

IBM Knowledge Catalog – Data Quality

Vishwas Balakrishna

Product Manager, Metadata Import | Metadata Enrichment | Data Quality | Workflows



Agenda

1. What is data quality?

- 1.1 Definition & perspectives
- 1.2 Challenges & risks
- 1.3 Key market trends
- 1.4 Business centric metrics

2. Data quality on Cloud Pak for Data

- 2.1 Objectives
- 2.2 Product vision & strategy
- 2.3 Unified data quality solution
- 2.4 Entry points

3. Roadmap

- 3.1 Data quality timeline

01

What is data quality?

Definition

Data quality solutions are the set of processes and technologies for **identifying, understanding, preventing, escalating and correcting issues** in data that supports effective decision making and governance across all business processes”
- Gartner DQ report

Data quality in governance is required to accelerate insights and address regulatory compliance.

BUSINESS NEEDS

Organizations struggle to deliver timely, trusted, quality data for business insights.



Data is useful only if its context, content and quality is trusted and continuously evaluated.

REGULATORY COMPLIANCE

Organizations struggle to keep pace with data privacy and industry regulations.



To ensure data is regulatory compliant, it is critical to address quality issues.

Challenges & risks

Data quality is one fundamental challenge inhibiting enterprises from becoming data-driven

Source: [Capgemini Research Institute](#)

45%

Data access

Multiple locations, clouds, applications and data silos hinder timely delivery of right data to the right users.

Only **45%** of enterprise data useful for analysis was analyzed or fed into AI.¹

80%

Data quality

Poor data quality and consistency lowers trust in data.

80% of business executives surveyed do not trust their data.²

26%

Data literacy

Lack of a consistent understanding of data across the organization can hinder data utilization.

26% of survey respondents cite expanding companywide data literacy as a high or critical priority.³

18%

Data protection

Inadequate data protection increases the risk of non-compliance.

Only **18%** of survey respondents report that their firm excelled at protecting organizational data.⁴

“Bad data quality costs organizations an average of \$12.9 million per year.”

Gartner report *2021*

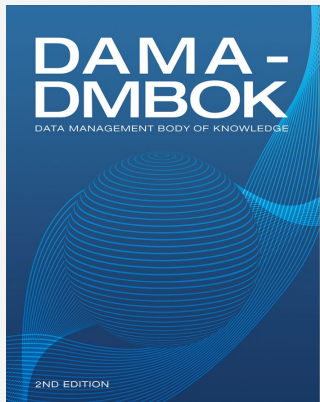
Managing data quality consumes **up to 20% of a data management team's time***. Often this is time spent on generic and unfocused data quality improvement efforts.



Only by **identifying critical data elements** and their associated service level objectives can organizations ensure efforts are spent on the right topics, such as data feeding regulatory reporting requirements.

Measuring data quality

Data quality dimensions for business centric metrics



The Data Management Association (DAMA) International published a paper that describes **6 core dimensions** of data quality.

Accuracy

Data values are as close as possible to real values.

Consistency

Data values within a column comply with a rule.

Uniqueness

Distinct values appear only once within a column of data.

Completeness

All required data values are present in the data.

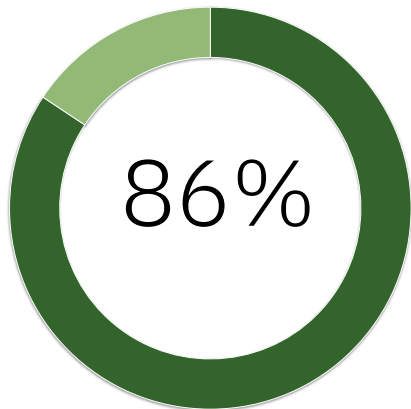
Timeliness

Data represent the reality from a required point in time.

Validity

Data conforms to the format, type, or range of its definition.

Customers are looking for detailed data quality info



Finding

86% of all data consumer and data provider want to see data quality on individual column level in their dataset or table.

Top attributes and metrics

Consistency 90%

The number of inconsistencies, the percentage of data that is the same across multiple systems.

Accuracy 88%

The ratio of data to errors. incorrect or unlikely values (Misspellings, Incorrect Numbers, etc.).

Completeness 86%

Missing or incomplete data, potential missing records within a master data entity.

Insights

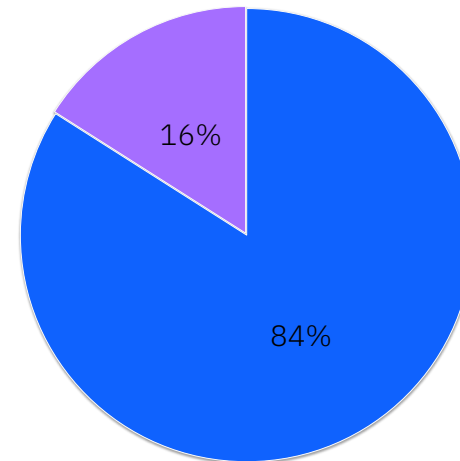
Customers are looking for a holistic and collaborative view for all stakeholders

“Holistic approach to data management. Working in a collaboration to identify data issues and come up with remediations or improvement initiatives as a team.”

Master Data Governance consultant

“The Data Quality metrics across business focused Data Consumers and more technical focused Data Providers should be combined in a single tool or platform.”

