AWS Glue Data Quality

Deliver high-quality data across
your data lakes and pipelines

Agenda

Challenges

Why AWS Glue Data Quality?

Use cases

Key customer case studies

Get started

Modern data platforms

Data warehouse

Operational data stores

Data lakes

Data mesh

AWS Glue helps you build these data platforms

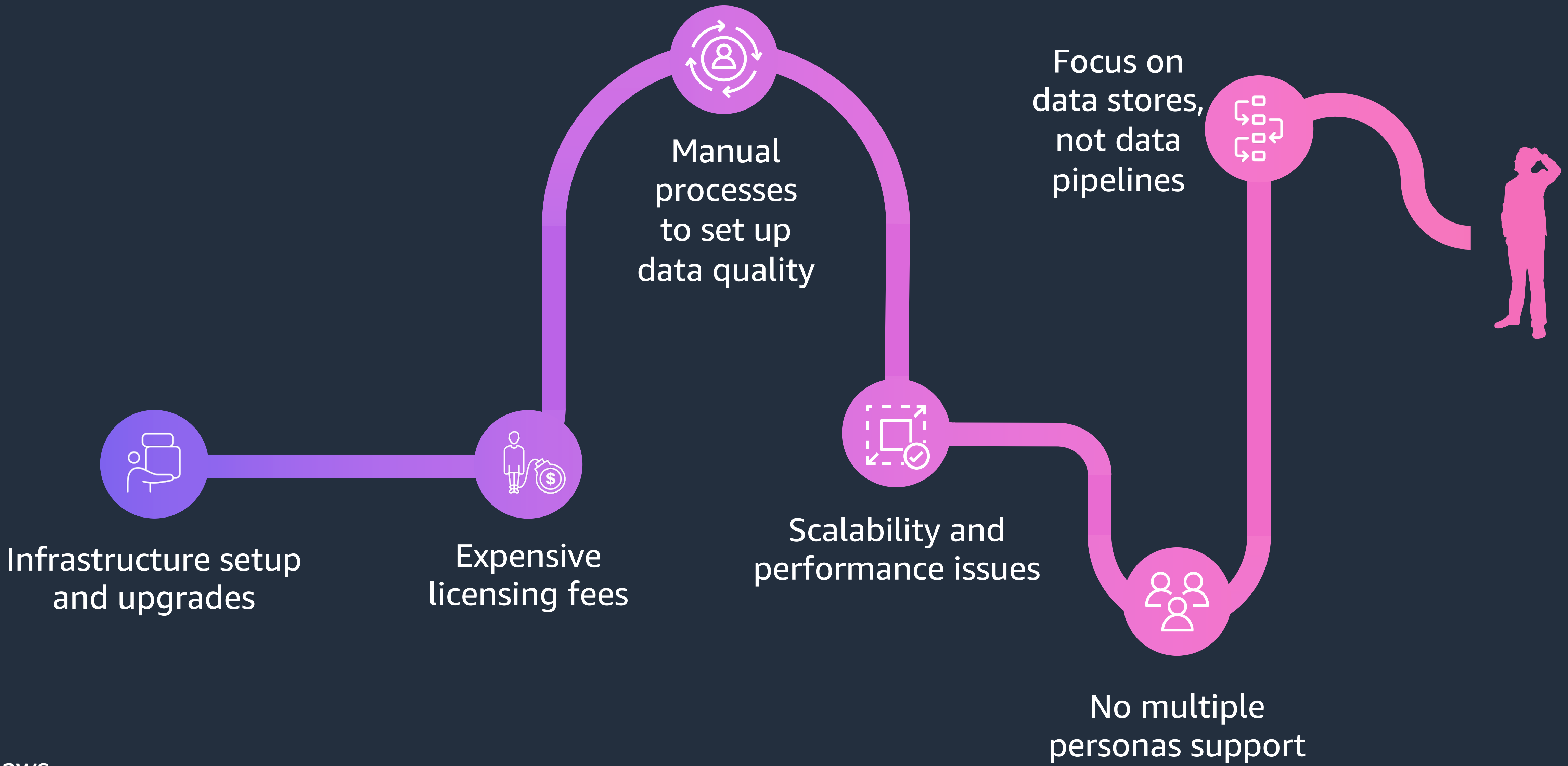
AWS Glue is a serverless data integration and ETL service

