
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

ANT222

Monitor and manage data quality in your data lake with AWS Glue

Brian Ross

Senior Software Development Manager
AWS

Shiv Narayanan

Product Manager
AWS

Sudeep Jos

VP, Chief Data & Information Architect
Travelers

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

MODERN DATA PLATFORMS



Data Warehouse



Operational Data Stores



Data Lakes



Data Mesh

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue helps you build these data platforms

AWS Glue is a Serverless Data Integration Service

100K+

AWS Glue customers

100M+

AWS Glue jobs in a month

99.95%

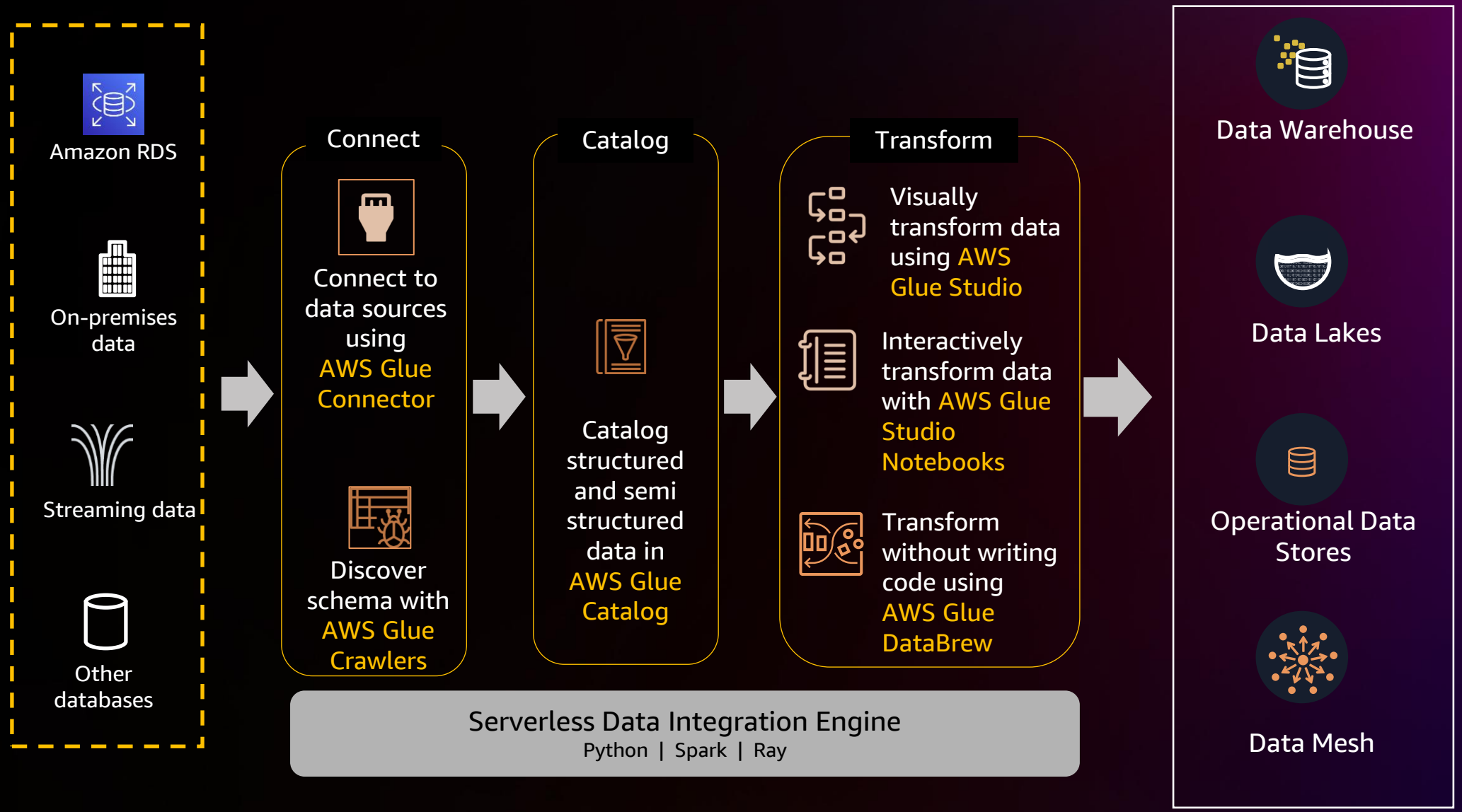
AWS Glue API reliability

100%

open-source API compliant

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue helps you build these data platforms