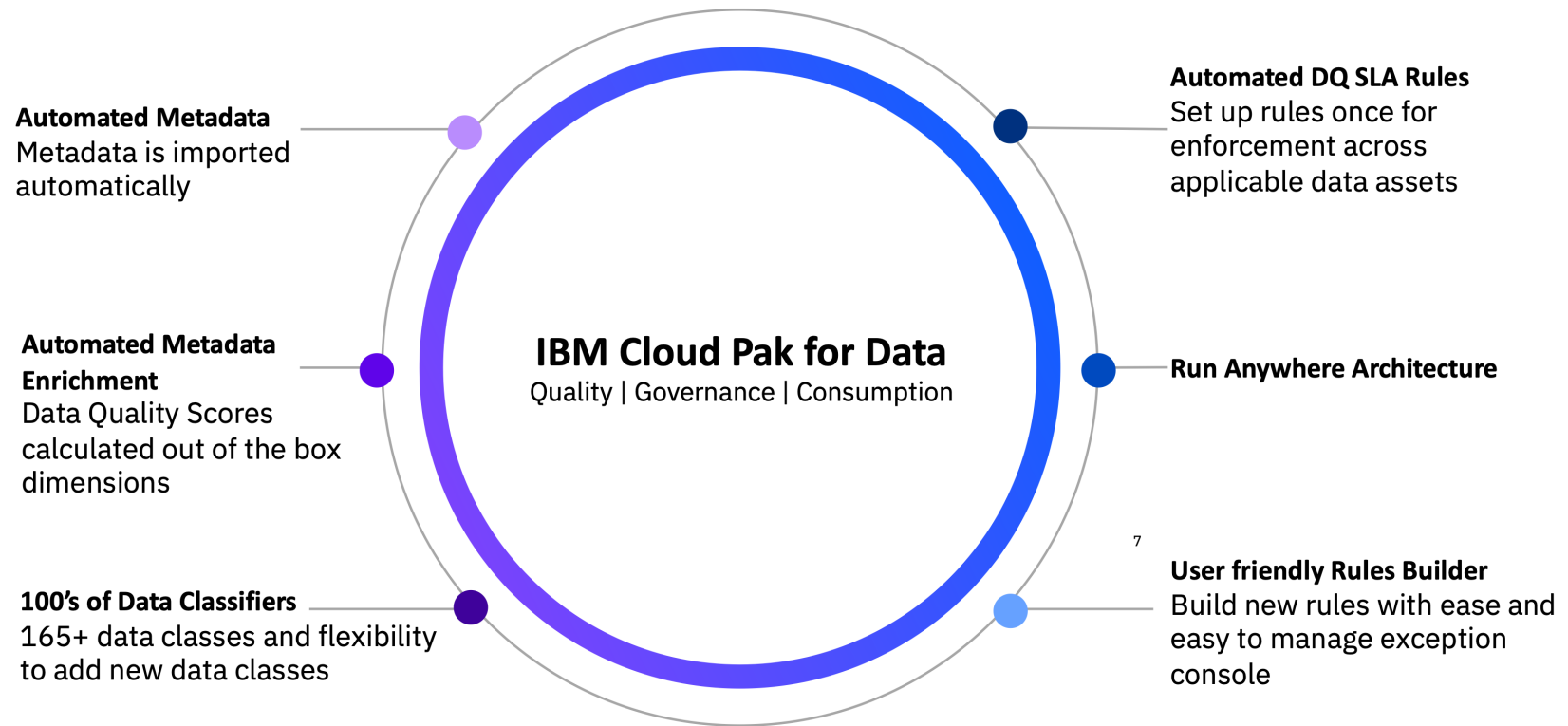
■ Agree ■ Neutral



02

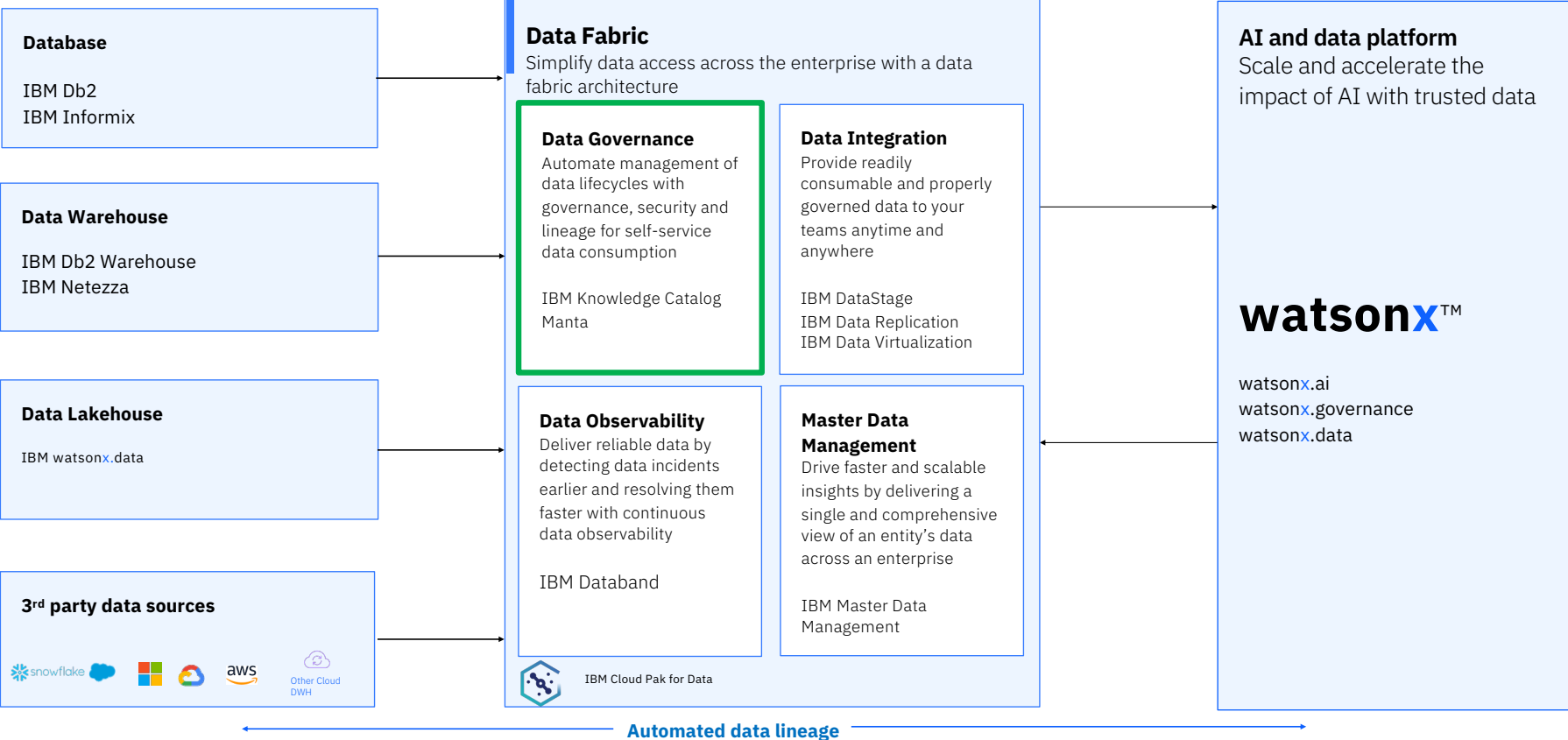
Data quality on CPD

Why IBM ? Key Differentiators



Product vision & strategy

Investments in a **Data Fabric** will accelerate and scale organizations' **AI** initiatives

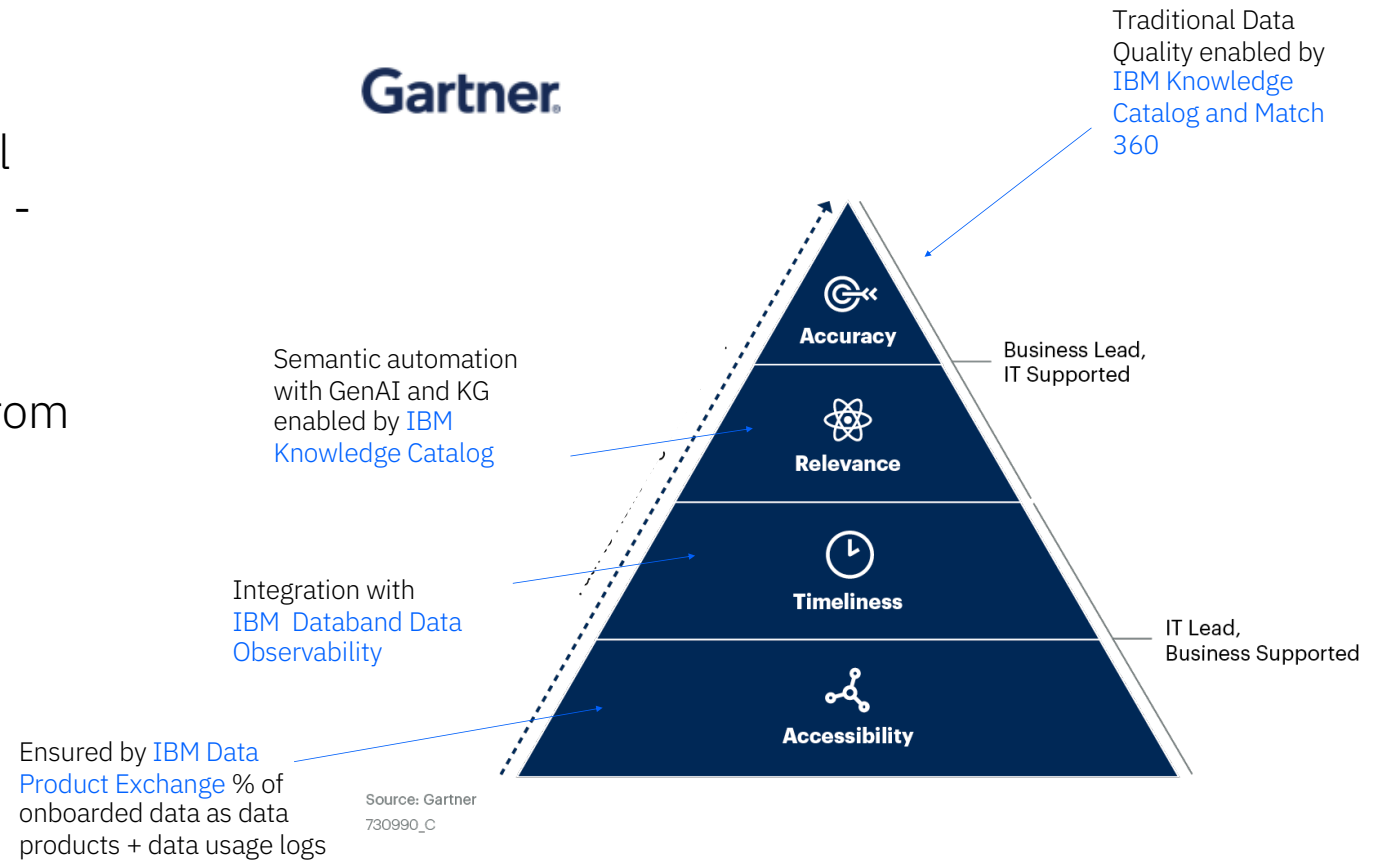


Gain deeper visibility into your data and its journey from source to end-use for regulatory compliance and AI use cases with Manta, an IBM company

IBM Data Quality capabilities for Data Fabric

As reported by Gartner, Data Fabric goes beyond traditional focus of data quality solutions - Data Quality is more than just Accuracy.

Effective Data Fabric design from IBM delivers on 4 core quality dimensions by design.