100K+	25M+	100%
AWS Glue customers	AWS Glue jobs per month	Open-source API compliant

There are options to manage data quality today

1. Old-guard data quality tools
2. Open-source frameworks
3. Niche solutions

Managing quality with these tools can be difficult



Infrastructure setup and upgrades

Expensive licensing fees

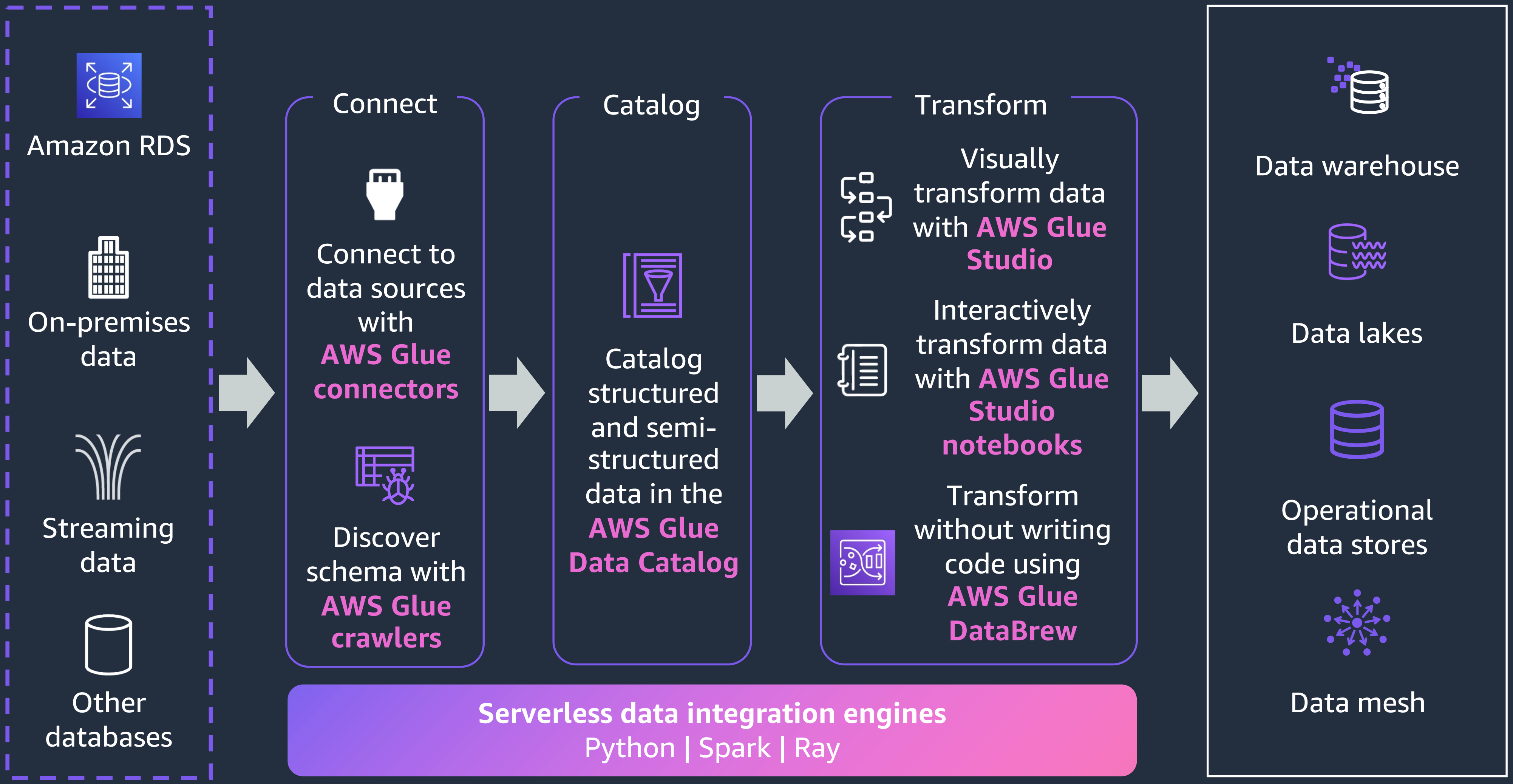
Manual processes to set up data quality

Scalability and performance issues

No multiple personas support

Focus on data stores, not data pipelines

AWS Glue helps you build these data platforms



Amazon RDS

On-premises data

Streaming data

Other databases