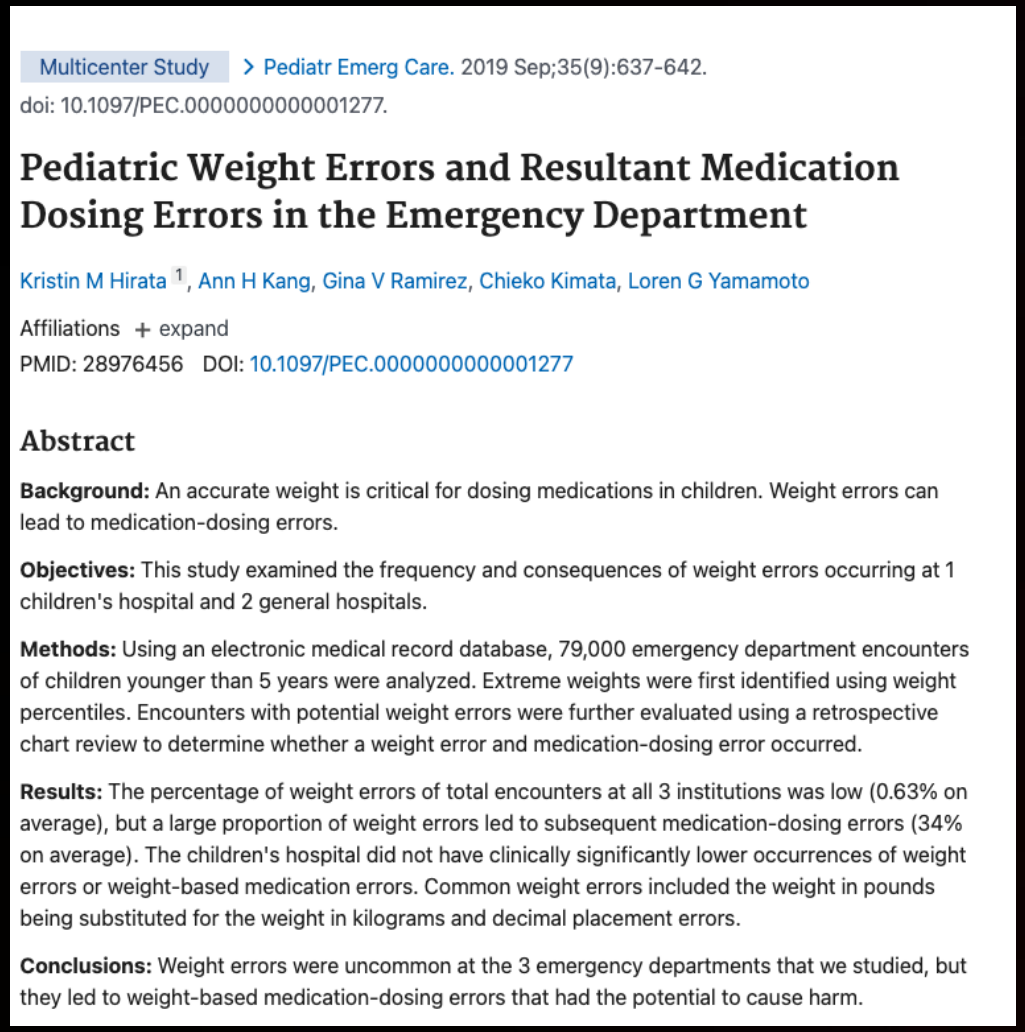


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



The quality of your analytics is only as good as the quality of your data

Bad data can have life-threatening consequences



Multicenter Study > Pediatr Emerg Care. 2019 Sep;35(9):637-642.
doi: 10.1097/PEC.0000000000001277.

Pediatric Weight Errors and Resultant Medication Dosing Errors in the Emergency Department

Kristin M Hirata [1], Ann H Kang, Gina V Ramirez, Chieko Kimata, Loren G Yamamoto

Affiliations + expand

PMID: 28976456 DOI: 10.1097/PEC.0000000000001277

Abstract

Background: An accurate weight is critical for dosing medications in children. Weight errors can lead to medication-dosing errors.

Objectives: This study examined the frequency and consequences of weight errors occurring at 1 children's hospital and 2 general hospitals.

Methods: Using an electronic medical record database, 79,000 emergency department encounters of children younger than 5 years were analyzed. Extreme weights were first identified using weight percentiles. Encounters with potential weight errors were further evaluated using a retrospective chart review to determine whether a weight error and medication-dosing error occurred.

Results: The percentage of weight errors of total encounters at all 3 institutions was low (0.63% on average), but a large proportion of weight errors led to subsequent medication-dosing errors (34% on average). The children's hospital did not have clinically significantly lower occurrences of weight errors or weight-based medication errors. Common weight errors included the weight in pounds being substituted for the weight in kilograms and decimal placement errors.