Unified data quality approach

Unified data quality approach as an [automated process](#)



Unified data quality approach as an [automated process](#)



UNDERSTAND DATA

Identify & classify critical data elements

Prepare governance artifacts and auto-assign business terms and classifications to data in Metadata enrichment.

Unified data quality approach as an [automated process](#)



UNDERSTAND DATA

Identify & classify critical data elements

Prepare governance artifacts and auto-assign business terms and classifications to data in Metadata enrichment.



ANALYSE DATA

Run data quality checks on data

Allow all source agnostic analysis on data assets to get data quality scores based on dimensions.



Unified data quality approach as an [automated process](#)



UNDERSTAND DATA

Identify & classify critical data elements

Prepare governance artifacts and auto-assign business terms and classifications to data in Metadata enrichment.



ANALYSE DATA

Run data quality checks on data

Allow all source agnostic analysis on data assets to get data quality scores based on dimensions.



DEFINE QUALITY EXPECTATIONS

Data quality governance with SLA rules

Define thresholds as acceptable criteria and decide what automated action to take in case of violations.



Unified data quality approach as an [automated process](#)



UNDERSTAND DATA

Identify & classify critical data elements

Prepare governance artifacts and auto-assign business terms and classifications to data in Metadata enrichment.



ANALYSE DATA

Run data quality checks on data

Allow all source agnostic analysis on data assets to get data quality scores based on dimensions.



DEFINE QUALITY EXPECTATIONS

Data quality governance with SLA rules

Define thresholds as acceptable criteria and decide what automated action to take in case of violations.



ONGOING MONITORING

Monitor data quality & SLA compliance

Review data quality scores and SLA compliance. Investigate cause of issues and see failed records.

Unified data quality approach as an [automated process](#)



UNDERSTAND DATA

Identify & classify critical data elements

Prepare governance artifacts and auto-assign business terms and classifications to data in Metadata enrichment.



ANALYSE DATA

Run data quality checks on data

Allow all source agnostic analysis on data assets to get data quality scores based on dimensions.



DEFINE QUALITY EXPECTATIONS

Data quality governance with SLA rules

Define thresholds as acceptable criteria and decide what automated action to take in case of violations.



ONGOING MONITORING

Monitor data quality & SLA compliance

Review data quality scores and SLA compliance. Investigate cause of issues and see failed records.

Unified data quality approach as an [automated process](#)



UNDERSTAND DATA

Identify & classify critical data elements

Prepare governance artifacts and auto-assign business terms and classifications to data in Metadata enrichment.



ANALYSE DATA

Run data quality checks on data

Allow all source agnostic analysis on data assets to get data quality scores based on dimensions.



DEFINE QUALITY EXPECTATIONS

Data quality governance with SLA rules

Define thresholds as acceptable criteria and decide what automated action to take in case of violations.



ONGOING MONITORING

Monitor data quality & SLA compliance

Understand data quality and review SLA compliance. Investigate issues. Subscribe to data quality events.

UNDERSTAND DATA

Identify & classify critical data elements

The platform experience combines a comprehensive range of capabilities. With its open ecosystem, users can perform metadata enrichment across any source.

New asset

Create a metadata enrichment

- Details
- Data scope
- Objective**
- Schedule
Optional
- Review

Enrichment objective

Profile data <input checked="" type="checkbox"/> Provides basic statistics about the asset content, assigns and suggests data classes, and suggests primary keys	Run basic quality analysis <input checked="" type="checkbox"/> Run predefined data quality checks to assess the general quality of your data Output: - Customize	Assign terms <input checked="" type="checkbox"/> Assigns and suggests business terms to tables and columns	Set relationships <input checked="" type="checkbox"/> Use profiling statistics and name similarities to provide primary and foreign keys and suggest relationships between assets and columns
--	--	--	---

Categories [ⓘ]

Selected categories Remove all

- Knowledge Accelerators
- Data Classes
Knowledge Accelerators /
- Common Data Classes
Knowledge Accelerators /Data Classes /
- Data Type Data Classes**
Knowledge Accelerators /Data Classes / ⊖
- Demographic Data Classes
Knowledge Accelerators /Data Classes /

Items per page: 5 ▾ 1-5 of 41 items 1 ▾ of 9 pages ◀ ▶

Sampling [ⓘ]

Basic <input checked="" type="radio"/> Minimum sample size to optimize for speed Analyze: 1,000 rows per table Classify: based on most frequent 100 values per column	Moderate Serves as a trade-off between speed and accuracy Analyze: 10,000 rows per table Classify: based on most frequent 100 values per column	Comprehensive Large sample size to optimize for accuracy Analyze: 100,000 rows per table Classify: all values per column	Custom Use customized sampling Sampling method: — Analyze: — Classify: — Customize
---	---	--	--

Unified data quality approach as an [automated process](#)



UNDERSTAND DATA

Identify & classify critical data elements

Prepare governance artifacts and auto-assign business terms and classifications to data in Metadata enrichment.



ANALYSE DATA

Run data quality checks on data

Allow all source agnostic analysis on data assets to get data quality scores based on dimensions.



DEFINE QUALITY EXPECTATIONS

Data quality governance with SLA rules

Define thresholds as acceptable criteria and decide what automated action to take in case of violations.



ONGOING MONITORING

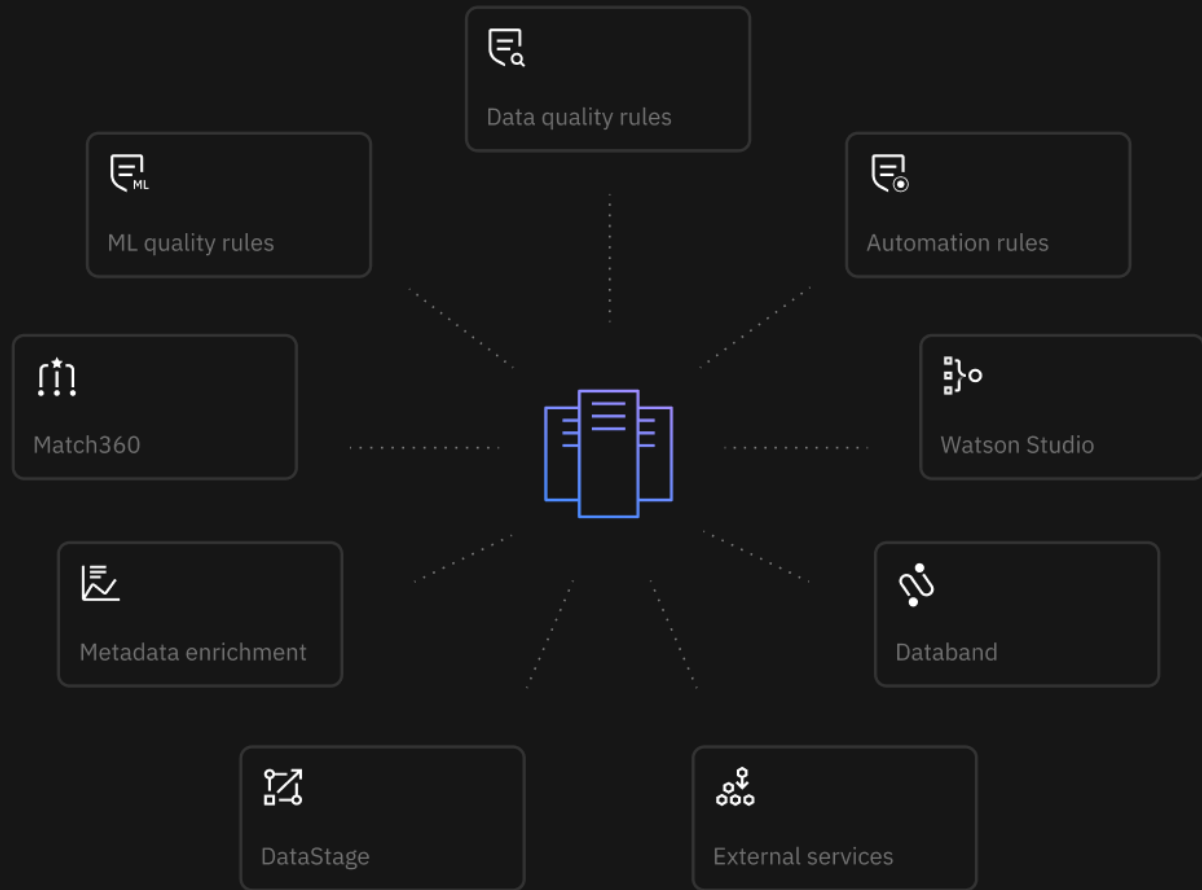
Monitor data quality & SLA compliance

Understand data quality and review SLA compliance. Investigate issues. Subscribe to data quality events.