Connect

Connect to data sources with **AWS Glue connectors**

Discover schema with **AWS Glue crawlers**

Catalog

Catalog structured and semi-structured data in the **AWS Glue Data Catalog**

Transform

Visually transform data with **AWS Glue Studio**

Interactively transform data with **AWS Glue Studio notebooks**

Transform without writing code using **AWS Glue DataBrew**

Serverless data integration engines
Python | Spark | Ray

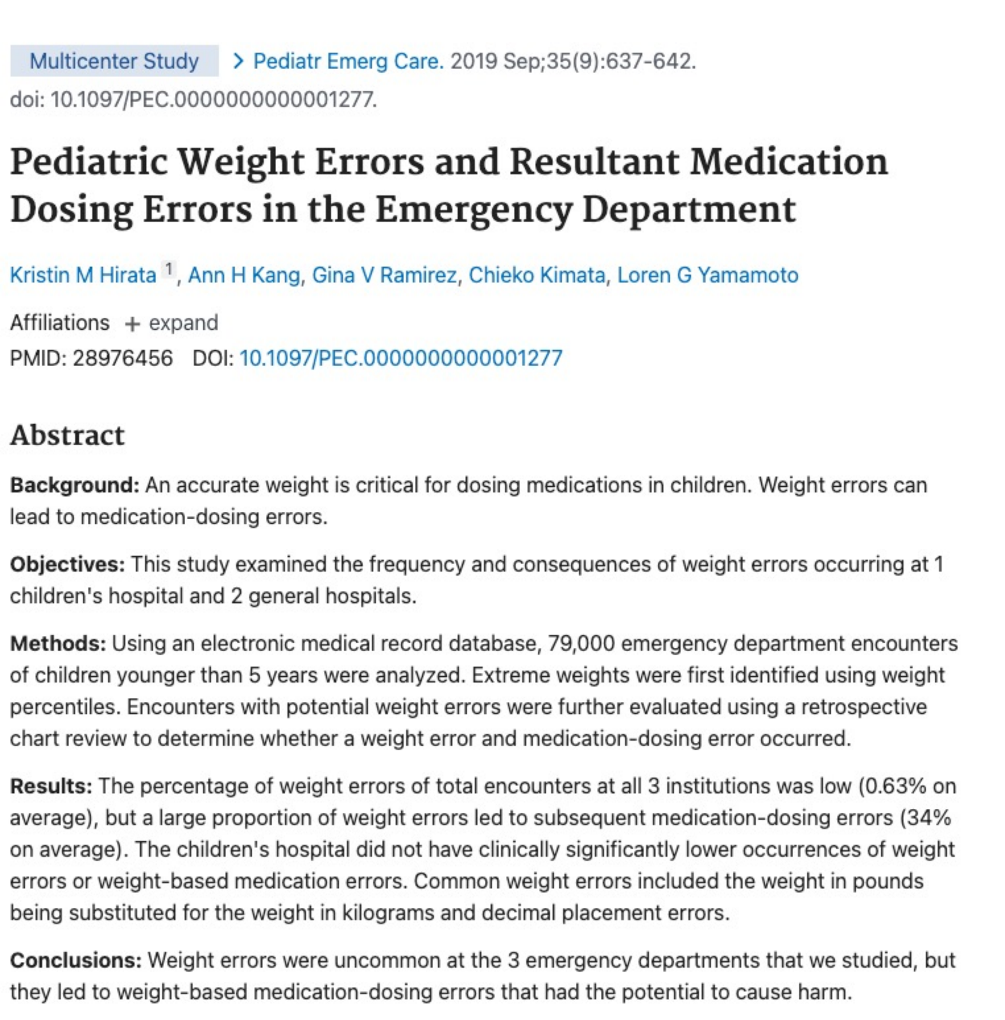
Data warehouse

Data lakes

Operational data stores

Data mesh

Bad data can have serious 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.

