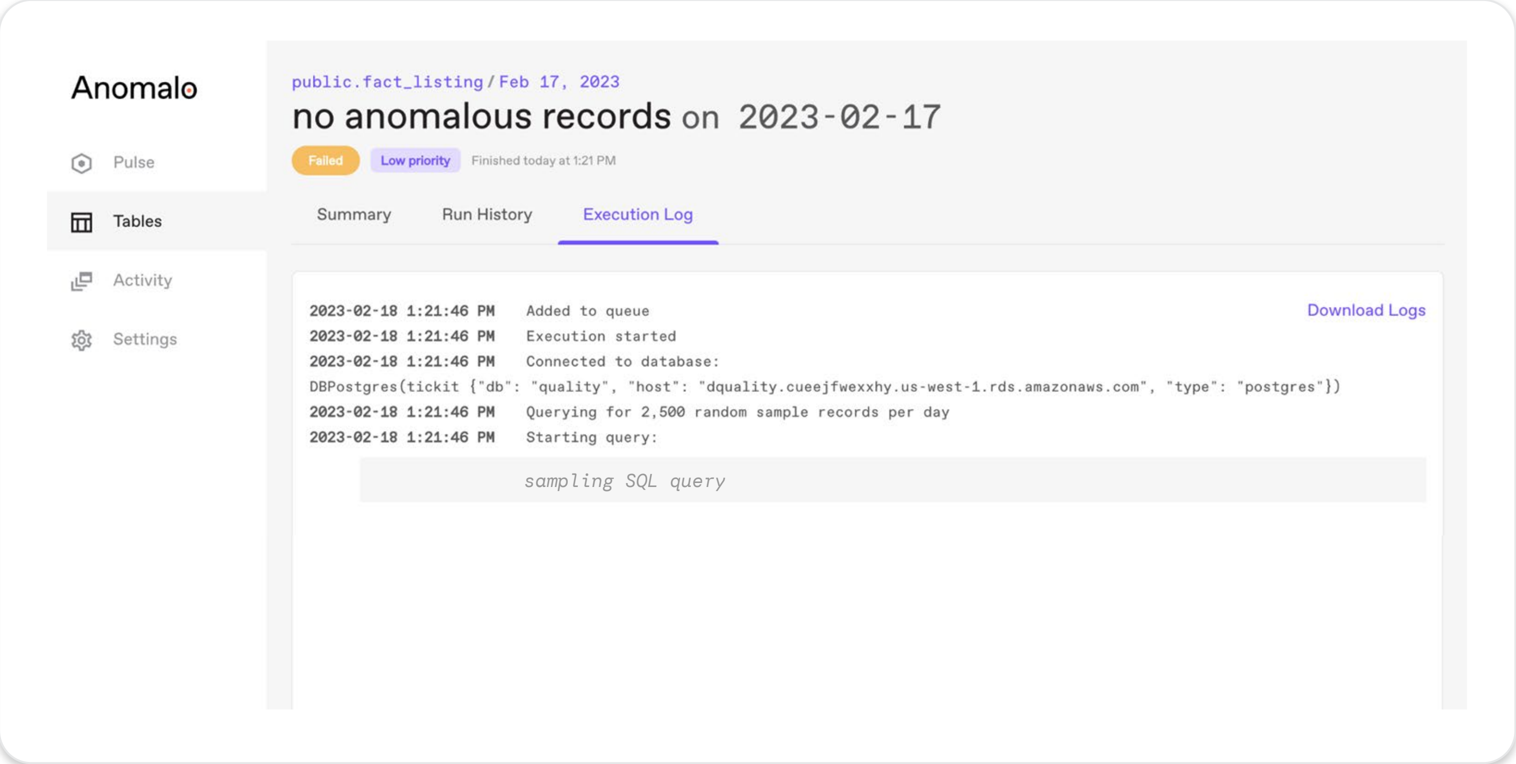
Re-Run The Check



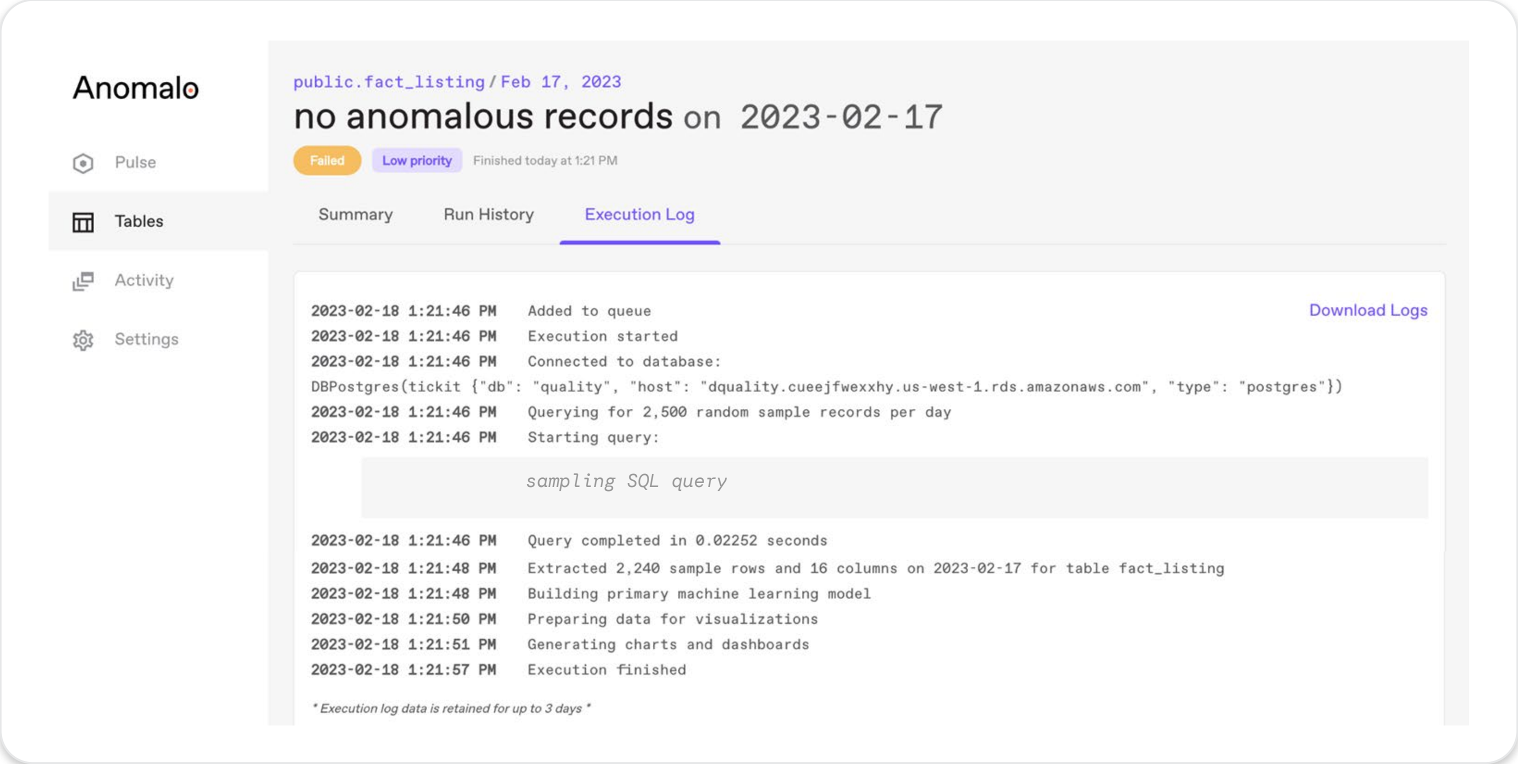
Re-Run The Check



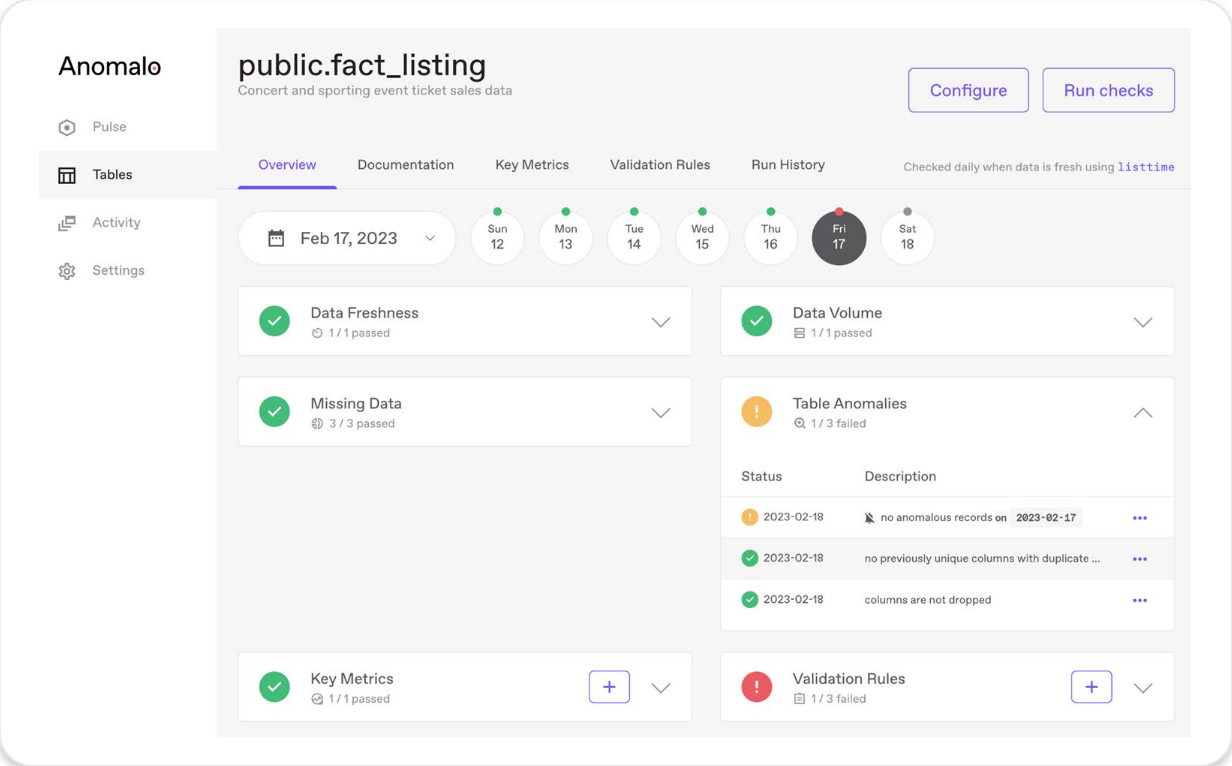
Watch The Log



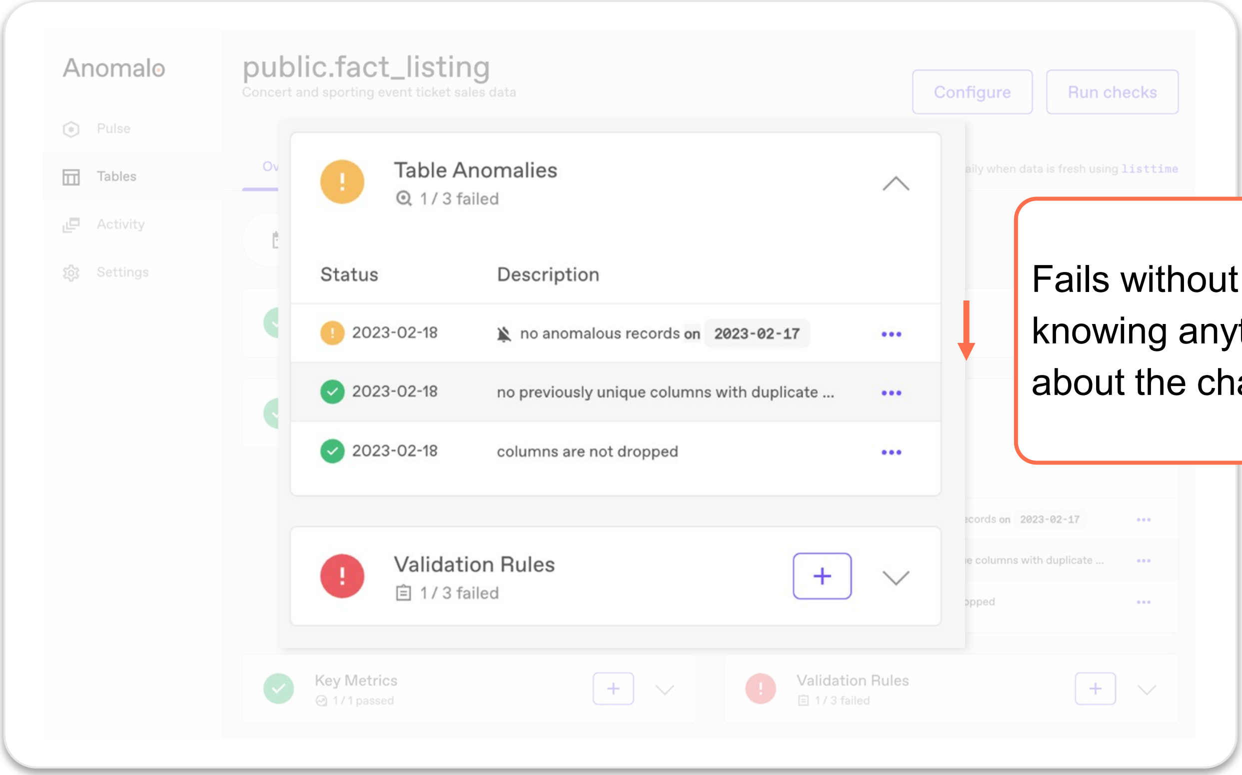
Watch The Log



Check Fails

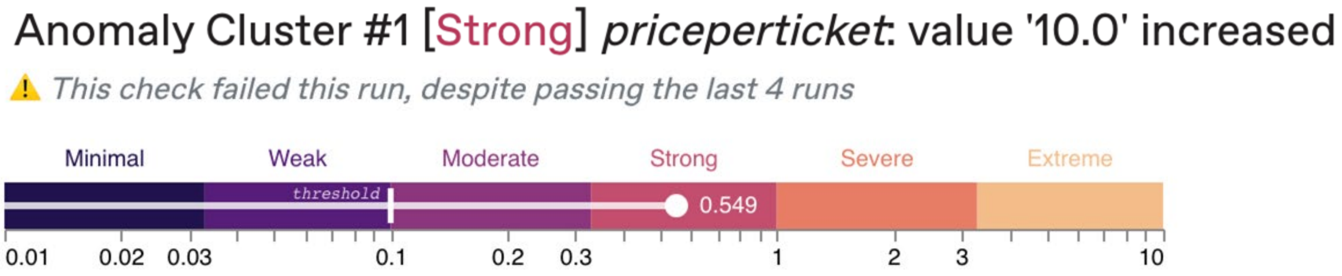

Anomalo
Pulse
Tables
Activity
Settings
public.fact_listing
Concert and sporting event ticket sales data
Configure
Run checks
Overview
Documentation
Key Metrics
Validation Rules
Run History
Checked daily when data is fresh using listtime
Feb 17, 2023
Sun 12
Mon 13
Tue 14
Wed 15
Thu 16
Fri 17
Sat 18
Data Freshness
1 / 1 passed
Data Volume
1 / 1 passed
Missing Data
3 / 3 passed
Table Anomalies
1 / 3 failed
Status
Description
2023-02-18
no anomalous records on 2023-02-17
2023-02-18
no previously unique columns with duplicate ...
2023-02-18
columns are not dropped
Key Metrics
1 / 1 passed
Validation Rules
1 / 3 failed

