datacockpit

a toolkit for data lake navigation & monitoring utilizing quality and usage information.


arpit **narechania**


surya **chakraborty**


shivam **agarwal**


atanu **sinha**


ryan **rossi**


fan **du**


jane **hoffswell**


shunan **guo**


eunyee **koh**


alex **endert**


shamkant **navathe**

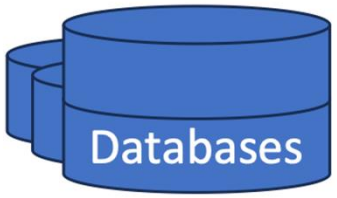
Adobe

Georgia Tech Visualization Lab

Georgia Tech

challenges

navigate
if unfamiliar



DATA LAKE

flag
irrelevant data

discover
relevant data

monitor
health of data

other tools


MONOSI


elementary
open source data observability


SQLLineage


Bigeye

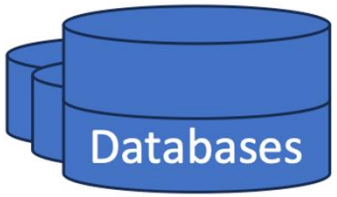

Datafold


Collibra


snowflake


MC
MONTE CARLO

opportunities



DATA LAKE

opportunities

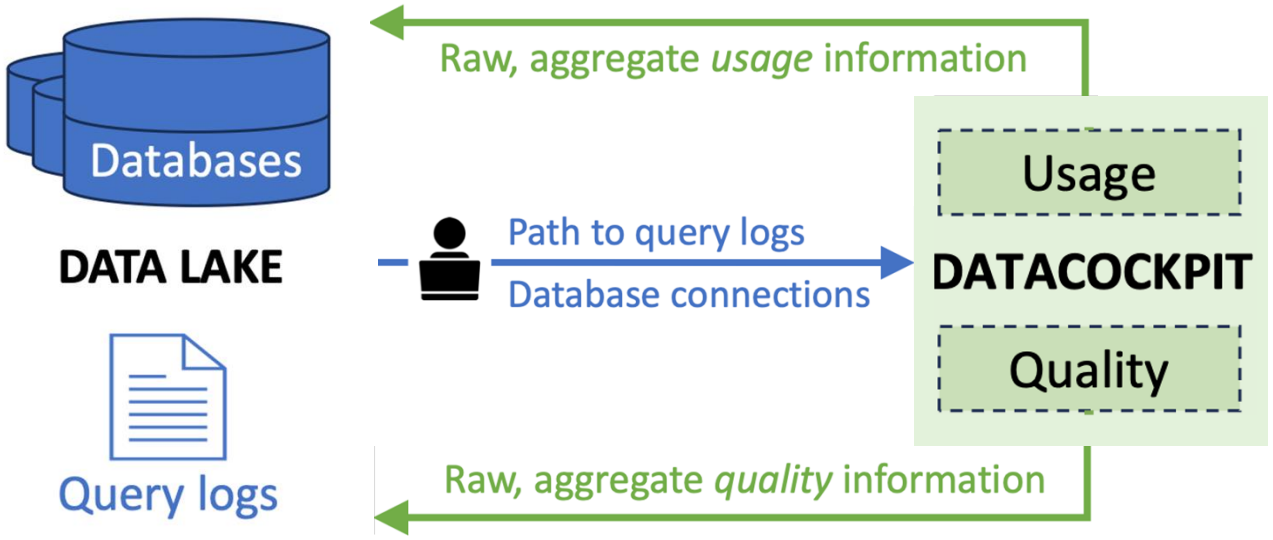
datacockpit

datacockpit

data usage

who used that data?
in what applications?
when?

datacockpit



data usage

who used that data?
in what applications?
when?

data quality

are there missing values?
what about errors?
is the dataset skewed?

datacockpit python™

data usage

who used that data?
in what applications?
when?

“The historical utilization characteristics of data across other users.”

data quality

are there missing values?
what about errors?
is the dataset skewed?

“The validity and appropriateness of data required to perform certain analytics.”

datacockpit python™

data usage

who used that data?
in what applications?
when?

“The historical utilization characteristics of data across other users.”

data quality

are there missing values?
what about errors?
is the dataset skewed?