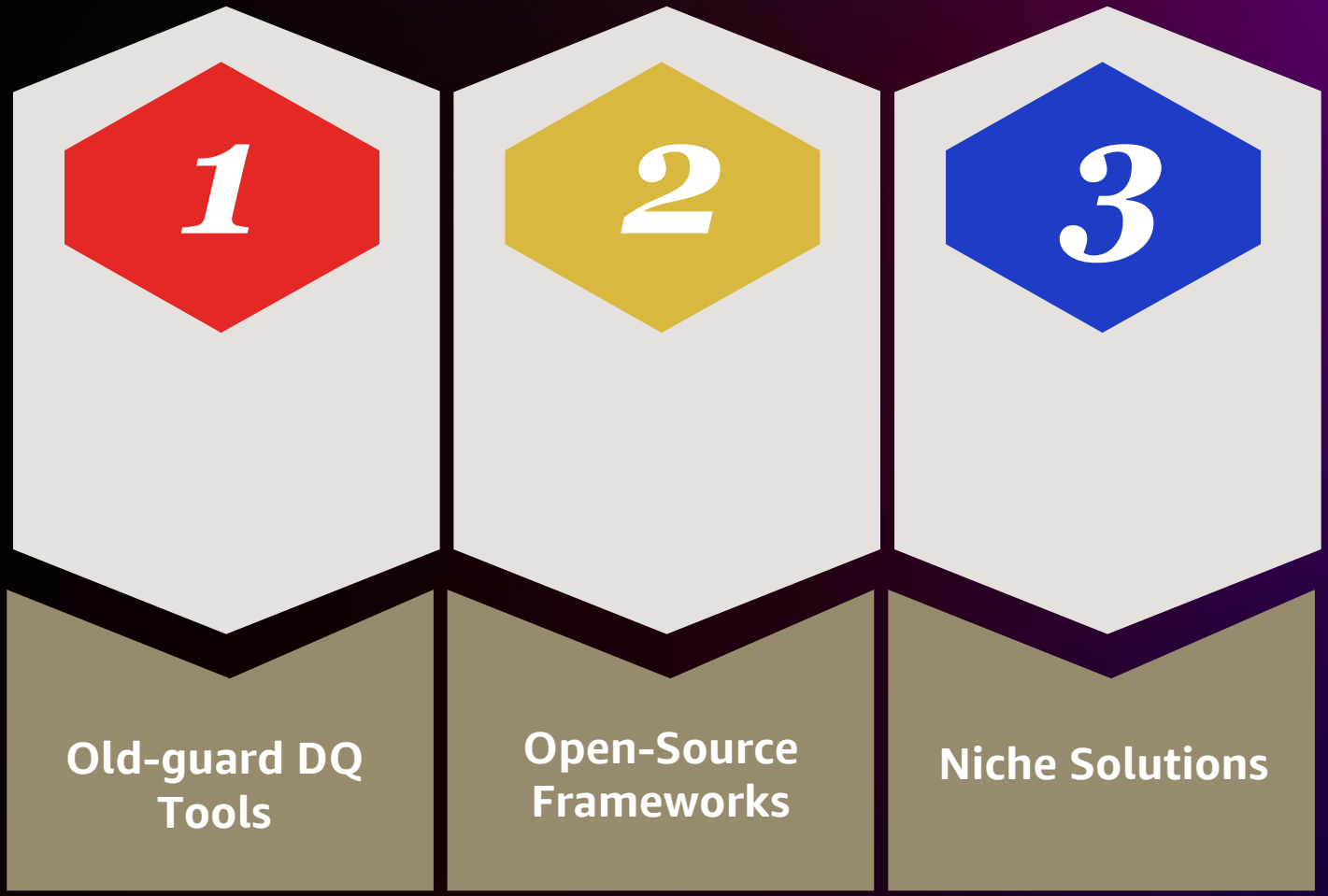
Conclusions: Weight errors were uncommon at the 3 emergency departments that we studied, but they led to weight-based medication-dosing errors that had the potential to cause harm.

Incorrect or absent recording of patient weights, though, can lead to medication dosage errors. Hirata and colleagues examined the frequency and consequences of weight errors that occurred across 79,000 emergency department encounters of children under the age of 5 years. They revealed that, although weight errors were relatively rare (0.63%), a large proportion of weight errors led to subsequent medication-dosing errors (34%).

©Hannelore Aerts, Dipak Kalra, Carlos Sáez, Juan Manuel Ramírez-Anguita, Miguel-Angel Mayer, Juan M Garcia-Gomez, Marta Durà-Hernández, Geert Thienpont, Pascal Coorevits. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 04.08.2021.