Check Fails

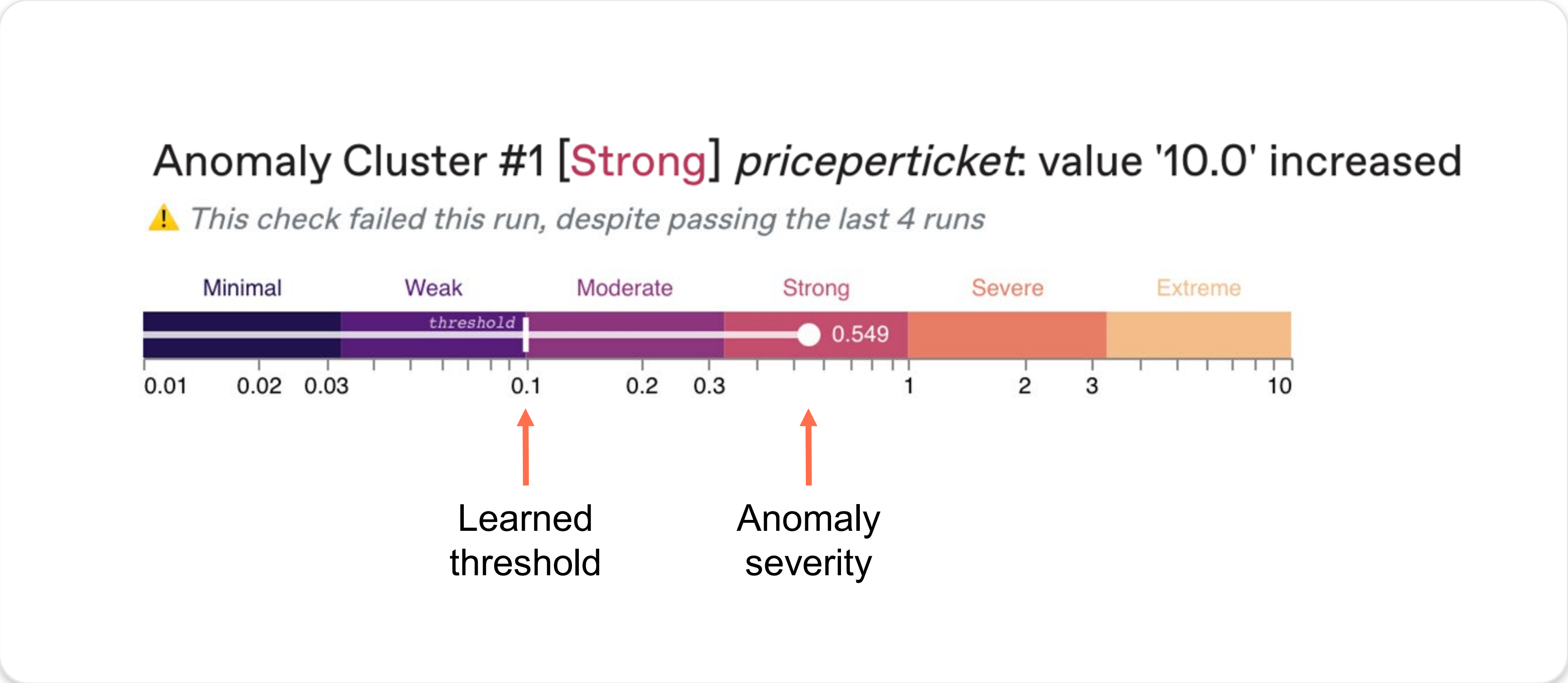


Fails without knowing anything about the change

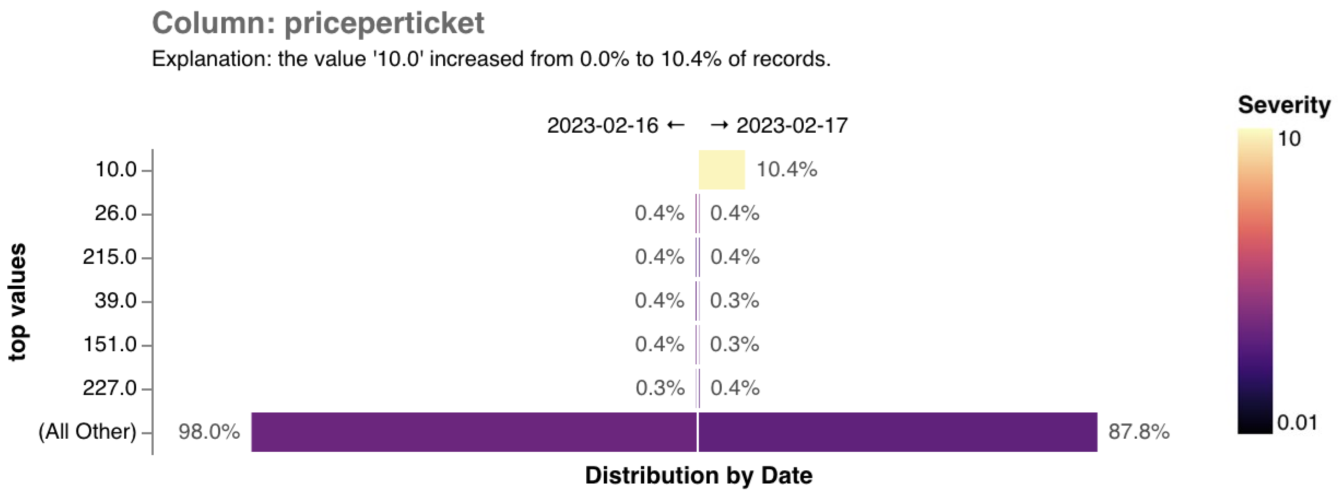
Severity and Explanation



Severity and Explanation



Distribution



Good and Bad Records

Sample of Most Anomalous Rows

A random sample of 100 rows with the highest anomaly scores for cluster: priceperticket