“The validity and appropriateness of data required to perform certain analytics.”

data usage

DATA LAKE

Query logs

_id	query	db	timestamp	user	app
1	SELECT prod.name, ...	org	1683743727	nroy	dash
2	INSERT INTO prod ...	org	1684901000	rbob	cli
3	UPDATE user SET role ...	org	1685329543	sguo	hr-app

data usage

DATA LAKE

Query logs

_id	query	db	timestamp	user	app
1	SELECT prod.name, ...	org	1683743727	nroy	dash
2	INSERT INTO prod ...	org	1684901000	rbob	cli
3	UPDATE user SET role ...	org	1685329543	sguo	hr-app

```
for log in usage_logs: # iterate over every usage log
  parsed_sql = SQLParser(log[“query”]) # parse the SQL query
  datasets = # from, join, updated (including those in subqueries)
  attributes = # select, update, where, group by, order by, having
  records = # exec query with(out) where, having, mark result set.
  user = log[“user”] # the user executing the query
  application = log[“app”] # the originating application
  timestamp = log[“timestamp”] # the query execution timestamp
```

data usage

DATA LAKE

Query logs

```
for log in usage_logs: # iterate over every usage log
    parsed_sql = SQLParser(log[“query”]) # parse the SQL query
    datasets = # from, join, updated (including those in subqueries)
    attributes = # select, update, where, group by, order by, having
    records = # exec query with(out) where, having, mark result set.
    user = log[“user”] # the user executing the query
    application = log[“app”] # the originating application
    timestamp = log[“timestamp”] # the query execution timestamp
```

_id	db	data	attribute	keyword	timestamp	user	app
1	org	prod	name	SELECT	1683743727	nroy	dash
2	org	prod	-	INSERT	1684901000	rbob	cli
3	org	user	role	UPDATE	1685329543	sguo	hr-app

data usage

DATA LAKE

Query logs

_id	db	data	attribute	keyword	timestamp	user	app
1	org	prod	name	SELECT	1683743727	nroy	dash
2	org	prod	-	INSERT	1684901000	rbob	cli
3	org	user	role	UPDATE	1685329543	sguo	hr-app

attribute	t_query_count	u_user_count	u_app_count	last_used	...
name	825	23	4	1683743727	...
role	2	1	1	1685329543	...

attribute	period	t_query_count	u_user_count	u_app_count	...
name	2023-05	43	5	5	...
role	2023-05	1	5	5	...

datacockpit python™

data usage

who used that data?
in what applications?
when?

“The historical utilization characteristics of data across other users.”

data quality

are there missing values?
what about errors?
is the dataset skewed?

“The validity and appropriateness of data required to perform certain analytics.”

datacockpit python™

data usage

who used that data?
in what applications?
when?

“The historical utilization characteristics of data across other users.”

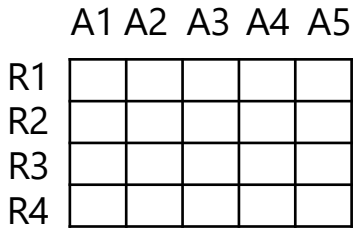
data quality

are there missing values?
what about errors?
is the dataset skewed?

“The validity and appropriateness of data required to perform certain analytics.”

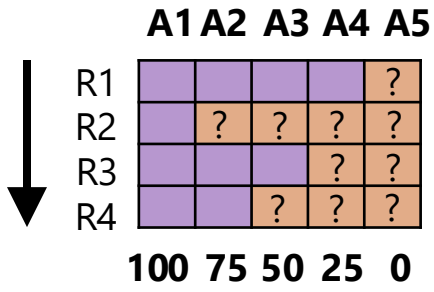
data quality: attribute completeness

percentage of non-missing values for each attribute (A1–A5)



non-missing ( )

missing (?)



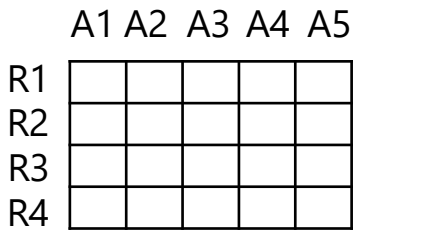
% non-missing values

Attribute Completeness

```
/* Compute percent of Non-Null values for all N attributes
 * where attr_i means the i^th attribute for i ∈ 1...N. */
SELECT CAST(COUNT(attr_1) AS FLOAT) * 100 /
            CAST (COUNT(*) AS FLOAT) AS ''attr_1'',
       CAST(COUNT(attr_2) AS FLOAT) * 100 /
            CAST (COUNT(*) AS FLOAT) AS ''attr_2'',
       ...,
       CAST(COUNT(attr_N) AS FLOAT) * 100 /
            CAST (COUNT(*) AS FLOAT) AS ''attr_N''
FROM dataset;
```

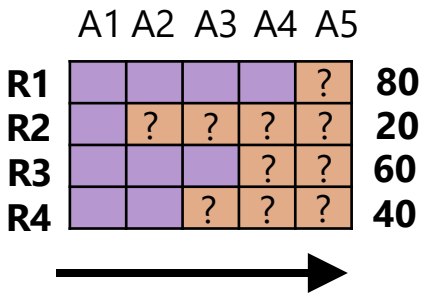
data quality: record completeness

percentage of non-missing values for each record (R1–R4)



non-missing ( )

missing (?)