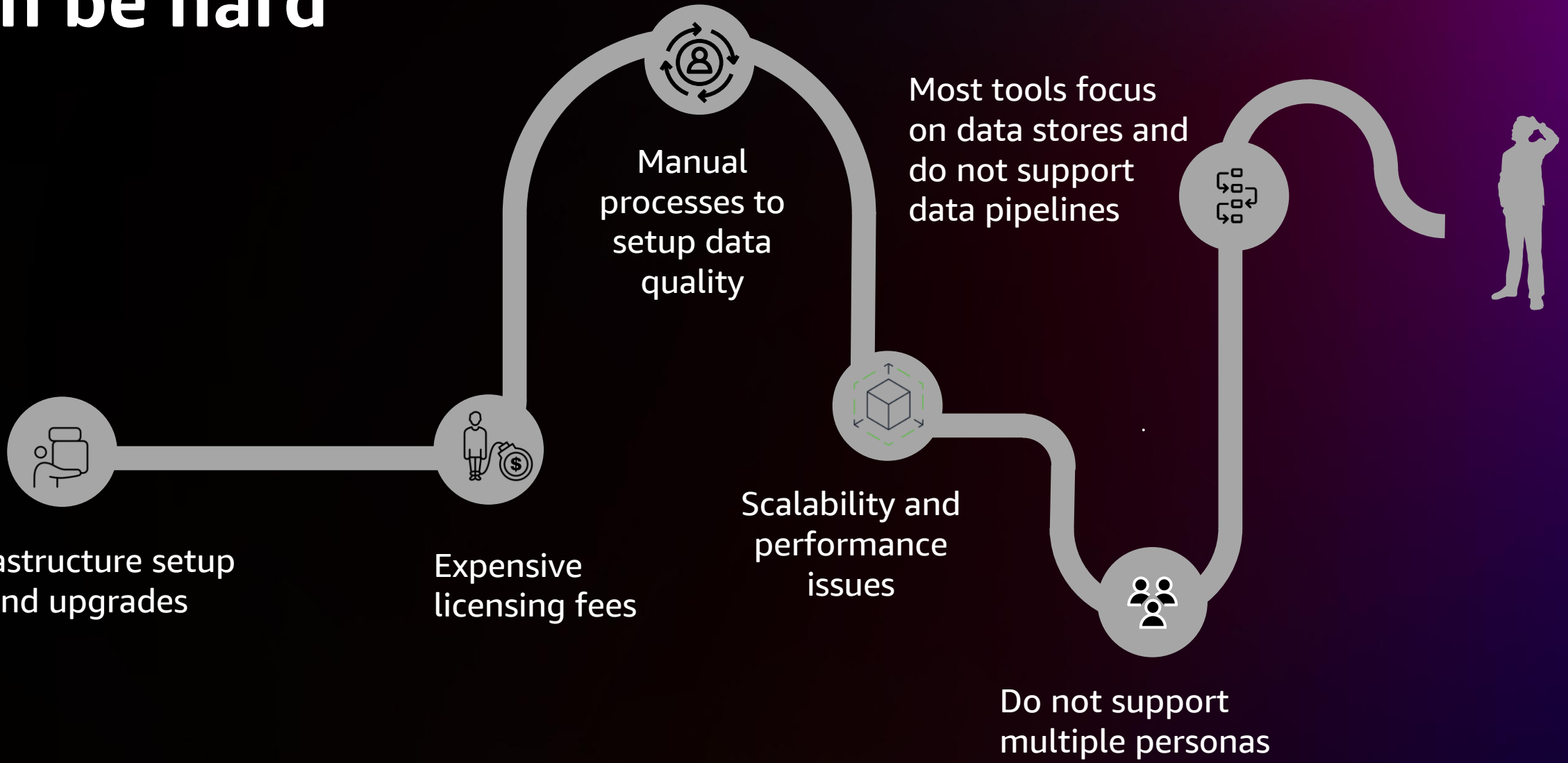
aws © 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

There are options to manage data quality today



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

However, managing quality with these tools can be hard



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



AWS GLUE DATA QUALITY

(PREVIEW)