Column	Row Value
catgroup	Concerts
catid	9
catname	Pop
eventid	4857
eventname	The Guess Who
listid	41089
listtime	2023-02-17 11:39:14
numtickets	8
priceperticket	10.0
sellerid	35078
totalprice	2376.0
venuecity	New York City
venueid	108
venuename	Shea Stadium
venueseats	0.0
venuestate	NY

sample row number: 1

column sort order: database order

Sample of Least Anomalous Rows

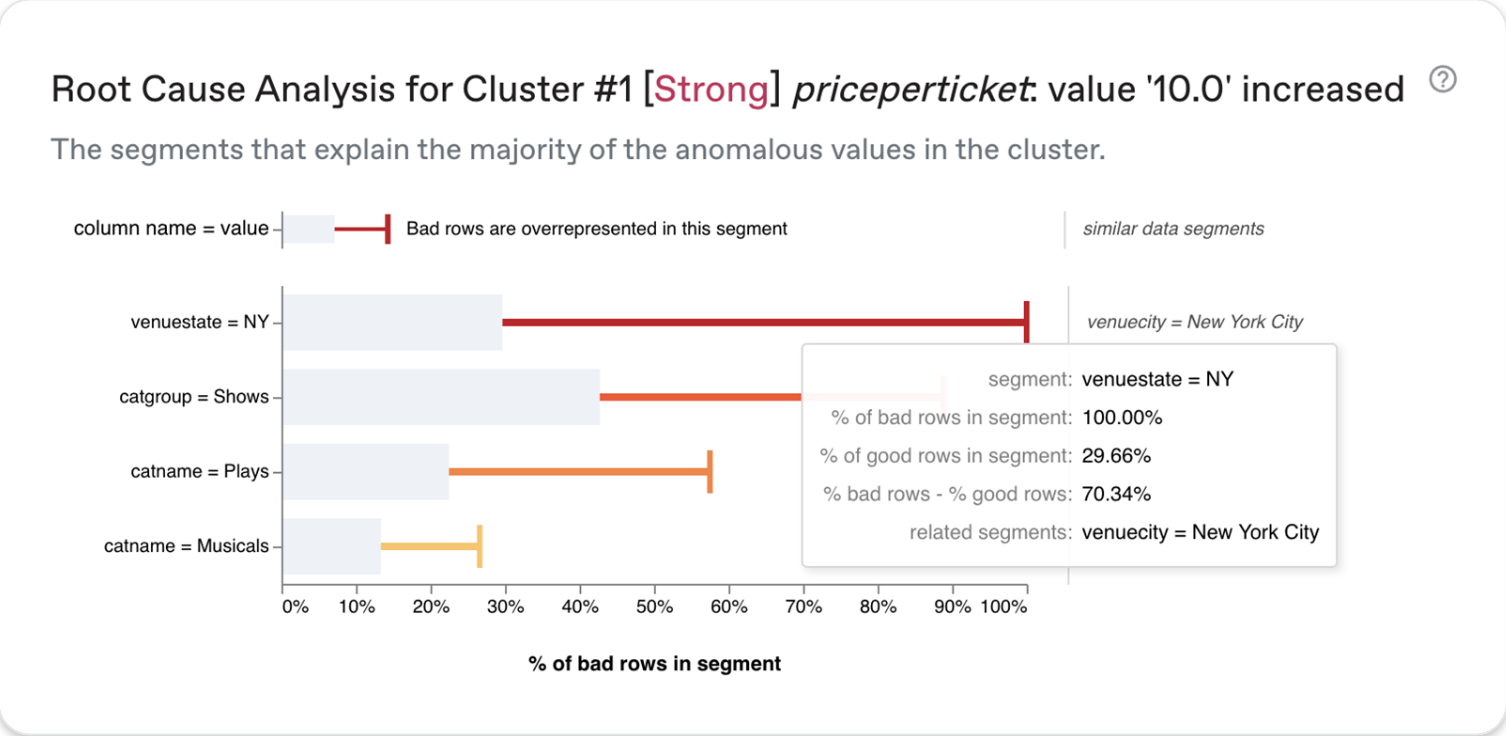
A random sample of 100 rows with the lowest anomaly scores for cluster: priceperticket

Column	Row Value
catgroup	Concerts
catid	9
catname	Pop
eventid	7066
eventname	Loretta Lynn
listid	5629
listtime	2023-02-17 01:35:54
numtickets	22
priceperticket	69.0
sellerid	18650
totalprice	1518.0
venuecity	New Orleans
venueid	43
venuename	New Orleans Arena
venueseats	0.0
venuestate	LA

sample row number: 1

column sort order: database order

Root Cause Analysis



DATA

TODAY

listid	listtime	sellerid	numtickets	...	venuecity	venuestate	venueseats
219247	2022-05-03 11:04:54	12619	4	...	New York City	NY	0
17151	2022-05-03 15:27:17	27222	3	...	Washington	DC	0
157318	2022-05-03 9:16:35	22899	9	...	New York City	NY	0
75570	2022-05-03 7:05:29	44301	4	...	Redwood City	CA	0
230609	2022-05-03 8:08:54	9663	4	...	Seattle	WA	0
217754	2022-05-03 18:19:40	4999	6	...	New York City	NY	0
7743	2022-05-03 19:36:02	4474	24	...	Green Bay	WI	72922
39587	2022-05-03 14:13:25	37990	22	...	Baltimore	MD	48876
...	...	...	...	...	...	...	...
43729	2022-05-03 10:11:43	39035	4	...	Houston	TX	72000

YESTERDAY

listid	listtime	sellerid	numtickets	...	venuecity	venuestate	venueseats
227155	2022-05-02 0:31:03	385	6	...	New York City	NY	0
41479	2022-05-02 14:19:31	1674	8	...	Mountain View	CA	22000
123523	2022-05-02 3:34:37	45271	8	...	Raleigh	NC	0
11438	2022-05-02 11:20:46	38064	1	...	Dayton	OH	0
118772	2022-05-02 9:51:39	41862	16	...	Kansas City	KS	0
143671	2022-05-02 0:13:16	4554	14	...	New York City	NY	0
101427	2022-05-02 17:26:38	3398	5	...	Philadelphia	PA	68532
228561	2022-05-02 3:40:54	4975	3	...	New York City	NY	0
...	...	...	...	...	...	...	...
158082	2022-05-02 20:25:23	47888	18	...	Milwaukee	WI	0

Take random samples from today and yesterday

DATA

TODAY

listid	listtime	sellerid	numtickets	...	venuecity	venuestate	venueseats
219247	2022-05-03 11:04:54	12619	4	...	New York City	NY	0
17151	2022-05-03 15:27:17	27222	3	...	Washington	DC	0
157318	2022-05-03 9:16:35	22899	9	...	New York City	NY	0
75570	2022-05-03 7:05:29	44301	4	...	Redwood City	CA	0
230609	2022-05-03 8:08:54	9663	4	...	Seattle	WA	0
217754	2022-05-03 18:19:40	4999	6	...	New York City	NY	0
7743	2022-05-03 19:36:02	4474	24	...	Green Bay	WI	72922
39587	2022-05-03 14:13:25	37990	22	...	Baltimore	MD	48876
...	...	...	...	...	...	...	...
43729	2022-05-03 10:11:43	39035	4	...	Houston	TX	72000

YESTERDAY

listid	listtime	sellerid	numtickets	...	venuecity	venuestate	venueseats
227155	2022-05-02 0:31:03	385	6	...	New York City	NY	0
41479	2022-05-02 14:19:31	1674	8	...	Mountain View	CA	22000
123523	2022-05-02 3:34:37	45271	8	...	Raleigh	NC	0
11438	2022-05-02 11:20:46	38064	1	...	Dayton	OH	0
118772	2022-05-02 9:51:39	41862	16	...	Kansas City	KS	0
143671	2022-05-02 0:13:16	4554	14	...	New York City	NY	0
101427	2022-05-02 17:26:38	3398	5	...	Philadelphia	PA	68532
228561	2022-05-02 3:40:54	4975	3	...	New York City	NY	0
...	...	...	...	...	...	...	...
158082	2022-05-02 20:25:23	47888	18	...	Milwaukee	WI	0

Take random samples from today and yesterday

Is there any material change between the two dates?