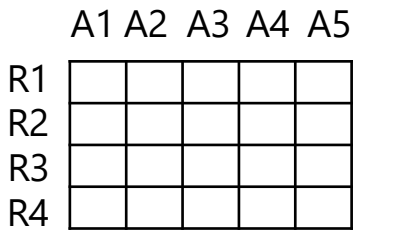
% non-missing values

Record Completeness

```
/* Compute the percent of Non-Null values for all dataset records
 * where attr_i means i^th attribute where i ∈ 1...N. */
SELECT PrimaryKey, (CAST(
    (CASE WHEN attr_1 IS NULL THEN 0 ELSE 1 END) +
    (CASE WHEN attr_2 IS NULL THEN 0 ELSE 1 END) +
    ... +
    (CASE WHEN attr_N IS NULL THEN 0 ELSE 1 END) AS
  FLOAT) * 100 / N) AS ''record_completeness''
FROM dataset;
```

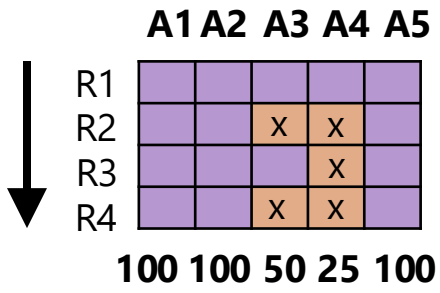
data quality: attribute correctness

percentage of correct values for each attribute (A1–A5)



correct ( )

incorrect (x)



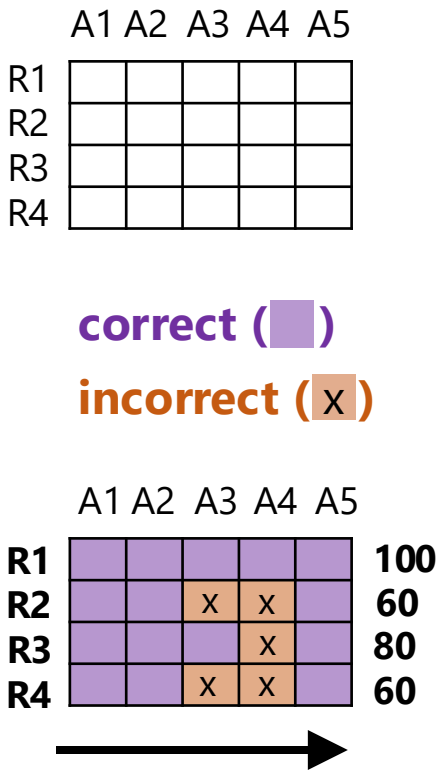
% correct values

Attribute Correctness

```
/* Compute the percent of correct values for all N attributes
 * where attr_i means i^th attribute for i ∈ 1...N. */
SELECT
  CAST(100 * CAST(SUM(CASE WHEN Age BETWEEN 0 AND 150
    THEN 1 ELSE 0 END) AS FLOAT) / COUNT(Age) AS FLOAT),
  CAST(100 * CAST(SUM(CASE WHEN Country IN ('CA','US')
    THEN 1 ELSE 0 END) AS FLOAT) / COUNT(Country) AS FLOAT),
    ...,
  CAST(100 * CAST(SUM(CASE WHEN attr_N satisfies a condition
    THEN 1 ELSE 0 END) AS FLOAT) / COUNT(attr_N) AS FLOAT)
FROM dataset;
```

data quality: record correctness

percentage of correct values for each record (R1–R4)

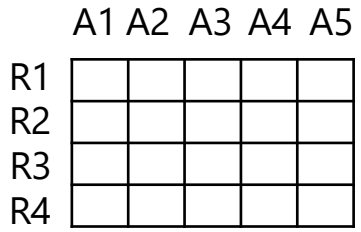


Record Correctness

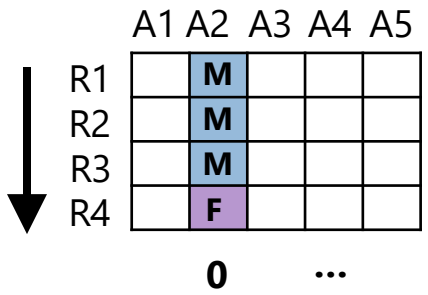
```
/* Compute the percent of correct values for all dataset records
 * where attr_i means i^th attribute where i ∈ 1...N. */
SELECT
  CAST(
    (CAST(100 * CAST(
        SUM(CASE WHEN Age BETWEEN 0 AND 150
        THEN 1 ELSE 0 END) AS FLOAT) / COUNT(Age)
            AS FLOAT) +
    CAST(100 * CAST(
        SUM(CASE WHEN Email LIKE '%_@_%.__%'
        THEN 1 ELSE 0 END) AS FLOAT) / COUNT(Email)
            AS FLOAT) +
    ...)
  / N AS FLOAT) AS ''record_correctness''
FROM dataset;
```

data quality: attribute objectivity

extent to which values conform to a target distribution for each attribute (A1–A5)