AWS Glue Data Quality Features and Benefits



Serverless, scalable, and performant



Data Quality Recommendations



Out-of-the-box data quality rules and actions



Data Quality Definition Language



Data Quality for data at rest and data in transit

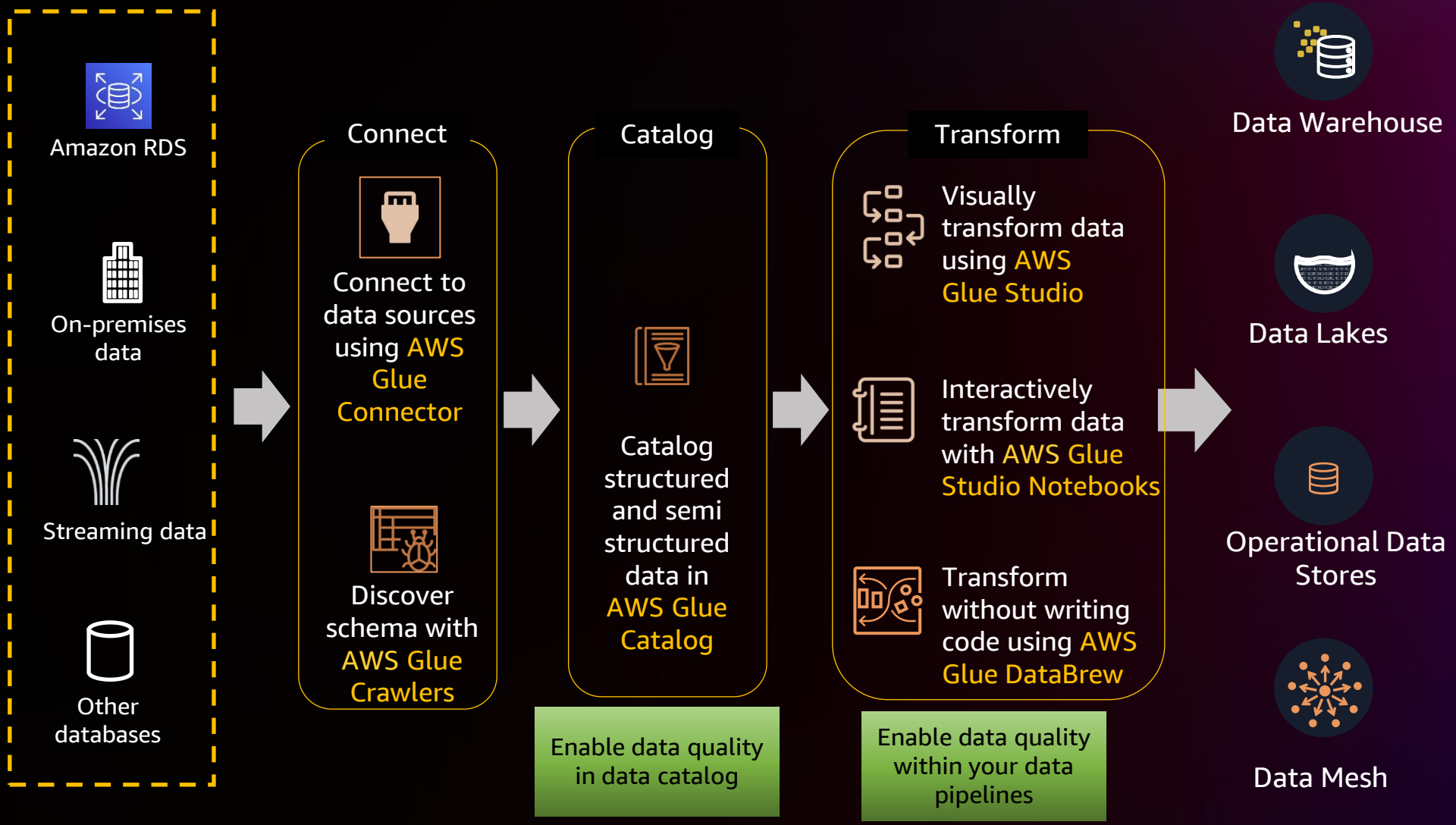


Supports multiple personas

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data Quality where you need it



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How it works: Data Quality on AWS Glue Data Catalog



Data Steward

Selects a dataset in AWS Glue Data Catalog

AWS Glue Data Quality analyzes data and recommends rules

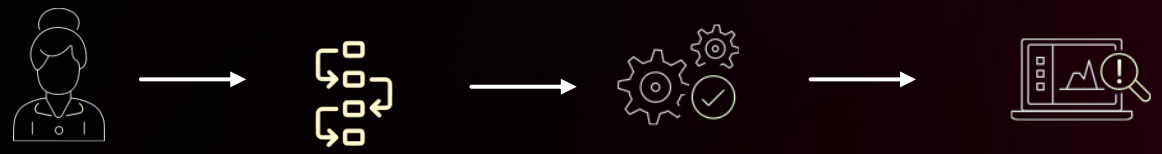
Data Steward refines the rules to create finalized rules

AWS Glue Data Quality Evaluates rules

Data Steward reviews results and alerts, takes appropriate action

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How it works: Data Quality on AWS Glue ETL Pipelines



Data Engineer

Selects an ETL Job and adds Data Quality rules and actions

AWS Glue Data Quality Evaluates rules

Engineer reviews results and alerts, takes appropriate action

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

DeeQu powers AWS Glue Data Quality

DeeQu is an open-source library developed by Amazon to manage and monitor data quality



Internally used by Amazon to manage data quality of a 60 PB data lake

Open-source package that allows the data quality rules you create to be executed in other environments

Many customers already use DeeQu to manage data quality

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

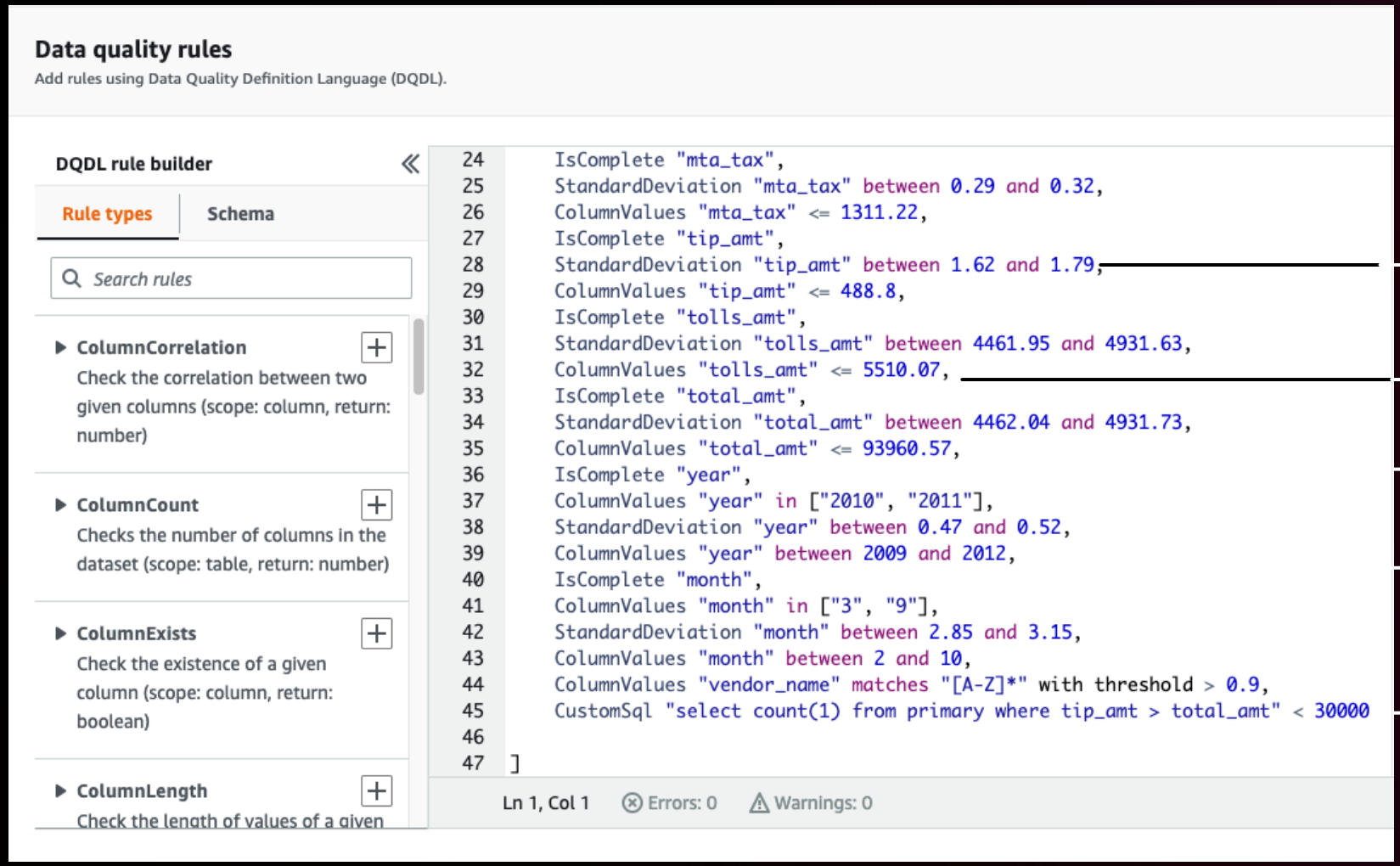
Defining Data Quality rules are hard and complex