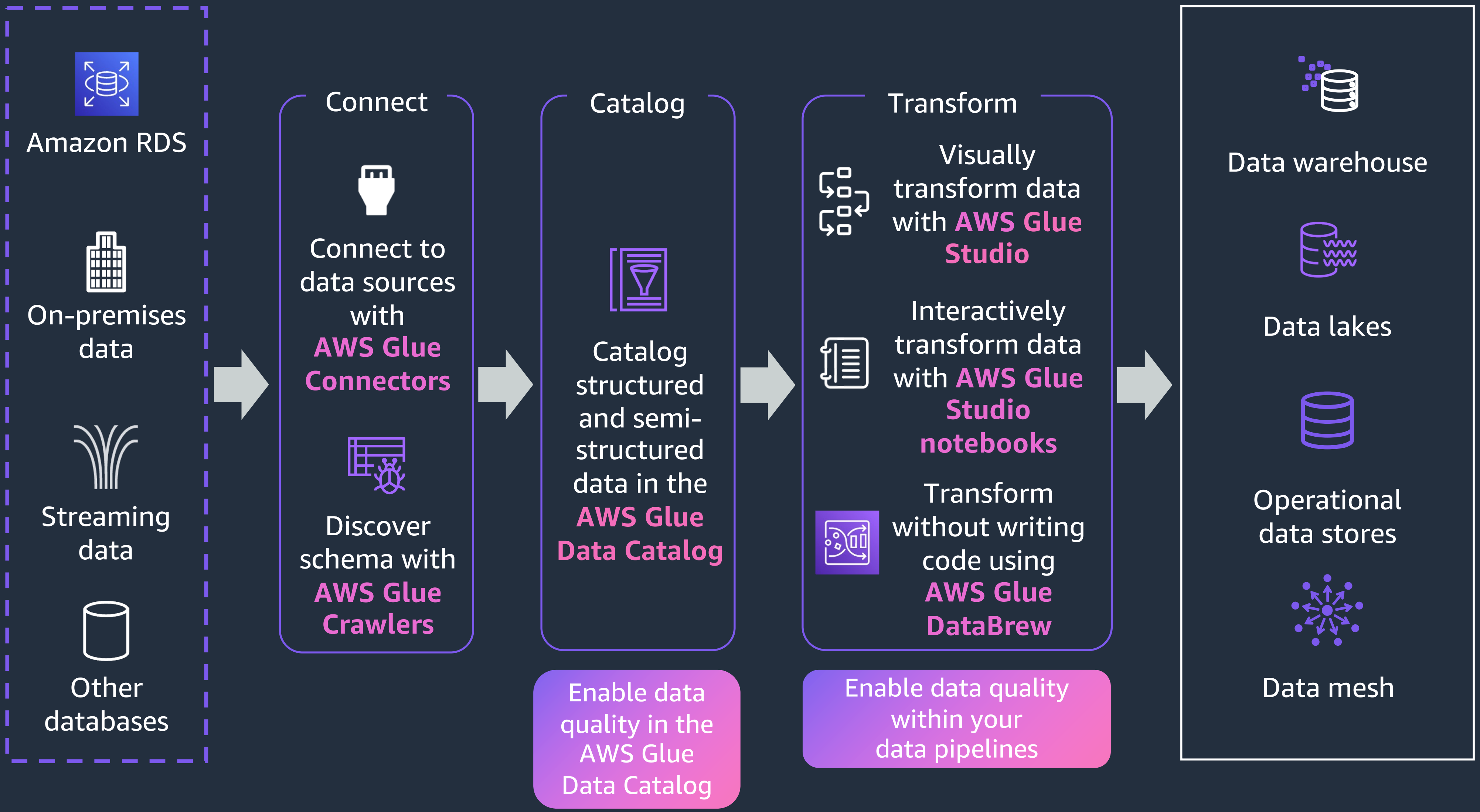
- Incorrect or absent recording of patient weights can lead to medication dosage errors
- Pediatric researchers Hirata and colleagues examined the frequency and consequences of weight errors that occurred across 79,000 emergency department encounters of children under the age of 5
- Although weight errors were relatively rare (0.63%), many weight errors led to subsequent medication dosing errors (34%)[1]