ANALYSE DATA

Run platform wide data quality checks

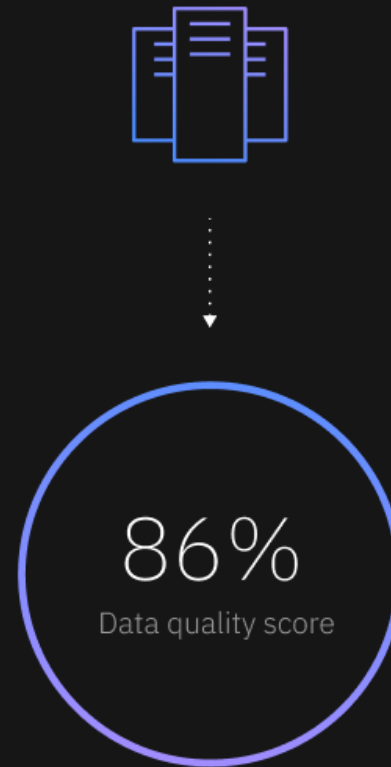
The platform experience combines a comprehensive range of capabilities. With its open ecosystem, users can perform quality analysis across any source.



ANALYSE DATA

Store data quality results centralised

Once the data quality score and results are measured, they're stored in a holistic architecture that's accessible to users across the platform.



ANALYSE DATA

Business centric data quality dimensions

IBM Cloud Pak for Data addresses a range of business requirements with out-of-the-box quality dimensions as DAMA industry standards.

This enables data teams to see a breakdown by dimension to conduct custom evaluations based on their unique business priorities.



ANALYSE DATA

Example: Data quality rule

Check for this criteria:

Column **Account number** in
data asset “**Customer_data**” to
have only **unique values**.

The screenshot displays the IBM Cloud Pak for Data interface. At the top, the header includes the IBM logo, the text "IBM Cloud Pak for Data", a search bar with the placeholder "Search for a resource or offering", and the user name "Jimmy Jimmereeno". Below the header, the breadcrumb "Projects / Credit card customers" is visible. The main navigation bar contains tabs for "Overview", "Assets", "Jobs", and "Manage", with "Assets" currently selected. A search bar labeled "Find assets" is positioned above the asset list. On the right side of this bar, there are buttons for "Import assets" and "New asset".

On the left side, a sidebar titled "9 assets" shows a tree view of "Asset types":

- Data access (2)
- Data (3)**
- Data quality (2)
- Flows (2)
 - DataStage flows (2)

The main content area displays a table of data assets:

Data		
Name		Last modified
BANK_CLIENTS application/octet-stream		1 month ago Modified by you
Customer_data Connection		1 month ago Modified by you

ANALYSE DATA

Example: Data quality rule

Check for this criteria:

Column **Account number** in
data asset “**Customer_data**” to
have only **unique values**.

IBM Cloud Pak for Data

Search for a resource or offering

Jimmy Jimmereeno

Projects / Credit card customers

New asset

Select a tool based on what type of asset you want and how you want to work.

Find tools by name or description

that are masked by advanced data protection rules.

training models, and creating deployments.

data and build and train a model, using a guided approach to machine learning that doesn't require coding.

Code editors

- Code package**
Organize and run a set of code and dependent files together.
- Federated Learning**
Create a federated learning experiment to train a common model on a set of remote data sources. Share training results without sharing data.
- Jupyter notebook editor**
Create a notebook in which you run Python, R, or Scala code to prepare, visualize, and analyze data, or build a model.
- Python function**
Python functions that can be deployed to a space.

Component editors

- Data quality definition**
Create abstract rule logic for data analysis that can be used in any number of data quality rules.
- Data quality rule**
Create rules to assess the quality of your data by evaluating and validating specific conditions.
- DataStage component**
Create reusable components for DataStage flows, such as subflows, libraries, and data definitions.
- Parameter set**
Collect multiple job parameters with specified values to reuse in jobs.

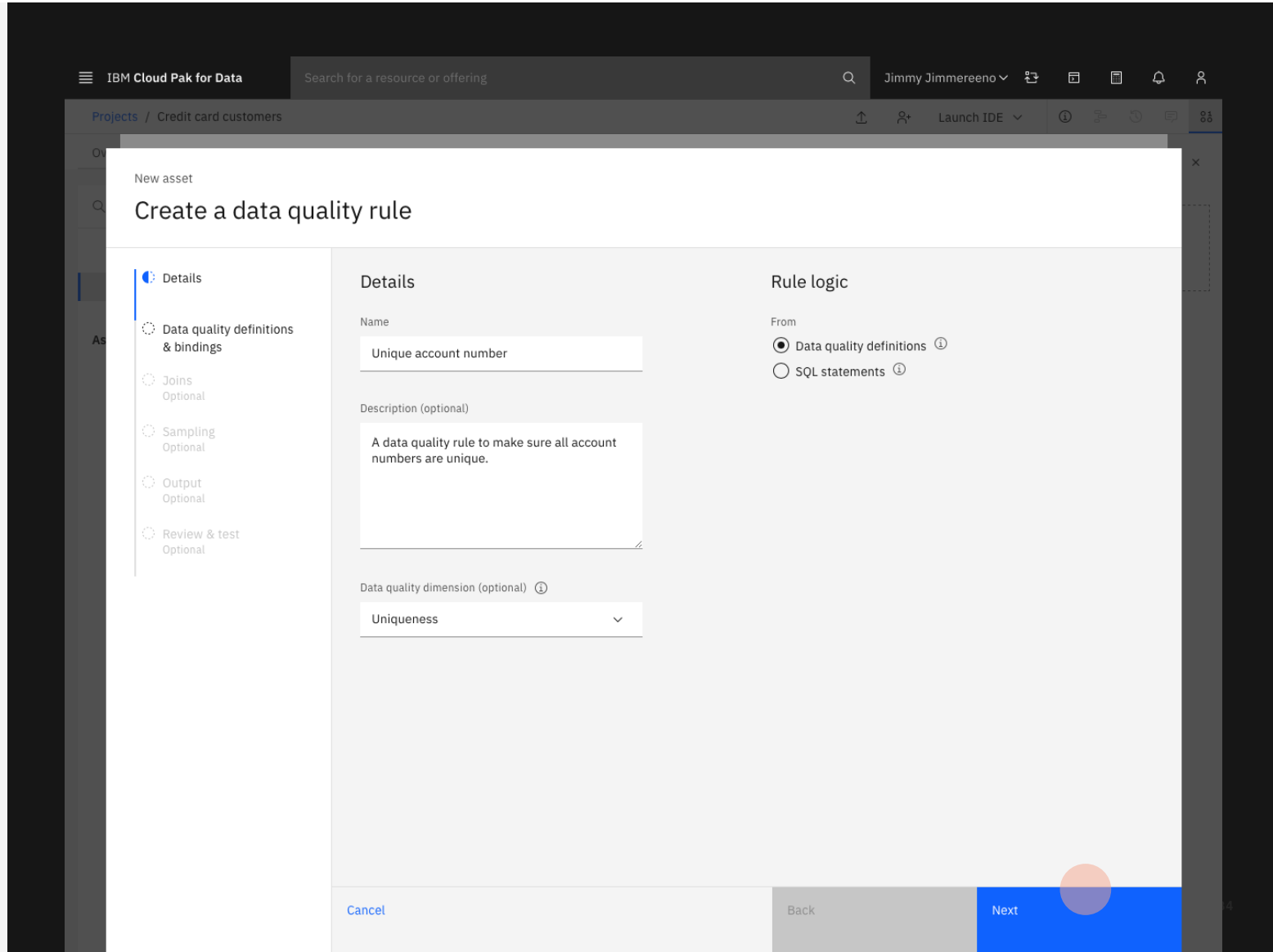
Show descriptions ⓘ

ANALYSE DATA

Example: Data quality rule

Check for this criteria:

Column **Account number** in data asset **“Customer_data”** to have only **unique values**.



ANALYSE DATA

Example: Data quality rule

Check for this criteria:

Column **Account number** in data asset **"Customer_data"** to have only **unique values**.

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there's a search bar and user information for Jimmy Jimmereeno. The main content area is titled 'New asset' and 'Create a data quality rule'. On the left, there's a navigation menu with options: Details (selected), Data quality definitions & bindings (active), Joins (Optional), Sampling (Optional), Output (Optional), and Review & test (Optional). The main panel is titled 'Data quality definitions & bindings' and shows a rule named 'UniqueAccountNum'. Below the name, there's a 'Rule expression with variables' section containing 'Account number EXISTS AND UNIQUE'. A table below indicates '1 / 1 variables bound'.