Male, Female



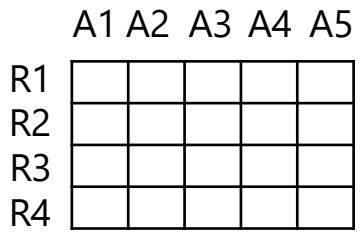
if Ms must equal Fs
objectivity = 0

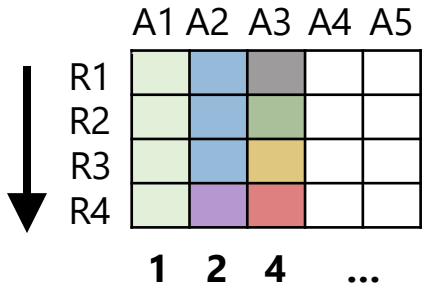
Attribute Objectivity

```
/* Compute if values for attributes are objective (100=Yes, 0=No)
 * where attr_i means i^th attribute for i ∈ 1...N. */
SELECT CAST(100 * (CASE WHEN
  (SELECT 100 * CAST(COUNT(Gender) AS float) /
    (SELECT CAST(COUNT(Gender) AS float) FROM dataset)
  FROM dataset
  WHERE Gender = ''Male'') > 60 /* where 60 = some threshold */
    THEN 0 ELSE 1 END) AS FLOAT) AS ''Is_Gender_Objective'', ...,
/* compare other attr_i distributions with a target or baseline */
FROM dataset;
```

data quality: attribute uniqueness

number of unique values for each attribute (A1–A5)





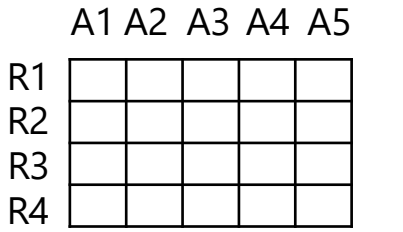
 # unique values

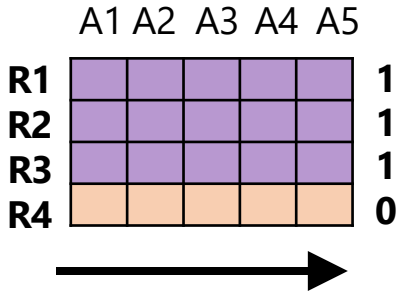
Attribute Uniqueness

```
/* Compute number of unique values for all N attributes.
 * where attr_i means i^th attribute where i ∈ 1...N */
SELECT COUNT(DISTINCT attr_1) AS ''attr_1'',
       ...,
       COUNT(DISTINCT attr_N) AS ''attr_N''
FROM dataset;
```

data quality: record freshness

whether ingestion was timely for each record (R1–R4)





fresh or not?

Record Freshness

```
/* Compute freshness of each record where ingestionDate is when
 * the record is inserted; freshnessThreshold is the unit of freshness
 * (e.g., 30 days); referenceDate is compared with ingestionDate. */
SELECT PrimaryKey,
       CASE WHEN DATEDIFF(referenceDate, ingestionDate)
        <= freshnessThreshold
            THEN 100 /* Within threshold */
            ELSE 0 /* Outside threshold */
            END
        AS ''freshness''
FROM dataset;
```

datacockpit python™

data usage

who used that data?
in what applications?
when?

“The historical utilization characteristics of data across other users.”

data quality

are there missing values?
what about errors?
is the dataset skewed?

“The validity and appropriateness of data required to perform certain analytics.”

datacockpit is an extension of datapilot

DataPilot: Utilizing Quality and Usage Information for Subset Selection during Visual Data Preparation

Arpit Narechania
Georgia Institute of Technology
Atlanta, USA
arpitnarechania@gatech.edu

Fan Du
Adobe Research
San Jose, USA
dufan2013@gmail.com

Atanu R Sinha
Adobe Research
Bengaluru, India
atr@adobe.com

Ryan A. Rossi
Adobe Research
San Jose, USA
ryrossi@adobe.com

Jane Hoffswell
Adobe Research
Seattle, USA
jhoffs@adobe.com

Shunan Guo
Adobe Research
San Jose, USA
sguo@adobe.com

Eunyee Koh
Adobe Research
San Jose, USA
eunyee@adobe.com

Shamkant B. Navathe
Georgia Institute of Technology
Atlanta, USA
sham@cc.gatech.edu

Alex Endert
Georgia Institute of Technology
Atlanta, USA
endert@gatech.edu

ABSTRACT

Selecting relevant data subsets from large, unfamiliar datasets can be difficult. We address this challenge by modeling and visualizing two kinds of auxiliary information: (1) *quality* – the validity and appropriateness of data required to perform certain analytical tasks; and (2) *usage* – the historical utilization characteristics of data across multiple users. Through a design study with 14 data workers, we integrate this information into a visual data preparation and analysis tool, DataPilot. DataPilot presents visual cues about "the good, the bad, and the ugly" aspects of data and provides graphical user interface controls as interaction affordances, guiding users to perform subset selection. Through a study with 36 participants, we investigate how DataPilot helps users navigate a large, unfamiliar tabular dataset, prepare a relevant subset, and build a visualization dashboard. We find that users selected smaller, effective subsets with higher quality and usage, and with greater success and confidence.

Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3581509

1 INTRODUCTION

Data are never truly raw [42] but still require processing through cleaning, integration, transformation, and selection before they can be utilized for their intended purposes [92]. Modern organizations often ingest all incoming data in their native form with the intent of performing analytics later [32]. The inherent information overload due to this "load-first" philosophy poses several challenges in data navigation and knowledge discovery [24, 33, 44]. For example, consider a user task, *"analyze a large e-commerce dataset and build a dashboard visualizing recent geographic trends for predicting future sales."* To perform this task, users must first identify relevant data attributes pertaining to customers' locations (e.g., "ZipCode") and then select the desired data records by applying a temporal filter

datacockpit is an extension of datapilot

DataPilot: Utilizing Quality and Usage Information for Subset Selection during Visual Data Preparation

Arpit Narechania
Georgia Institute of Technology
Atlanta, USA
arpitnarechania@gatech.edu

Fan Du
Adobe Research
San Jose, USA
dufan2013@gmail.com

Atanu R Sinha
Adobe Research
Bengaluru, India
atr@adobe.com

Ryan A. Rossi
Adobe Research
San Jose, USA
ryrossi@adobe.com

Jane Hoffswell
Adobe Research
Seattle, USA
jhoffs@adobe.com

Shunan Guo
Adobe Research
San Jose, USA
sguo@adobe.com

Eunyee Koh
Adobe Research
San Jose, USA
eunyee@adobe.com

Shamkant B. Navathe
Georgia Institute of Technology
Atlanta, USA
sham@cc.gatech.edu

Alex Endert
Georgia Institute of Technology
Atlanta, USA
endert@gatech.edu

ABSTRACT

Selecting relevant data subsets from large, unfamiliar datasets can be difficult. We address this challenge by modeling and visualizing two kinds of auxiliary information: (1) *quality* – the validity and appropriateness of data required to perform certain analytical tasks; and (2) *usage* – the historical utilization characteristics of data across multiple users. Through a design study with 14 data workers, we integrate this information into a visual data preparation and analysis tool, DataPilot. DataPilot presents visual cues about "the good, the bad, and the ugly" aspects of data and provides graphical user interface controls as interaction affordances, guiding users to perform subset selection. Through a study with 36 participants, we investigate how DataPilot helps users navigate a large, unfamiliar tabular dataset, prepare a relevant subset, and build a visualization dashboard. We find that users selected smaller, effective subsets with higher quality and usage, and with greater success and confidence.

Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3581509

1 INTRODUCTION

Data are never truly raw [42] but still require processing through cleaning, integration, transformation, and selection before they can be utilized for their intended purposes [92]. Modern organizations often ingest all incoming data in their native form with the intent of performing analytics later [32]. The inherent information overload due to this "load-first" philosophy poses several challenges in data navigation and knowledge discovery [24, 33, 44]. For example, consider a user task, *"analyze a large e-commerce dataset and build a dashboard visualizing recent geographic trends for predicting future sales."* To perform this task, users must first identify relevant data attributes pertaining to customers' locations (e.g., "ZipCode") and then select the desired data records by applying a temporal filter

datapilot is a visual data preparation and analysis tool that also computed quality and usage metrics but for one tabular dataset.

a user study

- to understand how users utilize these metrics to select data subsets and make visualization dashboards
- showed that datapilot users selected smaller, more effective subsets with higher quality and usage scores.

we then developed datacockpit to expand to multiple relational databases (our data lake).

datacockpit expanding to multiple databases…

recap: each quality and usage metric across each attribute and record of a table is first **normalized** to **a score between 0-100**.

then, an ***overall score*** is derived:

for usage: as the ***max*** of individual scores (***max*** because not all data is used across all applications and several low scores would demotivate users).

for quality: as the ***mean*** of individual scores.

finally, overall scores are **aggregated to higher granularities of data** (e.g., dataset-level) using a ***weighted average*** *(as some metrics are more important than others).*

$$\text{AGGRSCORE} = \frac{w_1 E_1 + w_2 E_2 + \ldots + w_n E_n}{w_1 + w_2 + \ldots + w_n}$$

pip install datacockpit

```
from datacockpit import DataCockpit
# Setup database connection with SQL engine and log-file CSV
dcp_obj = DataCockpit(engines=[array_of_sqlalchemy_engines],
        logs_paths=[array/of/your/logs/path.csv])

# Compute and persist quality & usage metrics
dcp_obj.compute_usage(levels=None, metrics=None)
dcp_obj.compute_quality(levels=None, metrics=None)

# Retrieve computed information for use in downstream apps
usage_info = dcp_obj.get_usage(**kwargs)
quality_info = dcp_obj.get_quality(**kwargs)
```