1. Developers need to write complex code to implement DQ rules
2. Business users need to read code to understand the rules
3. Migrating and reusing data quality rules is hard
4. Data Quality rules are written in proprietary languages

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data Quality Definition Language

Declarative language to author and reuse data quality rules



- Easy to author
- Simple to understand
- Complex rule support
- Reusable and easy to deploy
- Author SQL-based rules

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Out-of-the-box Rule Types

Consistency

Column Correlation
Column Data Type
Column Exists
Column Length
Column Names Match Pattern

Accuracy

Column Values
Custom Rule with SQL support
Mean
Row Count
Standard Deviation
Sum
Distinct Value Count

Rule Types

Completeness

Completeness
Completeness with threshold
Entropy
Data Freshness

Integrity

Is Primary Key
Uniqueness
Uniqueness with threshold
Unique Value Ratio

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue Data Quality Pricing

	AWS Glue Data Catalog	AWS Glue ETL
Get data quality recommendations	$0.44 per DPU Hour	-
Evaluate Data Quality	$0.44 per DPU Hour	$0.44 per DPU Hour $0.29 per FLEX DPU Hour

Examples

Scenario	Calculation	Cost
Get recommendations for a dataset Job Execution time: 5 min Number of DPUs: 2	5 min * $0.44/60 * 2 DPUs	$0.07
Evaluate data quality of the dataset Execution Time: 10 min Number of DPUs: 4	10 min * $0.44/60 * 4 DPUs	$0.29
Evaluate data quality of a dataset using FLEX DPU (non-SLA sensitive workload) Execution Time: 10 min Number of DPUs: 4	10 min *0.29/60 *4 DPUs	$0.19

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

“Augmenting Vyaire’s AWS cloud-native data architecture with AWS Glue Data Quality drives significant efficiency in our global data management process across IT and business resources,” said Gopal Ramamurthi, Vice President of Global Data Management and Digital Health at Vyaire Medical. “Based on our testing, we estimate AWS Glue Data Quality will save us 1,500 hours that would have been spent building custom solutions and manual rule creation across our IT and business resources.”





"United Airlines is building hundreds of data and analytics driven tools for their customers *and* employees, which makes managing and maintaining data quality critical to our operations.” said Ashok Srinivas, Director of ML Engineering and Sarang Bapat, Director of Data Engineering “We are excited about AWS Glue Data Quality, which will automatically identify, analyze, and take act *on* data quality issues in *a* matter of minutes. This will help us *make* informed, timely, and accurate decisions and save countless hours in manually identifying and fixing all data issues.”

“The time and complexity associated with installing open source packages has made it very challenging to implement data quality,” said Gaurav Singla, architect at RX Global. “We were able get started very quickly using AWS Glue Data Quality without any installations. The out-of-the-box data quality rules and scalability makes it easy for us to integrate data quality into our data pipelines to prevent bad data from entering our data lakes.”



aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Customer Experience: Travelers

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The Travelers Companies, Inc

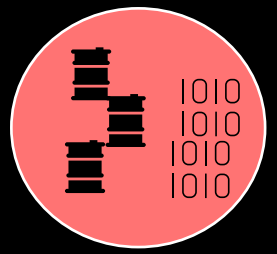
The Travelers Companies, Inc. (NYSE: TRV) is a leading provider of property casualty insurance for auto, home and business. *The company's diverse business lines offer its global customers a wide range of coverage sold primarily through independent agents and brokers.* A component of the Dow Jones Industrial Average, Travelers has approximately 30,000 employees and operations in the United States and selected international markets. The company generated revenues of approximately $35 billion in 2021.