DATA

TODAY

listid	listtime	sellerid	numtickets	...	venuecity	venuestate	venueseats
219247	2022-05-03 11:04:54	12619	4	...	New York City	NY	0
17151	2022-05-03 15:27:17	27222	3	...	Washington	DC	0
157318	2022-05-03 9:16:35	22899	9	...	New York City	NY	0
75570	2022-05-03 7:05:29	44301	4	...	Redwood City	CA	0
230609	2022-05-03 8:08:54	9663	4	...	Seattle	WA	0
217754	2022-05-03 18:19:40	4999	6	...	New York City	NY	0
7743	2022-05-03 19:36:02	4474	24	...	Green Bay	WI	72922
39587	2022-05-03 14:13:25	37990	22	...	Baltimore	MD	48876
...	...	...	...	...	...	...	...
43729	2022-05-03 10:11:43	39035	4	...	Houston	TX	72000

YESTERDAY

listid	listtime	sellerid	numtickets	...	venuecity	venuestate	venueseats
227155	2022-05-02 0:31:03	385	6	...	New York City	NY	0
41479	2022-05-02 14:19:31	1674	8	...	Mountain View	CA	22000
123523	2022-05-02 3:34:37	45271	8	...	Raleigh	NC	0
11438	2022-05-02 11:20:46	38064	1	...	Dayton	OH	0
118772	2022-05-02 9:51:39	41862	16	...	Kansas City	KS	0
143671	2022-05-02 0:13:16	4554	14	...	New York City	NY	0
101427	2022-05-02 17:26:38	3398	5	...	Philadelphia	PA	68532
228561	2022-05-02 3:40:54	4975	3	...	New York City	NY	0
...	...	...	...	...	...	...	...
158082	2022-05-02 20:25:23	47888	18	...	Milwaukee	WI	0

Take random samples from today and yesterday

Is there any material change between the two dates?

Can we predict which day each record came from?

DATA

TODAY

listid	listtime	sellerid	numtickets	...	venuecity	venuestate	venueseats	response
219247	2022-05-03 11:04:54	12619	4	...	New York City	NY	0	1
17151	2022-05-03 15:27:17	27222	3	...	Washington	DC	0	1
157318	2022-05-03 9:16:35	22899	9	...	New York City	NY	0	1
75570	2022-05-03 7:05:29	44301	4	...	Redwood City	CA	0	1
230609	2022-05-03 8:08:54	9663	4	...	Seattle	WA	0	1
217754	2022-05-03 18:19:40	4999	6	...	New York City	NY	0	1
7743	2022-05-03 19:36:02	4474	24	...	Green Bay	WI	72922	1
39587	2022-05-03 14:13:25	37990	22	...	Baltimore	MD	48876	1
...	...	...	...	...	...	...	...	...
43729	2022-05-03 10:11:43	39035	4	...	Houston	TX	72000	1

YESTERDAY

listid	listtime	sellerid	numtickets	...	venuecity	venuestate	venueseats	response
227155	2022-05-02 0:31:03	385	6	...	New York City	NY	0	0
41479	2022-05-02 14:19:31	1674	8	...	Mountain View	CA	22000	0
123523	2022-05-02 3:34:37	45271	8	...	Raleigh	NC	0	0
11438	2022-05-02 11:20:46	38064	1	...	Dayton	OH	0	0
118772	2022-05-02 9:51:39	41862	16	...	Kansas City	KS	0	0
143671	2022-05-02 0:13:16	4554	14	...	New York City	NY	0	0
101427	2022-05-02 17:26:38	3398	5	...	Philadelphia	PA	68532	0
228561	2022-05-02 3:40:54	4975	3	...	New York City	NY	0	0
...	...	...	...	...	...	...	...	...
158082	2022-05-02 20:25:23	47888	18	...	Milwaukee	WI	0	0

Encode Response as 1 = Today, 0 = Yesterday

Encode features automatically

DATA

FEATURES

TODAY

listid	listtime	sellerid	numtickets	...	venuecity	venuestate	venueseats	response
219247	2022-05-03 11:04:54	12619	4	...	New York City	NY	0	1
17151	2022-05-03 15:27:17	27222	3	...	Washington	DC	0	1
157318	2022-05-03 9:16:35	22899	9	...	New York City	NY	0	1
75570	2022-05-03 7:05:29	44301	4	...	Redwood City	CA	0	1
230609	2022-05-03 8:08:54	9663	4	...	Seattle	WA	0	1
217754	2022-05-03 18:19:40	4999	6	...	New York City	NY	0	1
7743	2022-05-03 19:36:02	4474	24	...	Green Bay	WI	72922	1
39587	2022-05-03 14:13:25	37990	22	...	Baltimore	MD	48876	1
...	...	...	...	...	...	...	...	...
43729	2022-05-03 10:11:43	39035	4	...	Houston	TX	72000	1

X_1	X_2	X_3	...	X_n	response
0	0.07	17	...	0.28	1
0	0.08	1	...	0.27	1
0	0.61	2	...	0.29	1
1	0.37	1	...	0.27	1
1	0.44	9	...	0.45	1
1	0.01	8	...	0.88	1
1	0.16	1	...	0.10	1
0	0.21	11	...	0.58	1
...	...	...	...	...	...
1	0.73	3	...	1.00	1

YESTERDAY

listid	listtime	sellerid	numtickets	...	venuecity	venuestate	venueseats	response
227155	2022-05-02 0:31:03	385	6	...	New York City	NY	0	0
41479	2022-05-02 14:19:31	1674	8	...	Mountain View	CA	22000	0
123523	2022-05-02 3:34:37	45271	8	...	Raleigh	NC	0	0
11438	2022-05-02 11:20:46	38064	1	...	Dayton	OH	0	0
118772	2022-05-02 9:51:39	41862	16	...	Kansas City	KS	0	0
143671	2022-05-02 0:13:16	4554	14	...	New York City	NY	0	0
101427	2022-05-02 17:26:38	3398	5	...	Philadelphia	PA	68532	0
228561	2022-05-02 3:40:54	4975	3	...	New York City	NY	0	0
...	...	...	...	...	...	...	...	...
158082	2022-05-02 20:25:23	47888	18	...	Milwaukee	WI	0	0

X_1	X_2	X_3	...	X_n	response
0	0.85	17	...	0.43	0
0	0.78	2	...	0.49	0
1	0.79	2	...	0.84	0
0	0.46	6	...	0.93	0
0	0.91	3	...	0.37	0
1	0.31	1	...	0.51	0
1	0.52	13	...	0.61	0
1	0.51	3	...	0.35	0
...	...	...	...	...	...
1	0.60	7	...	0.86	0

Build a supervised learning model

DATA

TODAY

listid	listtime	sellerid	numtickets	...	venuecity	venuestate	venueseats	response
219247	2022-05-03 11:04:54	12619	4	...	New York City	NY	0	1
17151	2022-05-03 15:27:17	27222	3	...	Washington	DC	0	1
157318	2022-05-03 9:16:35	22899	9	...	New York City	NY	0	1
75570	2022-05-03 7:05:29	44301	4	...	Redwood City	CA	0	1
230609	2022-05-03 8:08:54	9663	4	...	Seattle	WA	0	1
217754	2022-05-03 18:19:40	4999	6	...	New York City	NY	0	1
7743	2022-05-03 19:36:02	4474	24	...	Green Bay	WI	72922	1
39587	2022-05-03 14:13:25	37990	22	...	Baltimore	MD	48876	1
...	...	...	...	...	...	...	...	...
43729	2022-05-03 10:11:43	39035	4	...	Houston	TX	72000	1