Applications

navigate
if unfamiliar

discover
relevant data

flag
irrelevant data

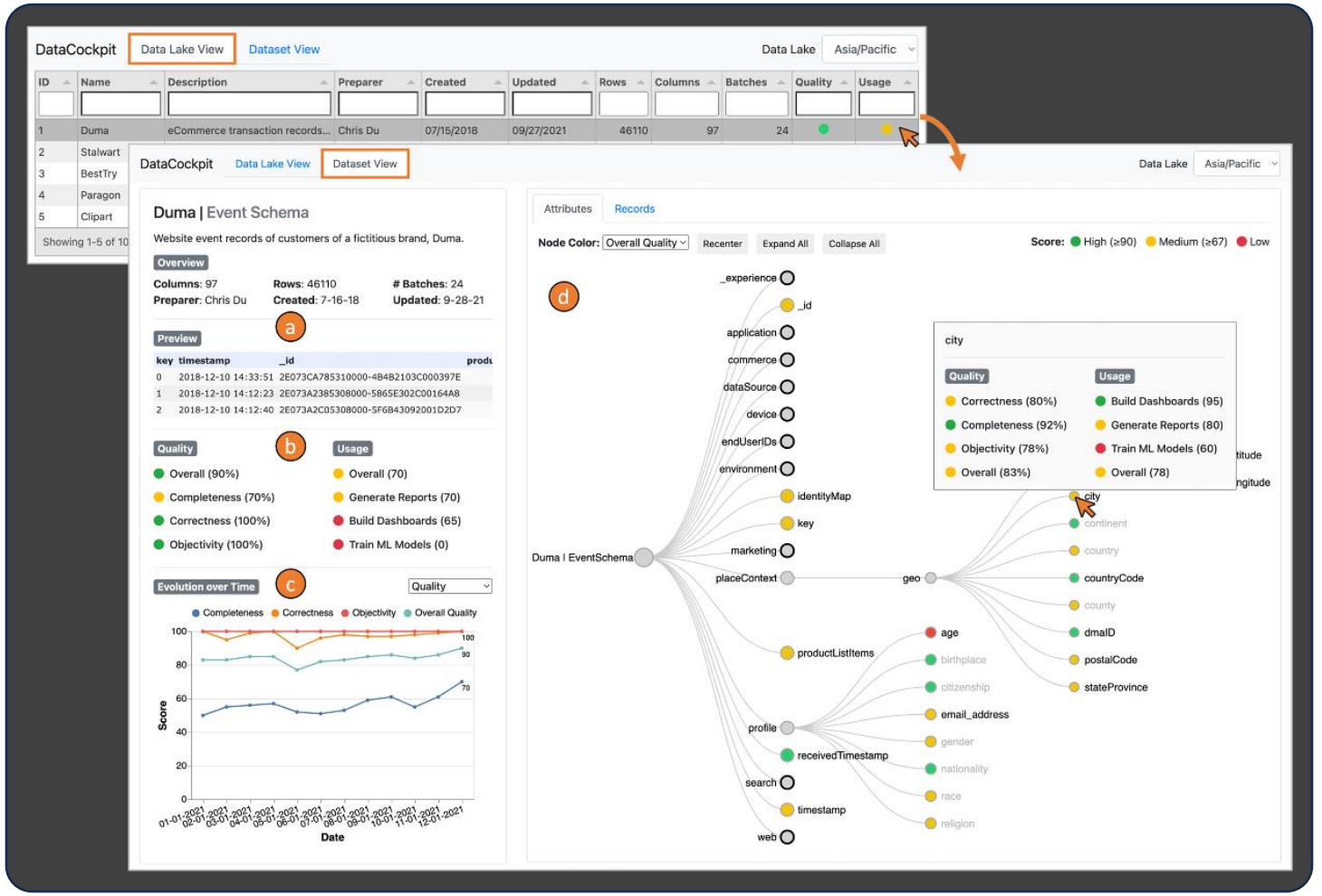
monitor
health of data

visual monitoring tool

JavaScript

navigate
if unfamiliar

discover
relevant data



flag
irrelevant data

monitor
health of data

DataCockpit | Data Lake View | Dataset View

Data Lake: Asia/Pacific

ID	Name	Prepared By	Created On	Last Updated On	Quality	Usage
1	Duma	Chris Du	07/15/2018	09/27/2021		
2	Stalwart	Pamela C. Michels	07/04/2015	05/13/2022		
3	BestTry	Gavin M. Elliott	07/19/2019	05/18/2020		
4	Paragon	John S. Page	12/10/2020	01/01/2021		
5	Clipart	Jeff M. Windsor	07/06/2022			
6	Depay	Christa B. Taylor	07/15/2018	09/27/2021		
7	Aloevera	Pamela C. Michels	07/04/2015	05/13/2022		
8	Shopium	Gavin M. Elliott	07/19/2019	05/18/2020		
9	Jing	John S. Page	12/10/2020	01/01/2021		
10	Ecom101	Jeff M. Windsor	07/06/2022			