Name	Binding type	Bind to	Complete
AccountNum Number	Column ▾	Account number Customer data	✓

At the bottom, there are three buttons: 'Cancel', 'Back', and 'Next'.

ANALYSE DATA

Example: Data quality rule

Check for this criteria:

Column **Account number** in data asset "Customer_data" to have only **unique values**.

The screenshot displays the IBM Cloud Pak for Data interface for configuring a data quality rule. The page title is "Project / Credit card customers" and the rule name is "Unique account number". The rule is currently in "Overview" mode, with a "Run history" tab also available. The "Run rule" button is highlighted with a red circle.

General

- Description:** A data quality rule to make sure all account numbers are unique.
- Data quality dimension:** Uniqueness
- Data quality definitions:** UniqueAccountNum
- Bound expression:** Account number EXISTS AND UNIQUE

Governance artifacts

- Business terms:** Customer bank accounts
- Governance rule:** Operations rule

Selected output

Name	Source	Data quality definition	Binding	Type
Data quality definition	-	-	-	Metric
first_name	name	ValidName	RANK_CLIENTS.first_name	Column

ANALYSE DATA

Example: Data quality rule

Check for this criteria:

Column **Account number** in data asset "Customer_data" to have only **unique values**.

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there is a navigation bar with the text 'IBM Cloud Pak for Data' and a search bar. Below the navigation bar, the breadcrumb 'Project / Credit card customers' is visible. The main content area is titled 'Unique account number' and is categorized as a 'Data quality rule'. There are two tabs: 'Overview' (selected) and 'Run history'. A notification box in the top right corner states 'Rule running: The data quality rule started running. View the status on the Run history tab.' The configuration is organized into sections: 'General', 'Governance artifacts', and 'Selected output'. The 'General' section includes 'Description', 'Data quality dimension', 'Data quality definitions', and 'Bound expression'. The 'Governance artifacts' section includes 'Business terms' and 'Governance rule'. The 'Selected output' section is a table with columns for Name, Source, Data quality definition, Binding, and Type.

Name	Source	Data quality definition	Binding	Type
Data quality definition	-	-	-	Metric
first_name	name	ValidName	RANK_CLIENTS.first_name	Column

ANALYSE DATA

Example: Data quality rule

Check for this criteria:

Column **Account number** in data asset "Customer_data" to have only **unique values**.

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there is a navigation bar with the text "IBM Cloud Pak for Data" and a search bar. The user's name "Jimmy Jimmereeno" is visible in the top right corner. Below the navigation bar, the breadcrumb "Project / Credit card customers" is shown. The main content area is titled "Unique account number" and is identified as a "Data quality rule". There are two tabs: "Overview" and "Run history", with "Run history" being the active tab. A "View run history" section is visible, followed by a table of run history entries. The table has columns for "Start time", "Job name", "Status", "Records tested", "Rule met", and "Rule not met". One entry is shown for "Nov 13, 2023 at 10:45 PM" with a "Success" status, 941 records tested, 509 (58.2%) rule met, and 432 (41.8%) rule not met. A "View output table" button is located in the top right of the table area.

Start time	Job name	Status	Records tested	Rule met	Rule not met
Nov 13, 2023 at 10:45 PM	Unique account number job	Success	941	509 (58.2%)	432 (41.8%)

ANALYSE DATA

Example:
Data quality rule

Check for this criteria:

Column **Account number** in data asset “**Customer_data**” to have only **unique values**.

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there is a search bar and the user name 'Jimmy Jimmereeno'. Below the search bar, the breadcrumb 'Project / Credit card customers' is visible. The main content area displays a window titled 'Output table of Unique account number'. Inside this window, there is a section 'Records that do not meet rule conditions' with a search input 'Find records' and a 'View output table asset' button. The table below lists records that fail the uniqueness check for the 'Account number' column.

Account number	UserID	Countryofoperation	System date
101579	21678	Australia	Mon 6 Sep 2019
101579	41234	Turkey	Mon 6 Sep 2019
101579	22345	Chile	Mon 6 Sep 2019
101579	12345	Republic of peru	Mon 6 Sep 2019
101579	90876	Egypt	Mon 6 Sep 2019
101579	21678	Australia	Mon 6 Sep 2019
101579	41234	Turkey	Mon 6 Sep 2019
101579	22345	Chile	Mon 6 Sep 2019
101579	12345	Republic of peru	Mon 6 Sep 2019
101579	90876	Egypt	Mon 6 Sep 2019
101579	21678	Australia	Mon 6 Sep 2019
101579	41234	Turkey	Mon 6 Sep 2019
101579	22345	Chile	Mon 6 Sep 2019
101579	12345	Republic of peru	Mon 6 Sep 2019
101579	90876	Egypt	Mon 6 Sep 2019
101579	21678	Australia	Mon 6 Sep 2019
101579	41234	Turkey	Mon 6 Sep 2019
101579	22345	Chile	Mon 6 Sep 2019
101579	12345	Republic of peru	Mon 6 Sep 2019
101579	90876	Egypt	Mon 6 Sep 2019
101579	90876	Egypt	Mon 6 Sep 2019
101579	21678	Australia	Mon 6 Sep 2019
101579	41234	Turkey	Mon 6 Sep 2019

Unified data quality approach as an [automated process](#)



UNDERSTAND DATA

Identify & classify critical data elements

Prepare governance artifacts and auto-assign business terms and classifications to data in Metadata enrichment.



ANALYSE DATA

Run data quality checks on data

Create and run source agnostic analysis to get data quality scores based on dimensions.



DEFINE QUALITY EXPECTATIONS

Data quality governance with SLA rules

Define thresholds as acceptable criteria and decide what automated action to take in case of violations.



ONGOING MONITORING

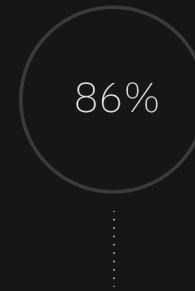
Monitor data quality & SLA compliance

Review data quality scores based on dimensions and SLA compliance. Investigate cause of issues and see failed records.

DEFINE DATA EXPECTATIONS

Data quality
governance with SLA
rules

Service Level Agreements are introduced to give meaningful and business-centric reflection of your data's quality.



DEFINE DATA EXPECTATIONS

Data quality
governance with SLA
rules

Service Level Agreements are introduced to give meaningful and business-centric reflection of your data's quality.