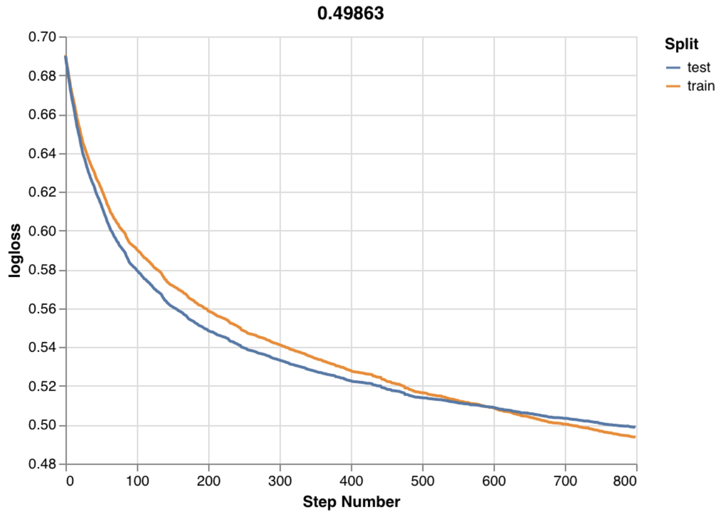
YESTERDAY

listid	listtime	sellerid	numtickets	...	venuecity	venuestate	venueseats	response
227155	2022-05-02 0:31:03	385	6	...	New York City	NY	0	0
41479	2022-05-02 14:19:31	1674	8	...	Mountain View	CA	22000	0
123523	2022-05-02 3:34:37	45271	8	...	Raleigh	NC	0	0
11438	2022-05-02 11:20:46	38064	1	...	Dayton	OH	0	0
118772	2022-05-02 9:51:39	41862	16	...	Kansas City	KS	0	0
143671	2022-05-02 0:13:16	4554	14	...	New York City	NY	0	0
101427	2022-05-02 17:26:38	3398	5	...	Philadelphia	PA	68532	0
228561	2022-05-02 3:40:54	4975	3	...	New York City	NY	0	0
...	...	...	...	...	...	...	...	...
158082	2022-05-02 20:25:23	47888	18	...	Milwaukee	WI	0	0

GRADIENT BOOSTING DECISION TREE

Model Train Test Split

Model performance over train and test split by iteration number

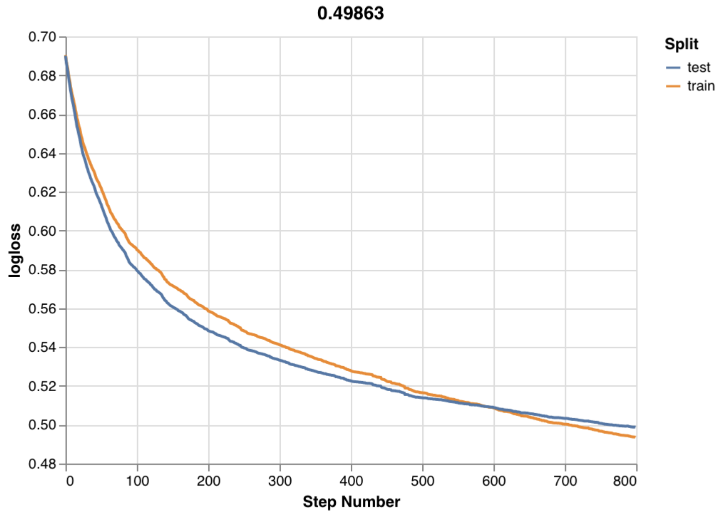


Use SHAP to credit changes to table values

GRADIENT BOOSTING DECISION TREE

Model Train Test Split

Model performance over train and test split by iteration number



SHAP VALUES (ANOMALY SCORES)

Most Anomalous Rows

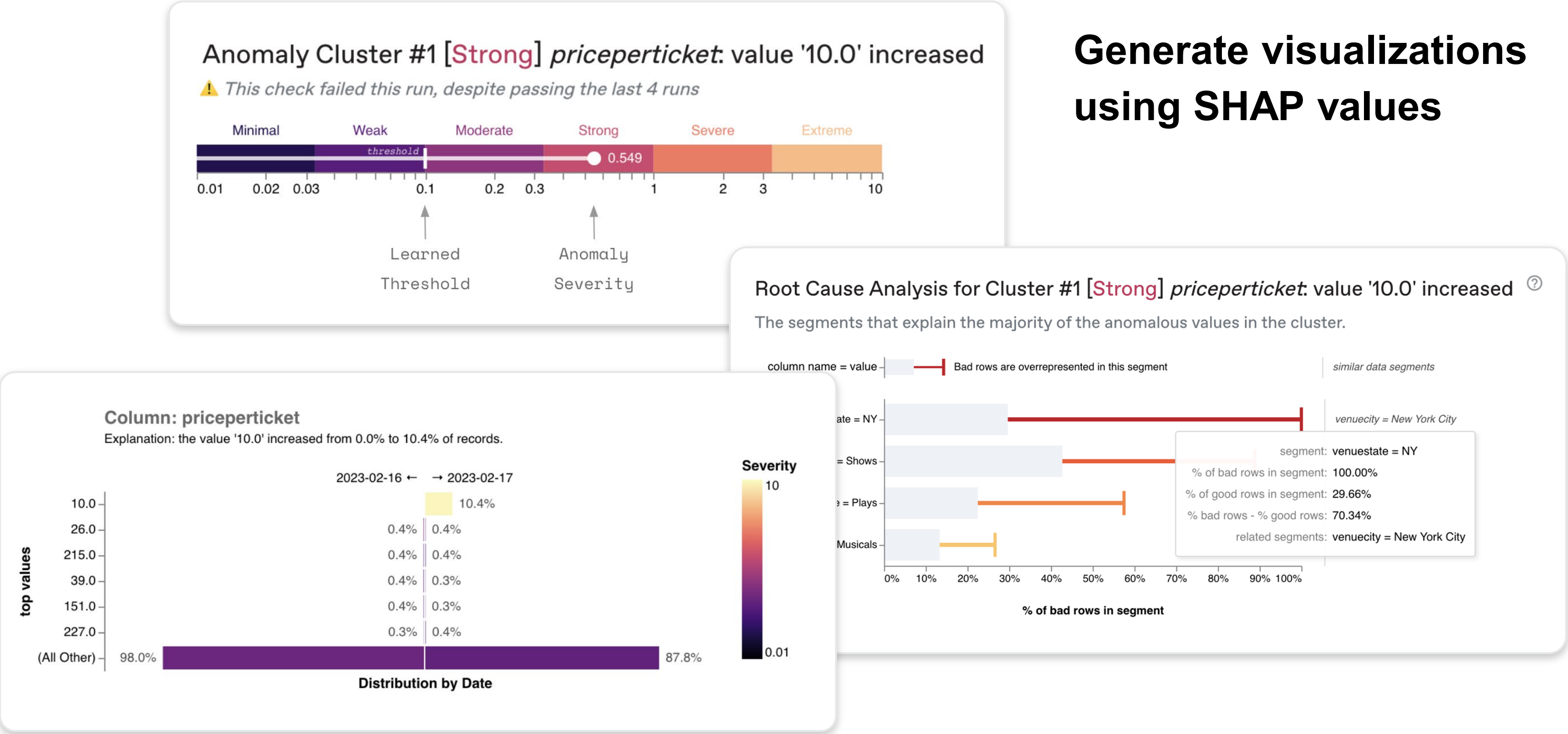
Shows the most anomalous rows of data, colored by anomaly scores and sorted by the maximum anomaly score in the row.