[1] Hirata, Kang, Kimata, Ramirez, and Yamamoto, NIH National Library of Medicine

AWS Glue Data Quality features and benefits

- Serverless, scalable, and high-performing
- Rule recommendations
- Built-in data quality rules and actions
- Data Quality Definition Language
- Data-at-rest and data-in-transit support
- Multiple-persona support

Data quality where you need it



Amazon RDS

On-premises data

Streaming data

Other databases

Connect

Connect to data sources with **AWS Glue Connectors**

Discover schema with **AWS Glue Crawlers**

Catalog

Catalog structured and semi-structured data in the **AWS Glue Data Catalog**

Enable data quality in the AWS Glue Data Catalog

Transform

Visually transform data with **AWS Glue Studio**

Interactively transform data with **AWS Glue Studio notebooks**

Transform without writing code using **AWS Glue DataBrew**

Enable data quality within your data pipelines

Data warehouse

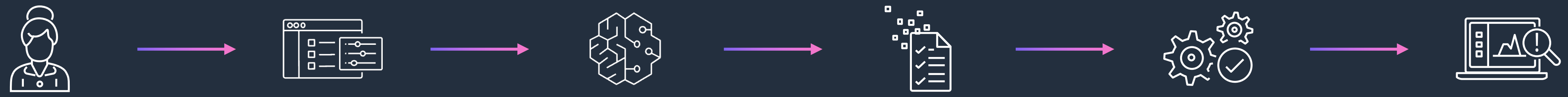
Data lakes

Operational data stores

Data mesh

AWS Glue Data Quality in the AWS Glue Data Catalog

HOW IT WORKS



Data steward

Selects a dataset in the AWS Glue Data Catalog

AWS Glue Data Quality analyzes data and recommends rules

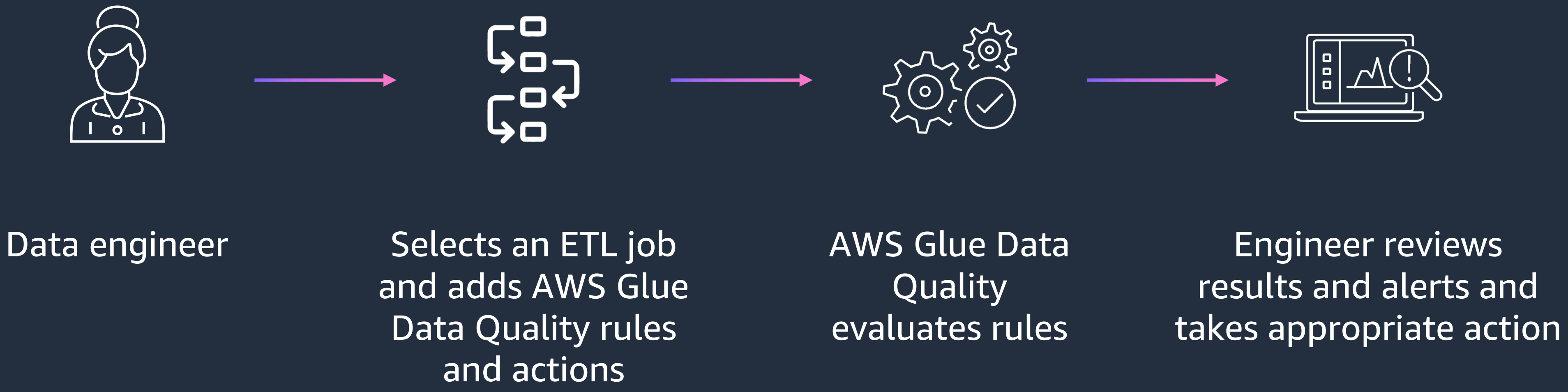
Data steward refines rules to create finalized rules