Showing 1-10 of 10 rows

First | Prev | 1 | Next | Last

applications and users

flag
irrelevant data

monitor
health of data

discover
relevant data

navigate
if unfamiliar

Data / System Administrators can efficiently find low quality, low recent usage datasets and mark them for archival (cold storage).

They can also monitor shifting trends in the data (e.g., “age” is being queried a lot).

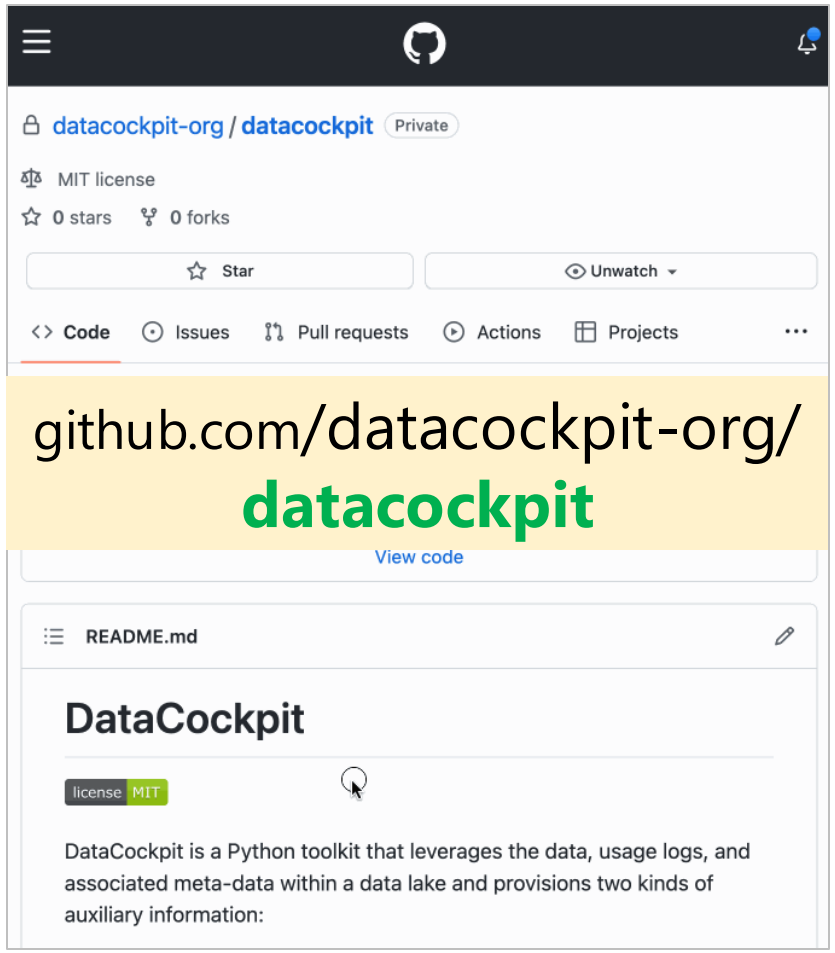
Experienced end-users can re-use ‘tried, tested, and effective’ attributes from past analysis and dashboards for future ones.

New users can efficiently learn these metrics across all datasets, flattening their learning curve.