column

	listid	listtime	sellerid	numtickets	priceperticket	totalprice	eventid	eventname	catid	catgroup	catname	venueid	venuename	venuecity	venuestate	venueseats
1	1060	2021-01-17...	44798	28	1960.0	1568.0	8571	Vampire Weekend	9	Concerts	Pop	84	Soldier Field	Chicago	IL	63000.0
2	31413	2021-01-17...	36758	1	1337.0	101.0	22	La Rondine	8	Shows	Opera	306	Lyric Opera House	Baltimore	MD	0.0
3	11686	2021-01-17...	3739	3	2487.0	306.0	6427	America	9	Concerts	Pop	8	The Home Depot...	Carson	CA	0.0
4	7207	2021-01-17...	13517	16	1687.0	320.0	6210	Mos Def	9	Concerts	Pop	15	McAfee Coliseum	Oakland	CA	63026.0
5	13558	2021-01-17...	22420	16	2099.0	416.0	5341	Rock To Win At...	9	Concerts	Pop	46	Nassau Veterans...	Uniondale	NY	0.0
6	11693	2021-01-17...	25615	28	2284.0	6916.0	862	A Catered Affair	6	Shows	Musicals	228	Eugene O'Neill...	New York City	NY	0.0
7	24081	2021-01-17...	8817	4	815.0	164.0	362	Tristan und Isolde	8	Shows	Opera	300	Kennedy Center...	Washington	DC	0.0
8	14725	2021-01-17...	27464	16	1656.0	1488.0	8086	Joe Satriani	9	Concerts	Pop	31	Pepsi Center	Denver	CO	0.0
9	40263	2021-01-17...	27486	22	547.0	6138.0	3464	For Colored...	7	Shows	Plays	228	Eugene O'Neill...	New York City	NY	0.0
10	222909	2021-01-17...	11687	7	69.0	17409.0	1122	Flower Drum Song	6	Shows	Musicals	231	Gerald...	New York City	NY	0.0
11	224164	2021-01-17...	21252	6	29.0	5478.0	3273	A Streetcar...	7	Shows	Plays	244	Royce Hall	Los Angeles	CA	0.0
12	224347	2021-01-17...	13152	7	244.0	12992.0	309	The Queen of...	8	Shows	Opera	301	Ellie Caulkins...	Denver	CO	0.0
13	16441	2021-01-17...	25047	2	1794.0	106.0	3379	Othello	7	Shows	Plays	230	Richard Rodgers...	New York City	NY	0.0
14	6132	2021-01-17...	10704	5	1854.0	200.0	5324	Etta James	9	Concerts	Pop	60	Rexall Place	Edmonton	AB	0.0
15	42133	2021-01-17...	2806	1	1247.0	216.0	919	The King and I	6	Shows	Musicals	222	Majestic Theatre	New York City	NY	0.0
16	234372	2021-01-17...	18759	4	101.0	4988.0	8693	.38 Special	9	Concerts	Pop	122	Saratoga...	Saratoga Springs	NY	0.0
17	233658	2021-01-17...	25374	7	55.0	9044.0	3954	Echo & the...	9	Concerts	Pop	24	Conseco Fieldhouse	Indianapolis	IN	0.0
18	123399	2021-01-17...	30914	9	1278.0	432.0	1364	A Catered Affair	6	Shows	Musicals	245	The Guthrie...	Minneapolis	MN	0.0
19	12686	2021-01-17...	20138	8	1822.0	456.0	7240	Bette Midler	9	Concerts	Pop	61	Xcel Energy Center	St. Paul	MN	0.0
20	212967	2021-01-17...	41156	7	115.0	13531.0	7044	Dolly Parton	9	Concerts	Pop	87	Hubert H....	Minneapolis	MN	64035.0
21	8368	2021-01-17...	1681	18	2488.0	3330.0	7456	Idina Menzel	9	Concerts	Pop	9	Dick's Sporting...	Commerce City	CO	0.0
22	1712	2021-01-17...	21627	3	1158.0	207.0	4156	Kansas	9	Concerts	Pop	105	Safeco Field	Seattle	WA	47116.0
23	40290	2021-01-17...	20313	7	1867.0	301.0	1339	South Pacific	6	Shows	Musicals	227	New Amsterdam...	New York City	NY	0.0
24	48650	2021-01-17...	7669	3	1191.0	267.0	5987	Rock The Bayou	9	Concerts	Pop	29	Amway Arena	Orlando	FL	0.0
25	178489	2021-01-17...	21218	24	830.0	576.0	3464	For Colored...	7	Shows	Plays	228	Eugene O'Neill...	New York City	NY	0.0

severity
1.5
1.0
0.5

Generate visualizations using SHAP values



Seasonality
This change happens every Monday

Clustering Across Columns
Groups of columns have the same change

Time-Correlated Features
ID or date columns always increase

Performance
Must scale to billions of rows, thousands of columns

Chaotic Tables
Some tables change 100x more often

Accuracy
Needs to be sensitive to real changes, but suppress false positives

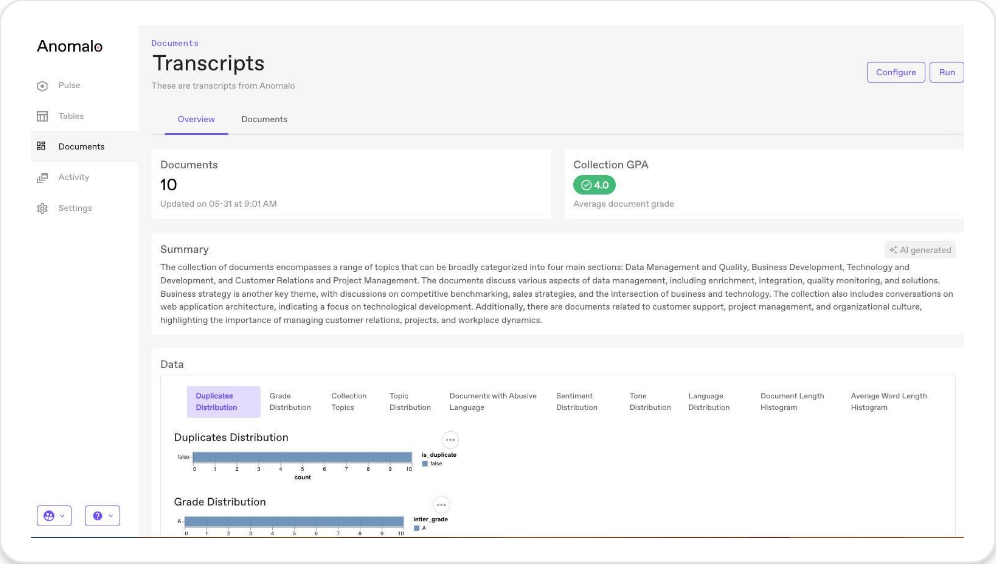
	Validation Rules	Metric Anomalies
Easy	❌	❌
Interpretable	✅	✅
Scalable	❌	❌
Sensitive	✅	⚠️
Comprehensive	❌	❌

	Validation Rules	Metric Anomalies	AI-Powered Monitoring
Easy	❌	❌	✅
Interpretable	✅	✅	✅
Scalable	❌	❌	✅
Sensitive	✅	⚠️	⚠️
Comprehensive	❌	❌	✅