AWS Glue Data Quality evaluates rules

Data steward reviews results and alerts and takes appropriate action

AWS Glue Data Quality on AWS Glue ETL pipelines

HOW IT WORKS



Data engineer

Selects an ETL job and adds AWS Glue Data Quality rules and actions

AWS Glue Data Quality evaluates rules

Engineer reviews results and alerts and takes appropriate action

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-petabyte data lake

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

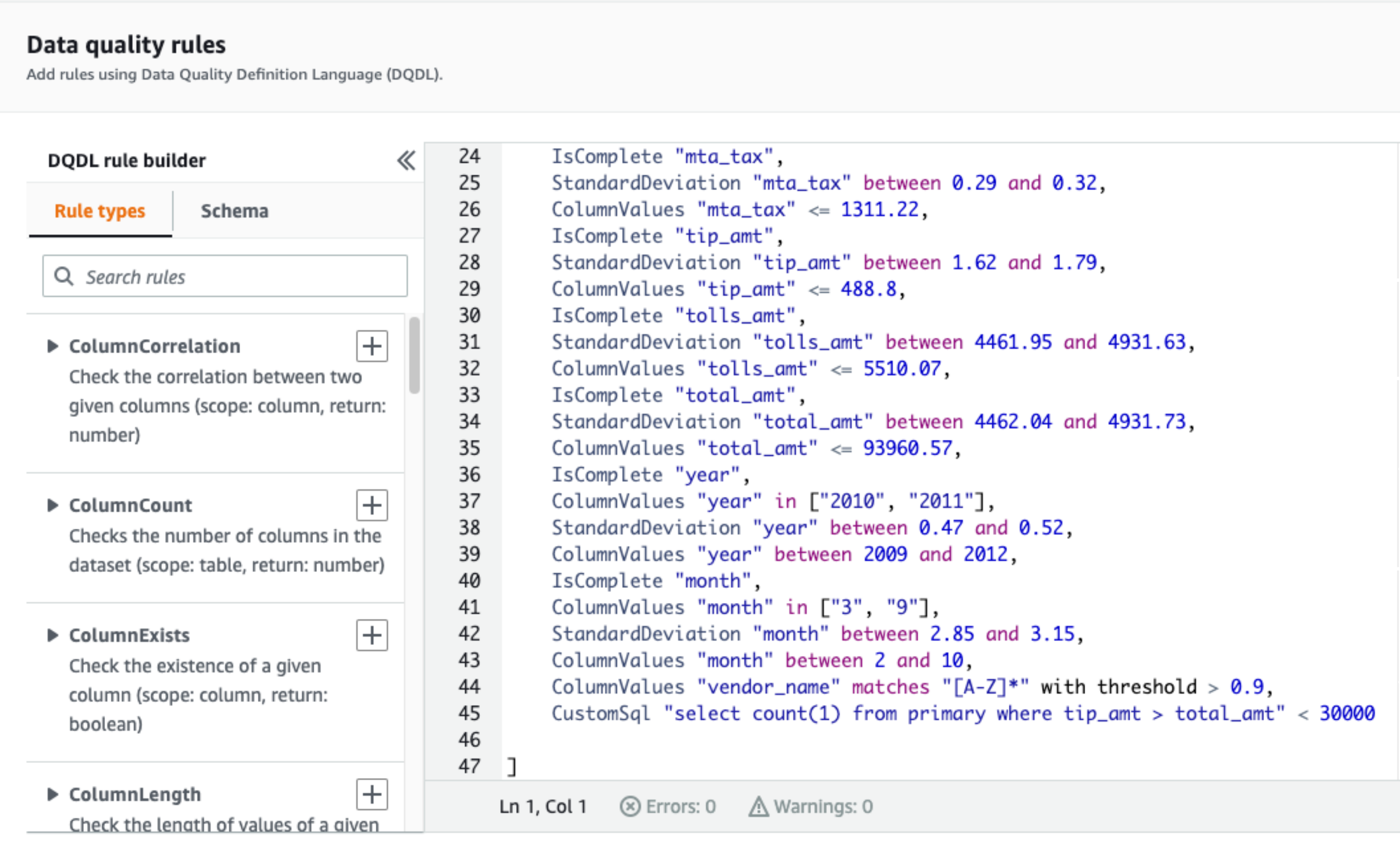
Many customers already use Deequ to manage data quality