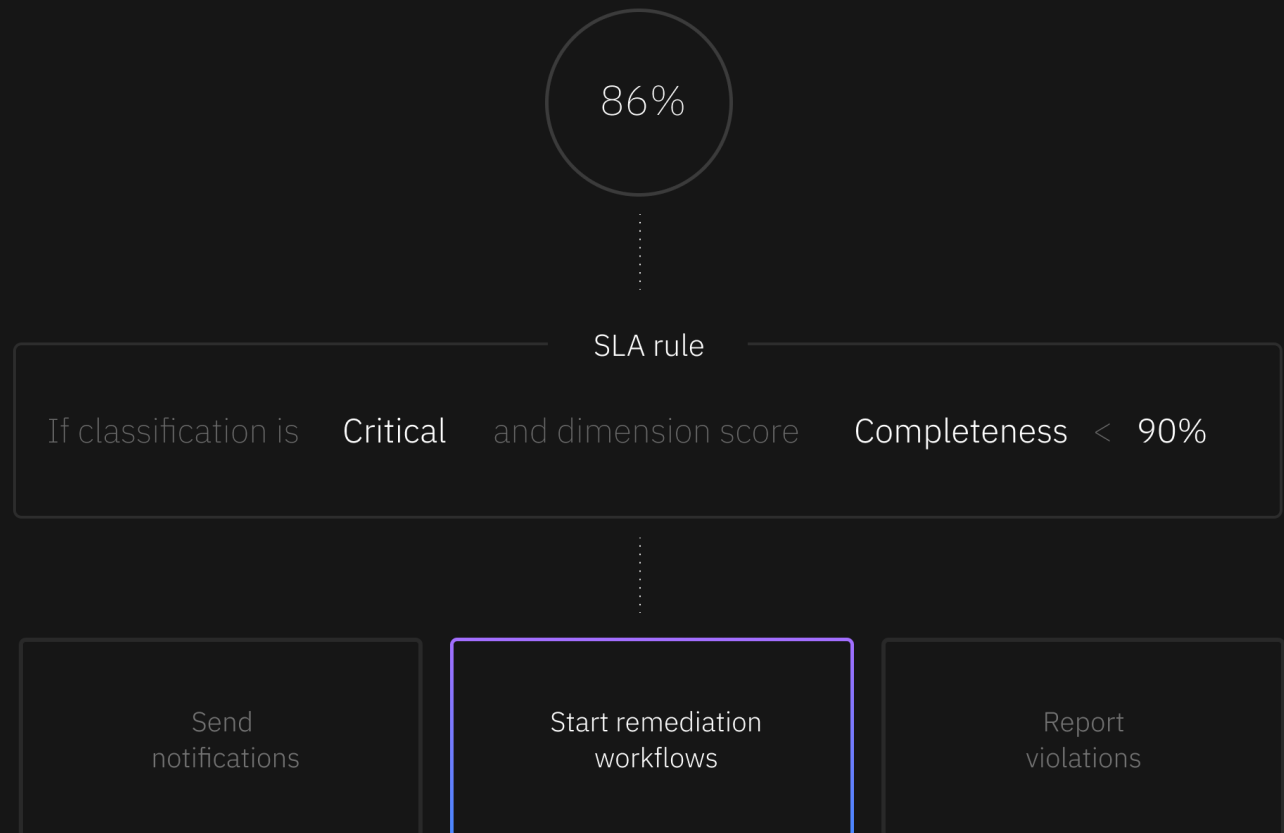
SLA rule

If classification is **Critical** and dimension score **Completeness** < 90%

DEFINE DATA EXPECTATIONS

Automated actions on violations

When SLA rule thresholds are trespassed, the system automatically triggers the best action to ensure data remains compliant with your regulatory expectations.



DEFINE DATA EXPECTATIONS

Data quality SLA rule example

Service Level Agreements are introduced to give meaningful and business-centric reflection of your data's quality.

The screenshot displays the IBM Cloud Pak for Data interface. The top navigation bar includes a search bar and a user profile icon labeled 'MB'. A left-hand navigation menu is visible, with the 'Rules' section highlighted. The main content area shows a table of rules with columns for 'Sort by', 'Name', 'Show', and 'All rule types'. A 'New rule' button is located in the top right corner of the table area.

Sort by:	Name	Show:	All rule types	Edit
	This is a rule to protect confidentially classified data		Data protection rule Last modified: Jan 24, 2023, 1:59 PM by admin	
	Obfuscate Passport Number, as national identification numbers must be protected to safeguard the rights and freedoms of the data subject.		Data protection rule Last modified: Nov 14, 2022, 5:52 AM by admin	
	Redact Social Security Number, as national identification numbers must be protected to safeguard the rights and freedoms of the data subject.		Data protection rule Last modified: Nov 14, 2022, 5:54 AM by admin	
	Vehicle Identification Number information may be used to identify an individual, therefore needs to be treated as protected data.		Data protection rule Last modified: Nov 14, 2022, 5:56 AM by admin	
rised by the	Where point (a) of Article 6(1) applies, in relation to the offer of information society services directly to a child, verify in such cases that consent is giv...	Knowledge Accelerator for...	Governance rule Last modified: Nov 14, 2022, 5:59 PM by admin	

DEFINE DATA EXPECTATIONS

Data quality SLA rule example

Service Level Agreements are introduced to give meaningful and business-centric reflection of your data's quality.

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there is a navigation bar with the text "IBM Cloud Pak for Data" and a search bar. Below this, the main content area is titled "Rules". On the right side of the "Rules" section, there is a blue button labeled "New rule".

The main content area displays a list of rules. The first rule is "Data quality SLA rule", which is highlighted with a red circle around the word "quality". The description for this rule is "A rule to monitor the data quality of critical data elements". Other rules listed include "Confidential info protection", "Mask Passport Number", "Mask Social Security Number", "Mask Vehicle Identification Number", and "Verify that consent is given or authorised by the holder of parental responsibility".

A detailed view of the "Data quality SLA rule" is shown on the right side of the screen. It includes the following information:

- Data quality SLA rule**
- A rule to monitor the data quality of critical data elements
- Data protection rule**: A rule to mask or deny access to data based on metadata.
- Data location rule**: A rule for sovereignty and location based enforcement
- Data governance rule**: A rule used as documentation

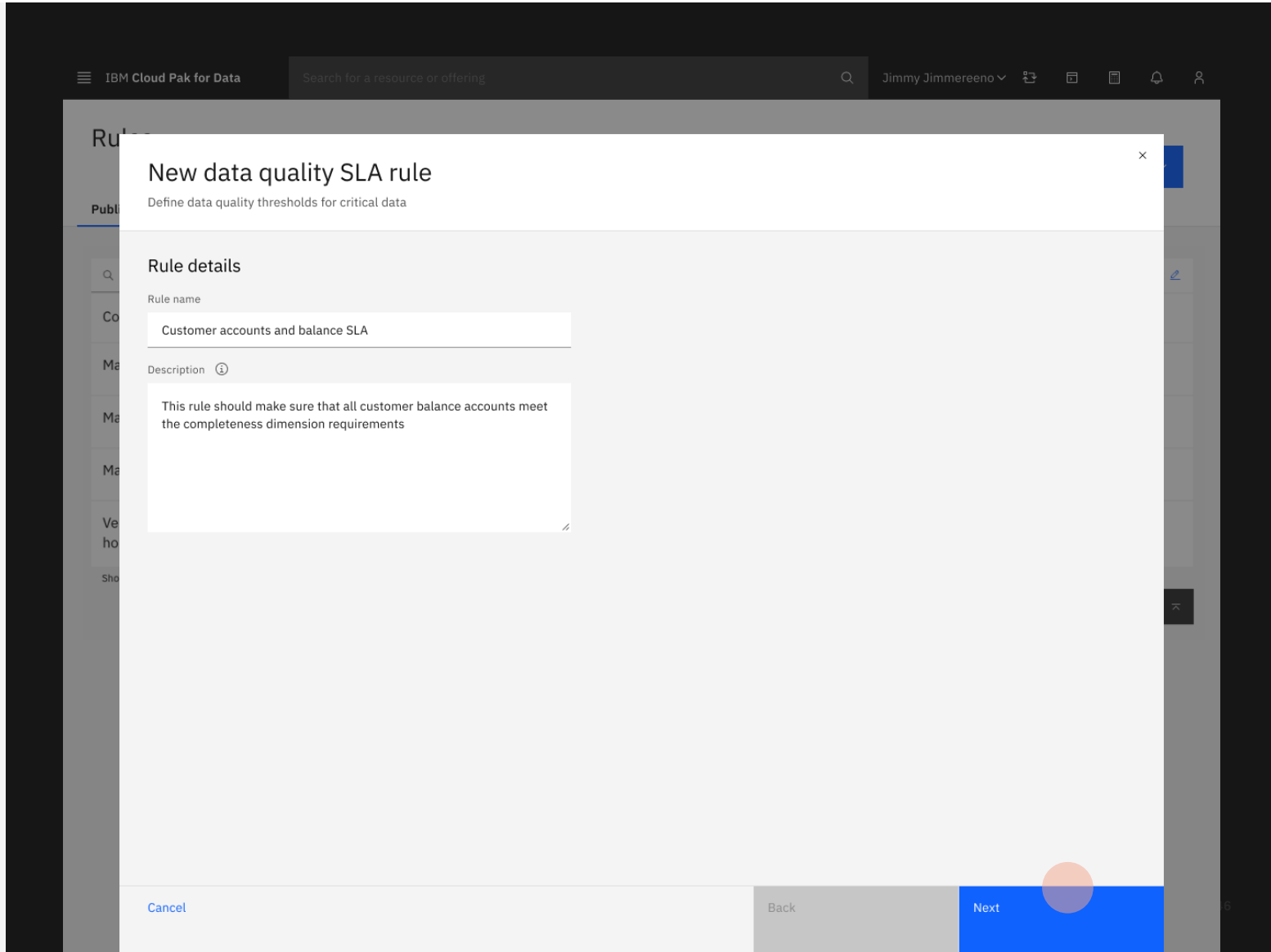
At the bottom of the detailed view, there is a small box containing the text: "Where point (a) of Article 6(1) applies, in relation to the offer of information society services directly to a child, verify in such cases that consent is giv...". To the right of this box, there is a "Knowledge Accelerator for..." button and a "Governance rule | Last modified: Nov 14, 2022, 5:59 PM by admin" label.

At the bottom left of the rules list, it says "Showing 5 of 5 accessible rules".

DEFINE DATA EXPECTATIONS

Data quality SLA rule example

Service Level Agreements are introduced to give meaningful and business-centric reflection of your data's quality.