Hiring Architects, Engineers, Data Scientists and many more roles...

Visit: https://careers.travelers.com/

© 2022 The Travelers Indemnity Company. All rights reserved

Acceleration of the Data-Centric World



Tsunami of Data



AI Driven Automation



Accelerated Digital Adoption



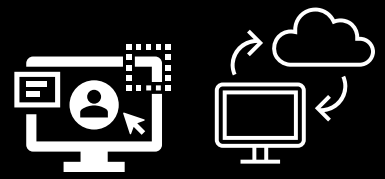
Insurance Industry Innovation

As Data and Analytics drive more business decisions, quality of data becomes critical for making accurate and responsible decisions

© 2022 The Travelers Indemnity Company. All rights reserved

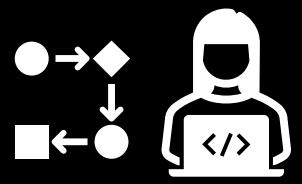
Three modes of Quality Monitoring

Preventive



At point of entry.
Preventing bad data from entering the ecosystem in the first place.

Proactive



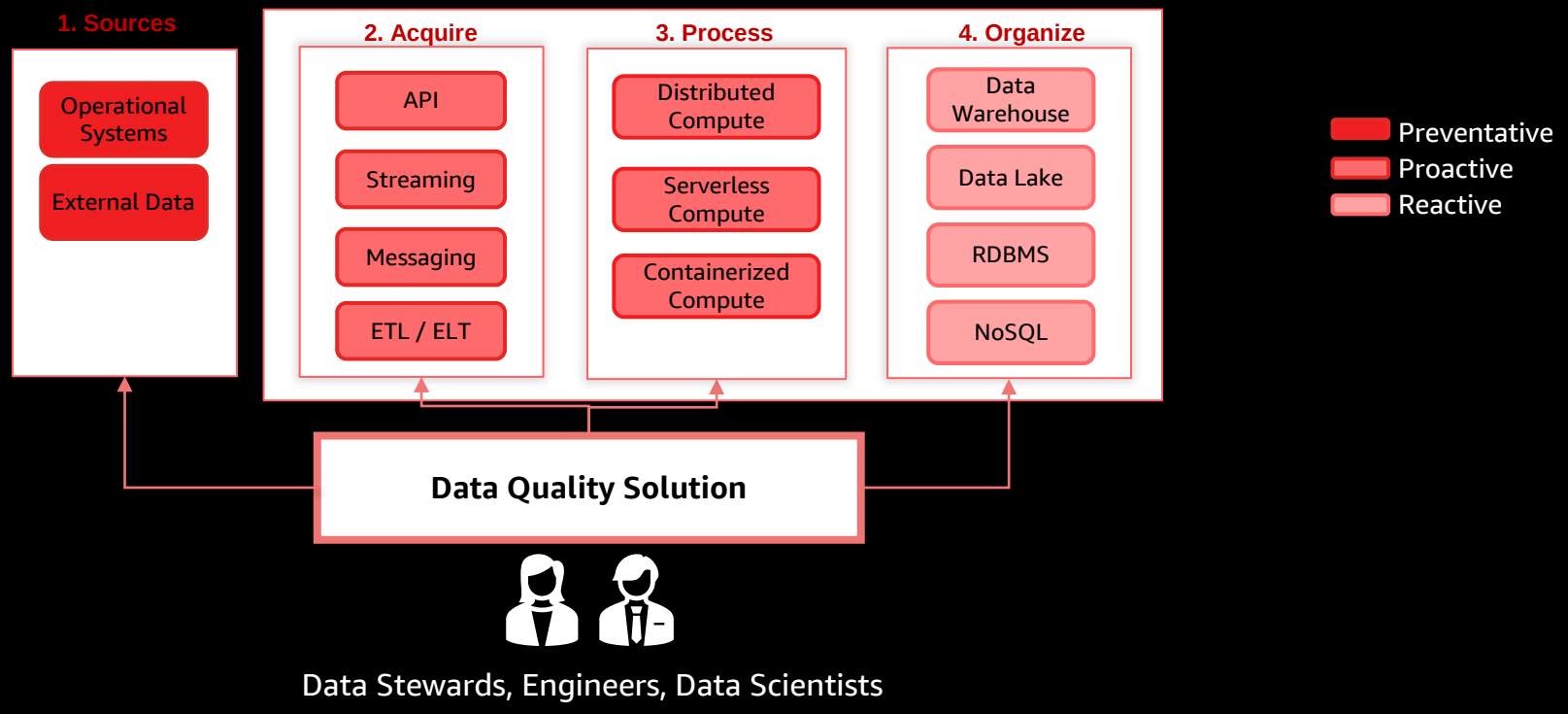
Monitoring of data processing pipelines in RT, NRT, Batch to observe/act on quality issues before landing in a mart/lake/warehouse

Reactive



Traditional monitoring of quality against at-Rest data assets.