Defining data quality rules can be difficult and complex

1. Developers must write complex code to implement data quality rules
2. Business users must read code to understand the rules
3. Migrating and reusing data quality rules can be difficult
4. Data quality rules are written in proprietary languages

Data Quality Definition Language (DQDL)

DECLARATIVE LANGUAGE TO AUTHOR AND REUSE DATA QUALITY RULES



Data quality rules

Add rules using Data Quality Definition Language (DQDL).

DQDL rule builder

▶ **ColumnCorrelation** ⊞
Check the correlation between two given columns (scope: column, return: number)

▶ **ColumnCount** ⊞
Checks the number of columns in the dataset (scope: table, return: number)

▶ **ColumnExists** ⊞
Check the existence of a given column (scope: column, return: boolean)

▶ **ColumnLength** ⊞
Check the length of values of a given

```
    IsComplete "mta_tax",
    StandardDeviation "mta_tax" between 0.29 and 0.32,
    ColumnValues "mta_tax" <= 1311.22,
    IsComplete "tip_amt",
    StandardDeviation "tip_amt" between 1.62 and 1.79,
    ColumnValues "tip_amt" <= 488.8,
    IsComplete "tolls_amt",
    StandardDeviation "tolls_amt" between 4461.95 and 4931.63,
    ColumnValues "tolls_amt" <= 5510.07,
    IsComplete "total_amt",
    StandardDeviation "total_amt" between 4462.04 and 4931.73,
    ColumnValues "total_amt" <= 93960.57,
    IsComplete "year",
    ColumnValues "year" in ["2010", "2011"],
    StandardDeviation "year" between 0.47 and 0.52,
    ColumnValues "year" between 2009 and 2012,
    IsComplete "month",
    ColumnValues "month" in ["3", "9"],
    StandardDeviation "month" between 2.85 and 3.15,
    ColumnValues "month" between 2 and 10,
    ColumnValues "vendor_name" matches "[A-Z]*" with threshold > 0.9,
    CustomSql "select count(1) from primary where tip_amt > total_amt" < 30000

]
```