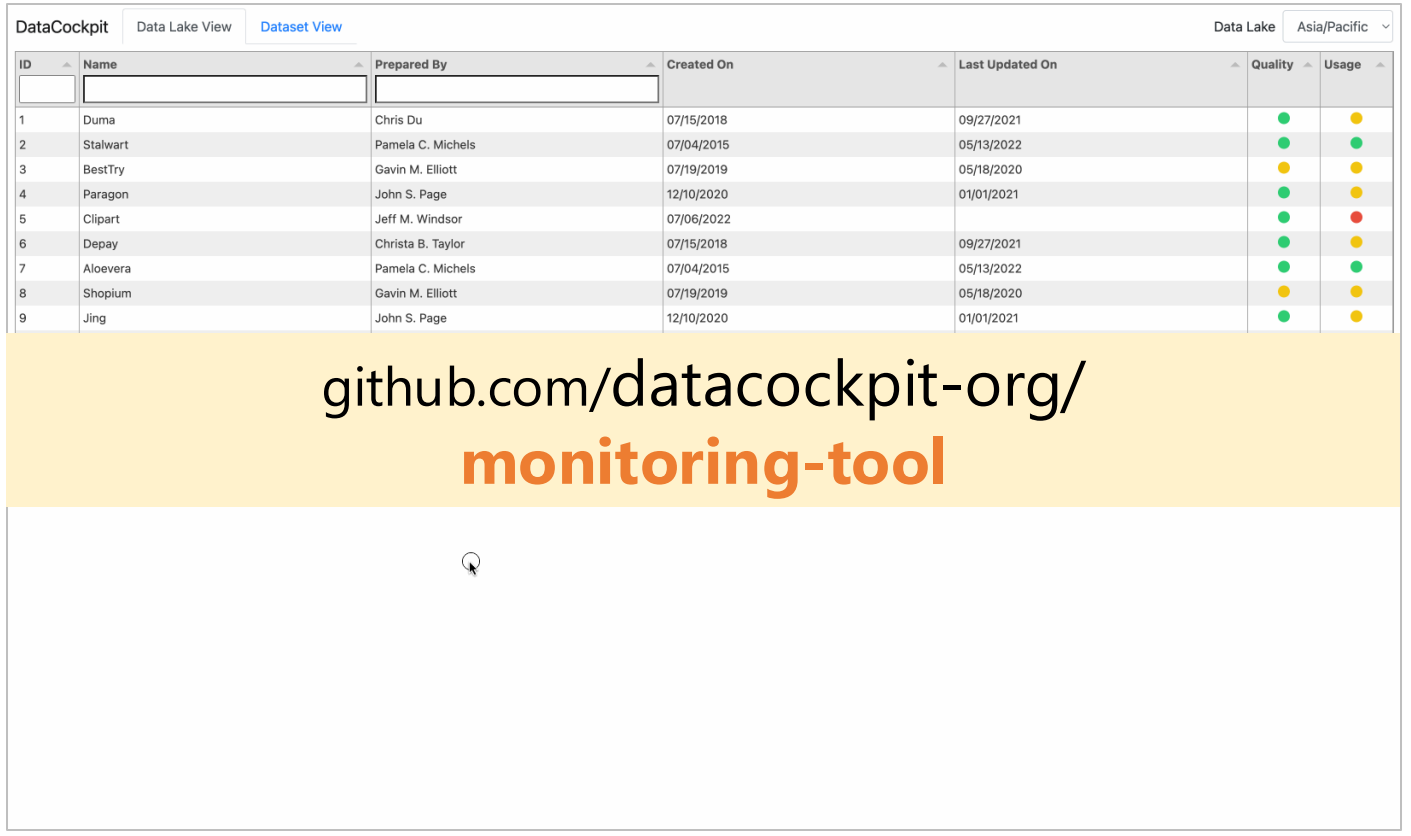
Developers can create usage and quality powered tracking and monitoring capabilities into their own new tools.

datacockpit

a toolkit for data lake navigation & monitoring utilizing quality and usage information.



github.com/datacockpit-org/
datacockpit



ID	Name	Prepared By	Created On	Last Updated On	Quality	Usage
1	Duma	Chris Du	07/15/2018	09/27/2021		
2	Stalwart	Pamela C. Michels	07/04/2015	05/13/2022		
3	BestTry	Gavin M. Elliott	07/19/2019	05/18/2020		
4	Paragon	John S. Page	12/10/2020	01/01/2021		
5	Clipart	Jeff M. Windsor	07/06/2022			
6	Depay	Christa B. Taylor	07/15/2018	09/27/2021		
7	Aloevera	Pamela C. Michels	07/04/2015	05/13/2022		
8	Shopium	Gavin M. Elliott	07/19/2019	05/18/2020		
9	Jing	John S. Page	12/10/2020	01/01/2021		

github.com/datacockpit-org/
monitoring-tool

what's next for datacockpit?

support **more quality metrics**, e.g., consistency.

provide **advanced analyses**, e.g., anomaly detection.

design a user interface to **make it easy for users to configure thresholds**, e.g., the criteria for what is considered "missing" or "incorrect".

become privacy-friendly by serving user requests as per an organization's data governance policies.

support more data sources (e.g., JSON-based, Parquet)

datacockpit

a toolkit for data lake navigation & monitoring utilizing quality and usage information.

arpit narechania
anarechania3
@gatech.edu

surya **chakraborty**
schakraborty76
@gatech.edu

shivam **agarwal**
s.agarwal
@gatech.edu

atanu sinha
atr
@adobe.com

ryan rossi
ryrossi
@adobe.com

fan du
dufan2013
@gmail.com

jane hoffswell
jhoffs
@adobe.com

shunan guo
sguo
@adobe.com

eunyee koh
eunyee
@adobe.com

alex endert
endert
@gatech.edu

shamkant navathe
shamkant.navathe
@cc.gatech.edu

github.com/datacockpit-org/
datacockpit

github.com/datacockpit-org/
monitoring-tool

Adobe

Georgia Tech Visualization Lab

Georgia Tech