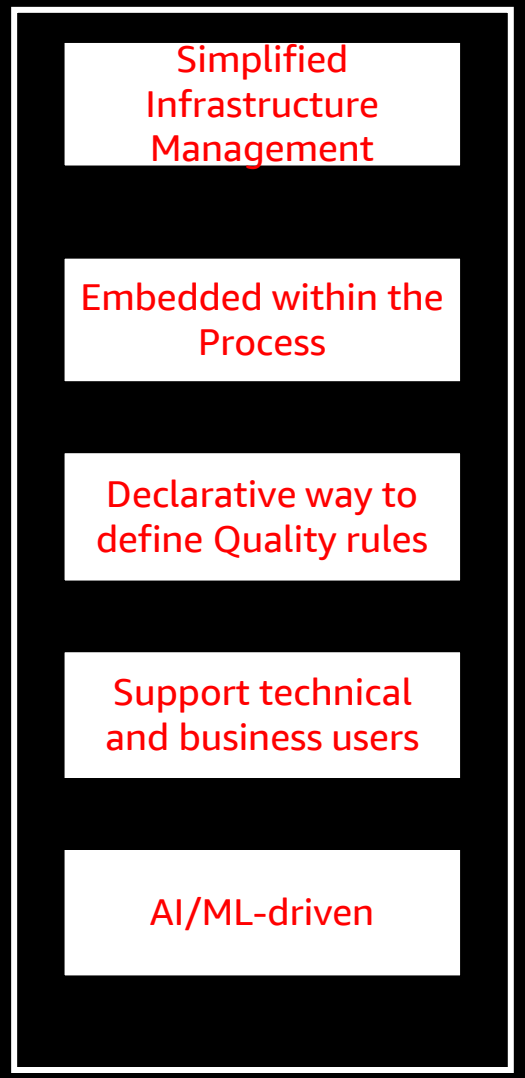
Vision: Pluggable Data Quality Solution



Having a flexible and scalable data quality solution that can cater to the Preventive, Proactive, and Reactive patterns of Data Quality monitoring will help accelerate the goal of increasing quality and trust in business decisions



AWS Glue Data Quality alignment with the vision



Eliminates infrastructure challenges through serverless architecture.

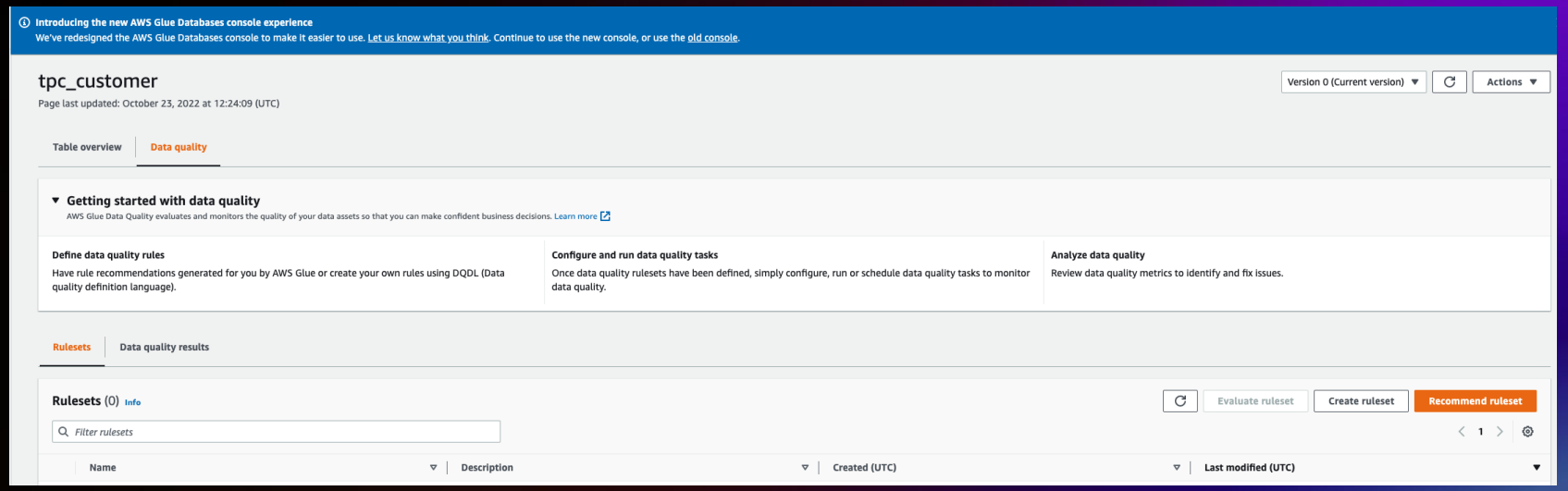
Out-of-the-box APIs, rules and actions allow us to integrate data quality within data pipelines and business processes.

Data Quality Definition Language (DQDL) allows creation of data quality rules without writing complex code.

Enables business users to setup data quality rules in AWS Glue Data Catalog while allowing data engineers to setup rules in data pipelines.

Proactive recommendations on quality rules

Call to Action



Try out AWS Glue Data Quality!

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Brian Ross
Senior Software Development Manager, AWS **Glue**

Shiv Narayanan
Product Manager, AWS Glue

Sudeep Joseph
VP, Chief Data & Information Architect, Travelers

Connect with us on LinkedIn by scanning the QR code



Please complete the session survey in the **mobile app**

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.