DEFINE DATA EXPECTATIONS

Data quality SLA rule example

Service Level Agreements are introduced to give meaningful and business-centric reflection of your data's quality.

The screenshot displays the IBM Cloud Pak for Data interface. At the top, the header includes the logo, a search bar, and the user name 'Jimmy Jimmereeno'. The main content area is a modal window titled 'New data quality SLA rule' with the subtitle 'Define data quality thresholds for critical data'. The window is divided into several sections: 'Rule conditions', 'Action if any condition is not met', and a footer with navigation buttons. The 'Rule conditions' section contains two conditions connected by 'and'. The first condition is 'Any data asset' with a dropdown menu set to 'with one of the names' and a tag 'Customer data'. Below it, 'must have a' is set to 'Completeness dimension score' with a threshold of 'equal to or greater than 90%'. The second condition is 'any column within' with a dropdown menu set to 'with one of the business terms' and tags 'Customer accounts' and 'Customer balance'. Below it, 'must have a' is set to 'Uniqueness dimension score' with a threshold of 'equal to or greater than 98%'. The 'Action if any condition is not met' section has a 'Remediation task' dropdown set to 'No workflow selected. No remediation task will be triggered.' The footer contains 'Cancel', 'Back', and 'Create rule' buttons.

DEFINE DATA EXPECTATIONS

Data quality SLA rule example

Service Level Agreements are introduced to give meaningful and business-centric reflection of your data's quality.

The screenshot shows the IBM Cloud Pak for Data interface with a modal window titled "New data quality SLA rule". The window's subtitle is "Define data quality thresholds for critical data".

Rule conditions

- Any data asset**
 - with one of the names (dropdown menu open, showing options: "with one of the names" (checked), "with one of the business terms", "with one of the classifications", "with one of the tags", "without further selection criteria", "with one of the business terms")
 - Customer data (tag)
 - equal to or greater than 90% (slider)
- Customer accounts (tag) and Customer balance (tag)
- must have a**
 - Uniqueness dimension score (dropdown)
 - equal to or greater than 98% (slider)

[Add quality criteria +](#)

[Add subcondition +](#)

Action if any condition is not met

Remediation task [Select +](#)

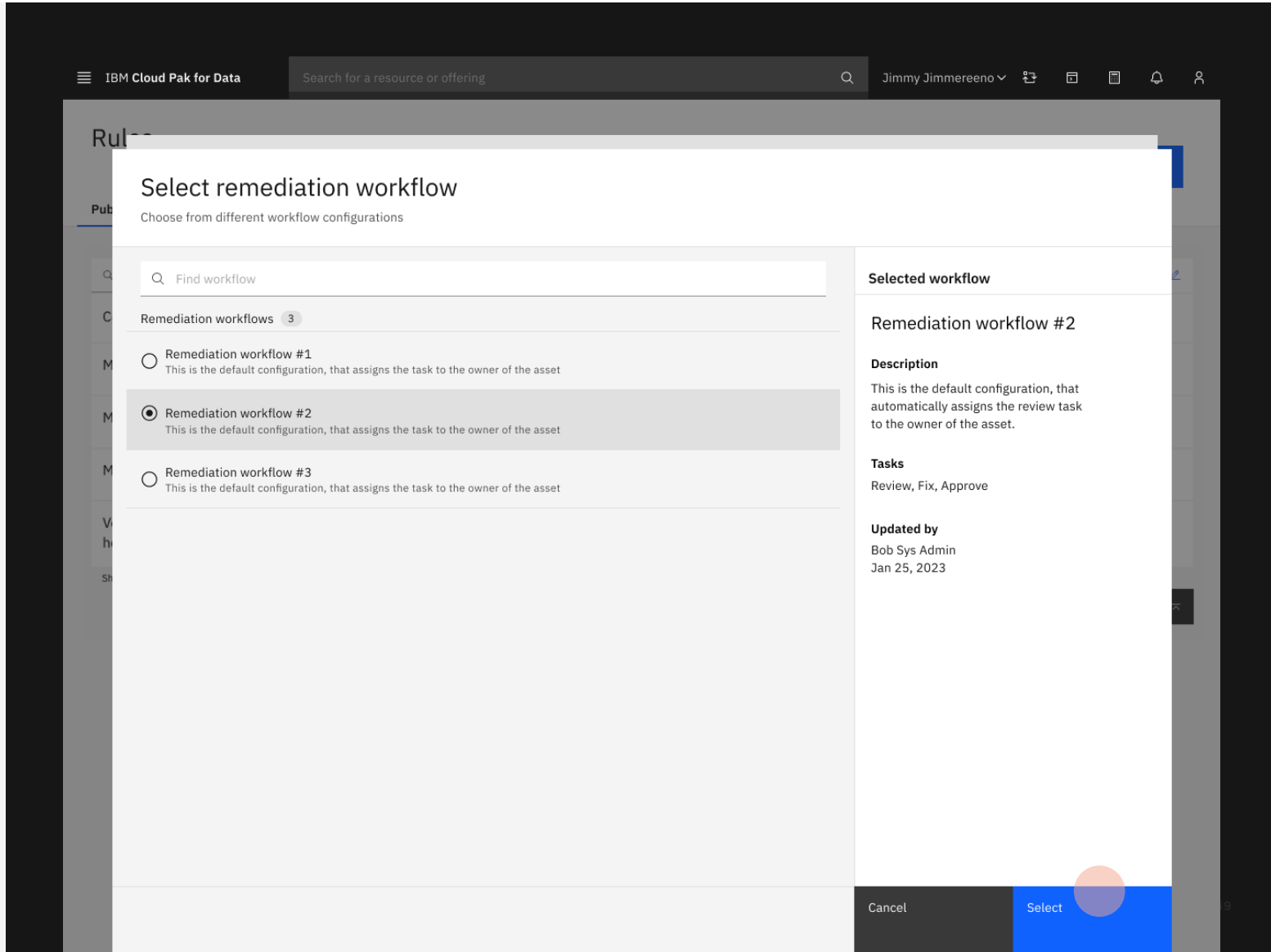
No workflow selected. No remediation task will be triggered.

Buttons at the bottom: [Cancel](#), [Back](#), [Create rule](#)

DEFINE DATA EXPECTATIONS

Data quality SLA rule example

Service Level Agreements are introduced to give meaningful and business-centric reflection of your data's quality.



DEFINE DATA EXPECTATIONS

Data quality SLA rule example

Service Level Agreements are introduced to give meaningful and business-centric reflection of your data's quality.

IBM Cloud Pak for Data

Search for a resource or offering

Jimmy Jimmereeno

New data quality SLA rule

Define data quality thresholds for critical data

Rule conditions

Any data asset

with one of the names

must have a

Completeness dimension score

[Add quality criteria +](#)

and

any column within

with one of the business terms

must have a

Uniqueness dimension score

[Add quality criteria +](#)

[Add subcondition +](#)

Action if any condition is not met

Remediation task [Edit](#)

Remediation workflow #2

[Cancel](#) [Back](#) [Create rule](#)

DEFINE DATA EXPECTATIONS

Data quality SLA rule example

Service Level Agreements are introduced to give meaningful and business-centric reflection of your data's quality.

The screenshot displays the IBM Cloud Pak for Data interface for configuring a Data Quality SLA rule. The page title is "Customer accounts and balance SLA" with a subtitle "Data quality SLA rule". A notification banner at the top right states "Rule successfully created" and "Data quality SLA rule has been successfully created." The interface includes buttons for "Edit rule" and "Delete rule".

Rule conditions

Any data asset

- with one of the names: Customer data
- must have a completeness dimension score equal to or greater than 90%

and

any column within

- with one of the business terms: Customer accounts, Customer balance
- must have a Uniqueness dimension score equal to or greater than 98%

Action if any condition is not met

Remediation task

Remediation workflow #2

Related rules

Buttons: Add rule +

<input type="checkbox"/>	Name	Description
--------------------------	------	-------------

Unified data quality approach as an [automated process](#)



UNDERSTAND DATA

Identify & classify critical data elements

Prepare governance artifacts and auto-assign business terms and classifications to data in Metadata enrichment.



ANALYSE DATA

Run data quality checks on data

Create and run source agnostic analysis to get data quality scores based on dimensions.



DEFINE QUALITY EXPECTATIONS

Data quality governance with SLA rules

Define thresholds as acceptable criteria and decide what automated action to take in case of violations.