Ln 1, Col 1 ⊗ Errors: 0 ⚠ Warnings: 0

→ Easier to author

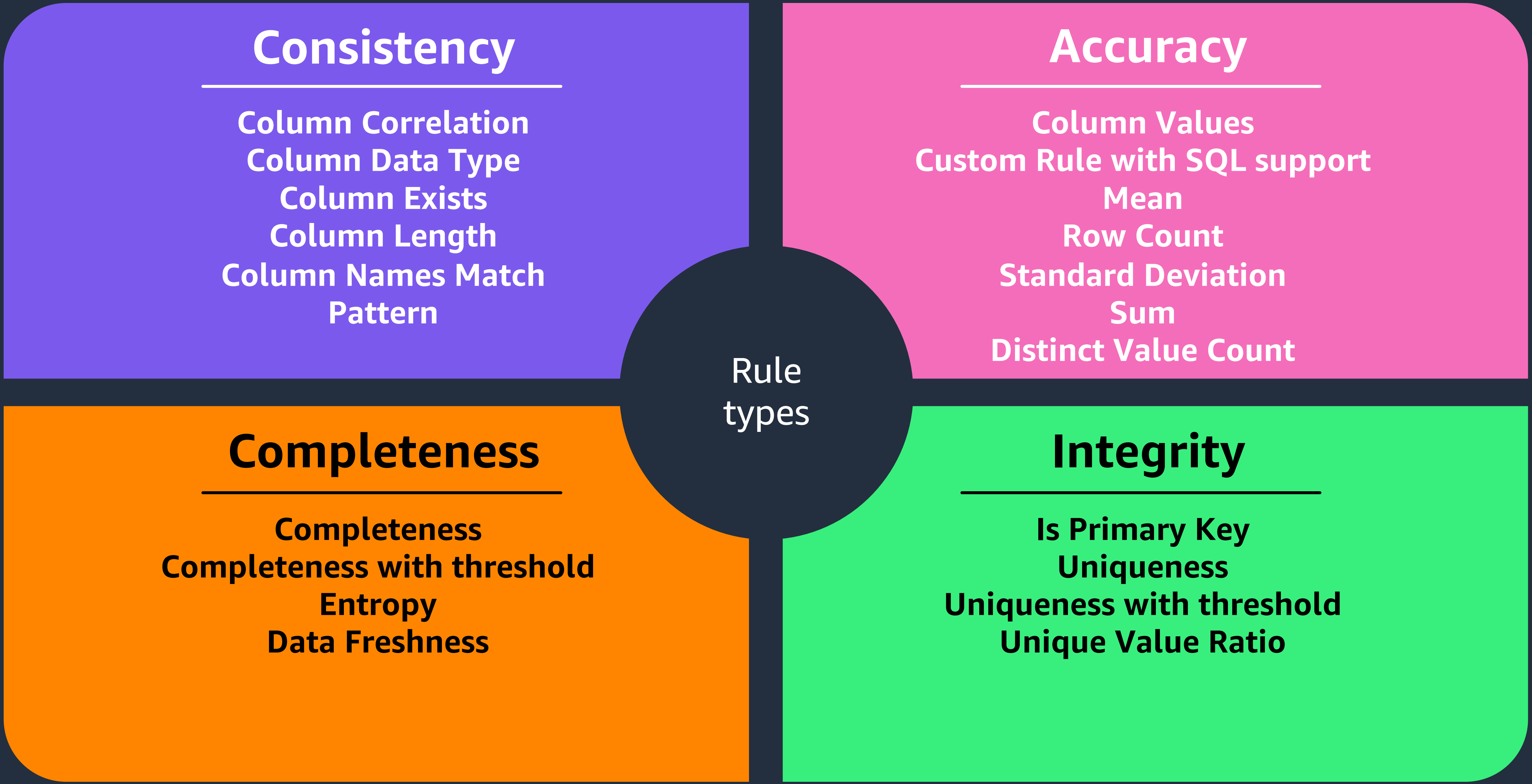
→ Intuitive

→ Complex rule support

→ Reusable and easier to deploy

→ Author SQL-based rules

Built-in rule types



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

Completeness

Completeness
Completeness with threshold
Entropy
Data Freshness

Integrity

Is Primary Key
Uniqueness
Uniqueness with threshold
Unique Value Ratio

Vyaire increases efficiency with AWS Glue Data Quality

> “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. 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.”

Gopal Ramamurthi

Vice President of Global Data Management and Digital Health, Vyaire Medical

vyaire™ MEDICAL

United Airlines to use AWS Glue Data Quality to reduce manual effort

> "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**."

Ashok Srinivas and Sarang Bapat

Director of ML Engineering and Director of Data Engineering, United Airlines

UNITED AIRLINES

RX Global simplifies data quality with AWS Glue

> “The time and complexity associated with installing open-source packages has made it very challenging to implement data quality. 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.”

Gaurav Singla

Architect, RX Global

RX

Get started with AWS Glue Data Quality

Review documentation

Access it in the AWS Glue console

Thank you!