Private Beta

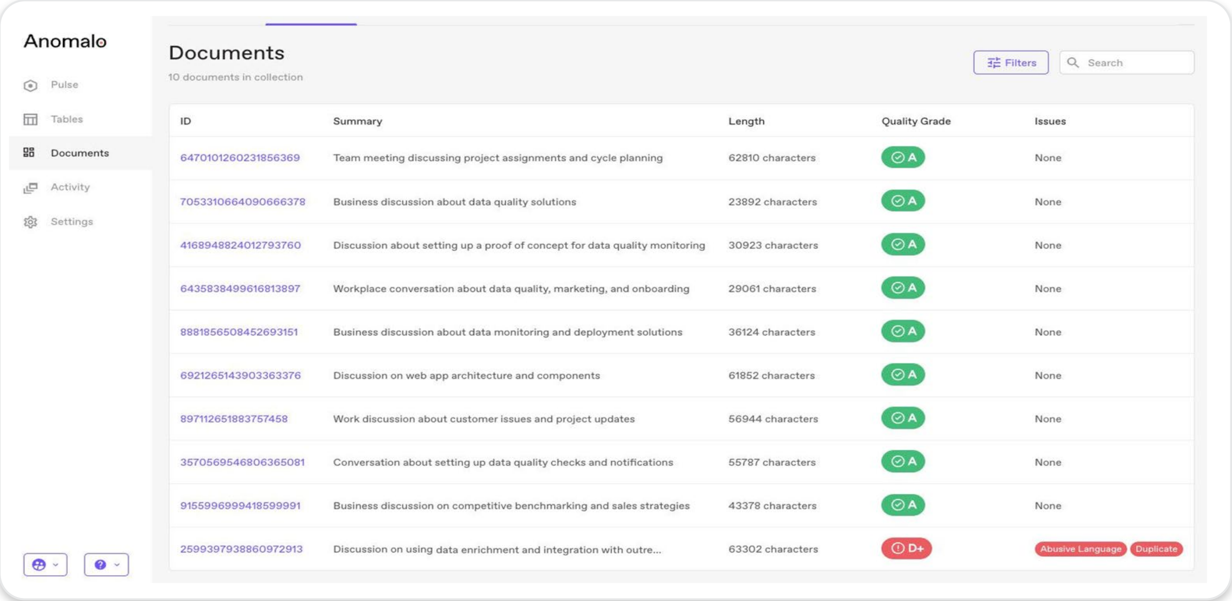
Anomalo’s Unstructured Monitoring Product

- 90% of Enterprise Data is Unstructrued
- **Now:** Unstructured text documents can be curated and evaluated with Anomalo
- Bring high-quality data to GenAI models
- Scale despite complexity, volume, and velocity



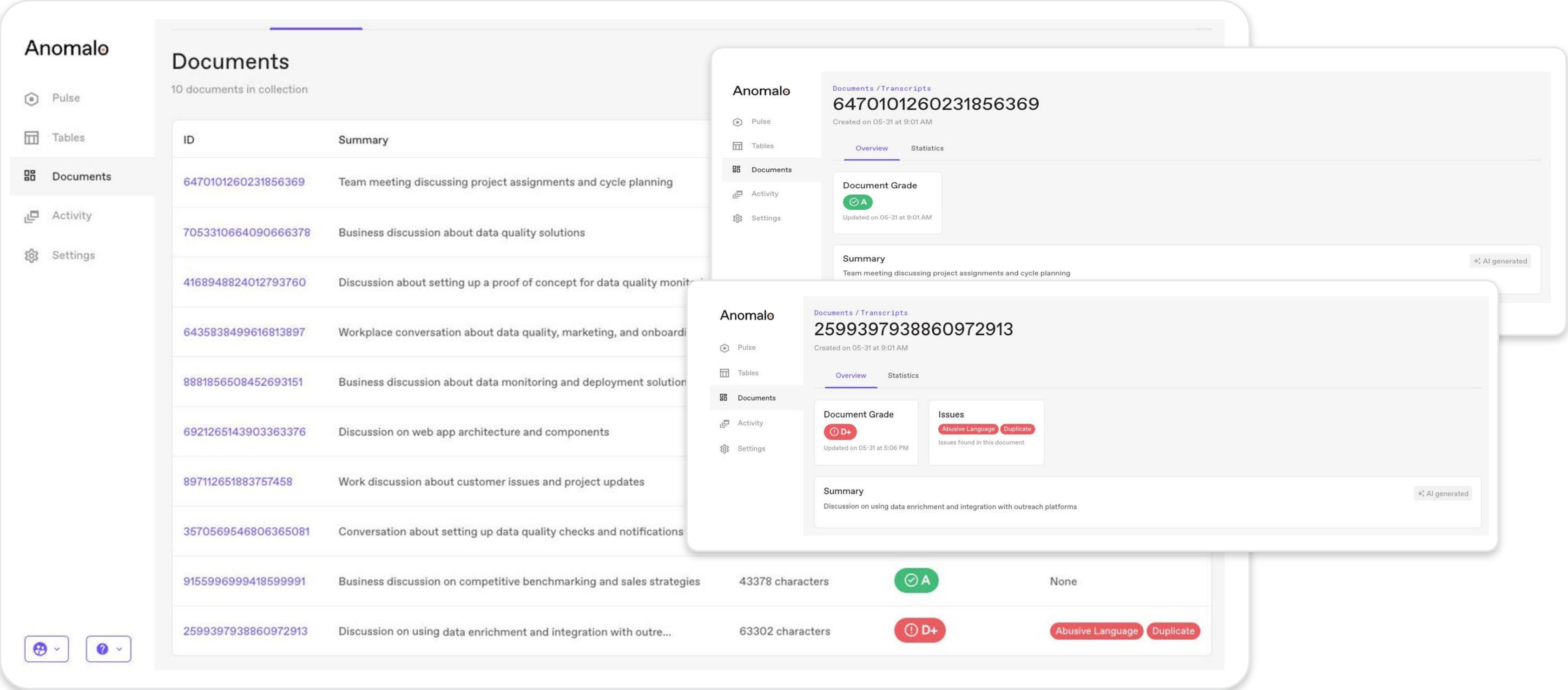
Apply for Private Beta at
anomalo.com/privatebeta

Monitor for length, duplicates, topics, tone, language, abusive language, PII and sentiment.



ID	Summary	Length	Quality Grade	Issues
6470101260231856369	Team meeting discussing project assignments and cycle planning	62810 characters	A	None
7053310664090666378	Business discussion about data quality solutions	23892 characters	A	None
4168948824012793760	Discussion about setting up a proof of concept for data quality monitoring	30923 characters	A	None
6435838499616813897	Workplace conversation about data quality, marketing, and onboarding	29061 characters	A	None
8881856508452693151	Business discussion about data monitoring and deployment solutions	36124 characters	A	None
6921265143903363376	Discussion on web app architecture and components	61852 characters	A	None
897112651883757458	Work discussion about customer issues and project updates	56944 characters	A	None
3570569546806365081	Conversation about setting up data quality checks and notifications	55787 characters	A	None
9155996999418599991	Business discussion on competitive benchmarking and sales strategies	43378 characters	A	None
2599397938860972913	Discussion on using data enrichment and integration with outre...	63302 characters	D+	Abusive Language, Duplicate

Quickly evaluate the quality of a document



Anomalo

Thank You

Visit Anomalo.com/privatebeta to learn more and come **visit us at Booth 46**