ONGOING MONITORING

Monitor data quality & SLA compliance

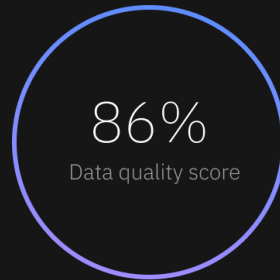
Understand data quality and review SLA compliance. Investigate issues. Subscribe to data quality events.

ONGOING MONITORING

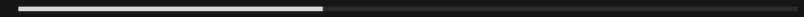
Monitoring data quality on asset level

Unified data quality monitoring for data providers and consumers on data asset level in projects and catalogs.

Asset data quality score



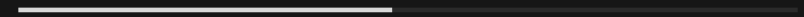
Accuracy



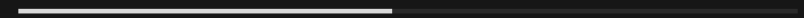
Completeness



Consistency



Timeliness



SLA rule compliance

██████████	▲	██████████	██████████
██████████	✓	██████████	██████████
██████████	✓	██████████	██████████

Data quality checks

██████████	██████████	██████████
██████████	██████████	██████████
██████████	██████████	██████████

ONGOING MONITORING

Monitoring data quality on asset level

Located on asset level in a project in this example

The screenshot displays the IBM Cloud Pak for Data interface. At the top, the header includes the product name, a search bar, and the user's name 'Jimmy Jimmereeno'. The main content area is titled 'Projects / Credit card customers' and features tabs for 'Overview', 'Assets', 'Jobs', and 'Manage'. The 'Assets' tab is active, showing a search bar for 'Find assets', an 'Import assets' button, and a 'New asset' button. On the left, a sidebar lists '9 assets' and 'Asset types' including 'Data access' (2), 'Data' (3), 'Data quality' (2), and 'Flows' (2). The 'Data' asset type is selected. The main panel shows a table of data assets:

Data		
Name		Last modified
BANK_CLIENTS application/octet-stream		1 month ago Modified by you
Customer_data Connection		1 month ago Modified by you

ONGOING MONITORING

Monitoring data quality on asset level

Data quality exists on a separate dedicated tab

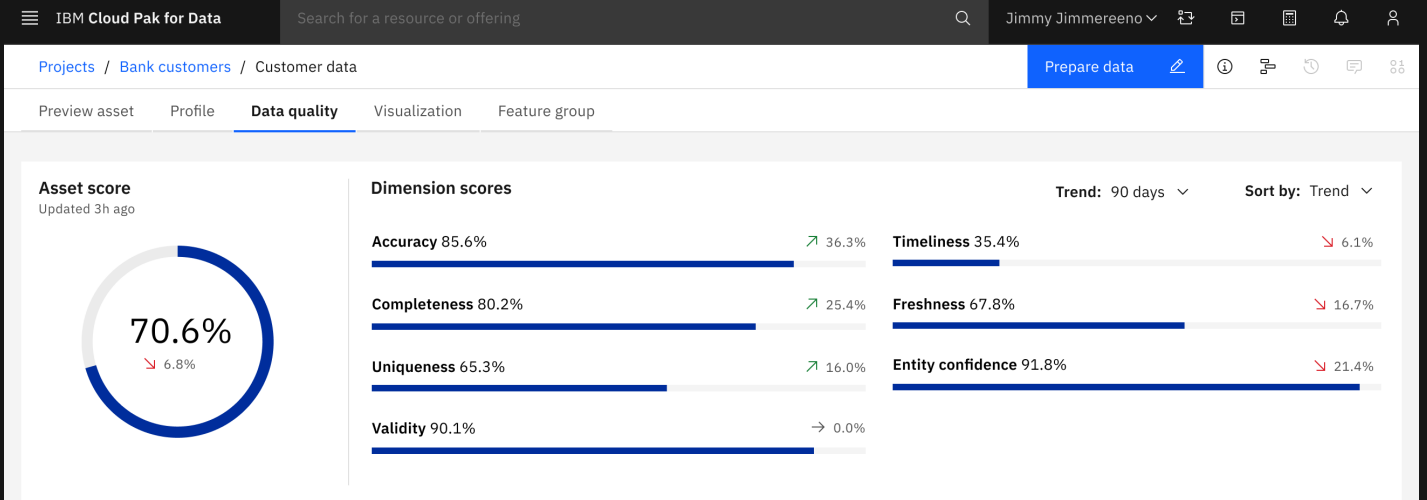
The screenshot shows the IBM Cloud Pak for Data interface. At the top, there is a search bar and the user name 'Jimmy Jimmereeno'. Below the search bar, the breadcrumb path is 'Projects / Credit card customers / Customer_data'. There are four tabs: 'Preview', 'Profile', 'Data Quality', and 'Lineage'. The 'Data Quality' tab is selected and highlighted with a red circle. Below the tabs, it says 'Profile size: 12 Columns' and 'The preview includes only a limited set of columns and rows.' followed by an information icon. On the right, it says 'Last refresh: just now' with a refresh icon. The main area contains a table with 11 columns: customer... String, company... String, contactN... String, contactTI... String, address String, city String, region String, postalCode String, country String, phone String, and fax String. The table lists 20 rows of customer data.

customer... String	company... String	contactN... String	contactTI... String	address String	city String	region String	postalCode String	country String	phone String	fax String
ALFKI	Alfreds Futterkiste	Maria Anders	Sales Representa	Obere Str. 57	Berlin	NULL	12209	Germany	030-0074321	030-0076545
ANATR	Ana Trujillo Empa	Ana Trujillo	Owner	Avda. de la Cons	México D.F.	NULL	05021	Mexico	(5) 555-4729	(5) 555-3745
ANTON	Antonio Moreno T	Antonio Moreno	Owner	Mataderos 2312	México D.F.	NULL	05023	Mexico	(5) 555-3932	NULL
AROUT	Around the Horn	Thomas Hardy	Sales Representa	120 Hanover Sq.	London	NULL	WA1 1DP	UK	(171) 555-7788	(171) 555-6750
BERGS	Berglunds snabbk	Christina Berglur	Order Administr	Berguvsvägen 8	Luleå	NULL	S-958 22	Sweden	0921-12 34 65	0921-12 34 67
BLAUS	Blauer See Delika	Hanna Moos	Sales Representa	Forsterstr. 57	Mannheim	NULL	68306	Germany	0621-08460	0621-08924
BLONP	Blondesdsl père	Frédérique Citea	Marketing Manag	24	place Kléber	Strasbourg	NULL	67000	France	88.60.15.31
BOLID	Bólido Comidas pi	Martín Sommer	Owner	C/ Araquil	67	Madrid	NULL	28023	Spain	(91) 555 22 82
BONAP	Bon app'	Laurence Lebihai	Owner	12	rue des Boucher	Marseille	NULL	13008	France	91.24.45.40
BOTTM	Bottom-Dollar Ma	Elizabeth Lincoln	Accounting Man	23 Tsawassen Bl	Tsawassen	BC	T2F 8M4	Canada	(604) 555-4729	(604) 555-3745
BSBEV	B's Beverages	Victoria Ashwort	Sales Representa	Fauntleroy Circu	London	NULL	EC2 5NT	UK	(171) 555-1212	NULL
CACTU	Cactus Comidas p	Patricio Simpson	Sales Agent	Cerrito 333	Buenos Aires	NULL	1010	Argentina	(1) 135-5555	(1) 135-4892
CENTC	Centro comercial l	Francisco Chang	Marketing Manag	Sierras de Grana	México D.F.	NULL	05022	Mexico	(5) 555-3392	(5) 555-7293
CHOPS	Chop-suey Chines	Yang Wang	Owner	Hauptstr. 29	Bern	NULL	3012	Switzerland	0452-076545	NULL
COMMI	Comércio Mineiro	Pedro Afonso	Sales Associate	Av. dos Lusíadas	23	Sao Paulo	SP	05432-043	Brazil	(11) 555-7647
CONSH	Consolidated Holc	Elizabeth Brown	Sales Representa	Berkeley Garden	London	NULL	WX1 6LT	UK	(171) 555-2282	(171) 555-9199
DRACD	Drachenblut Delik	Sven Ottlieb	Order Administr	Walsersweg 21	Aachen	NULL	52066	Germany	0241-039123	0241-059428
DUMON	Du monde entier	Janine Labrune	Owner	67	rue des Cinquant	Nantes	NULL	44000	France	40.67.88.88
EASTC	Eastern Connectic	Ann Devon	Sales Agent	35 King George	London	NULL	WX3 6FW	UK	(171) 555-0297	(171) 555-3373
ERNSH	Ernst Handel	Roland Mendel	Sales Manager	Kirchgasse 6	Graz	NULL	8010	Austria	7675-3425	7675-3426

ONGOING MONITORING

Monitoring data quality on asset level

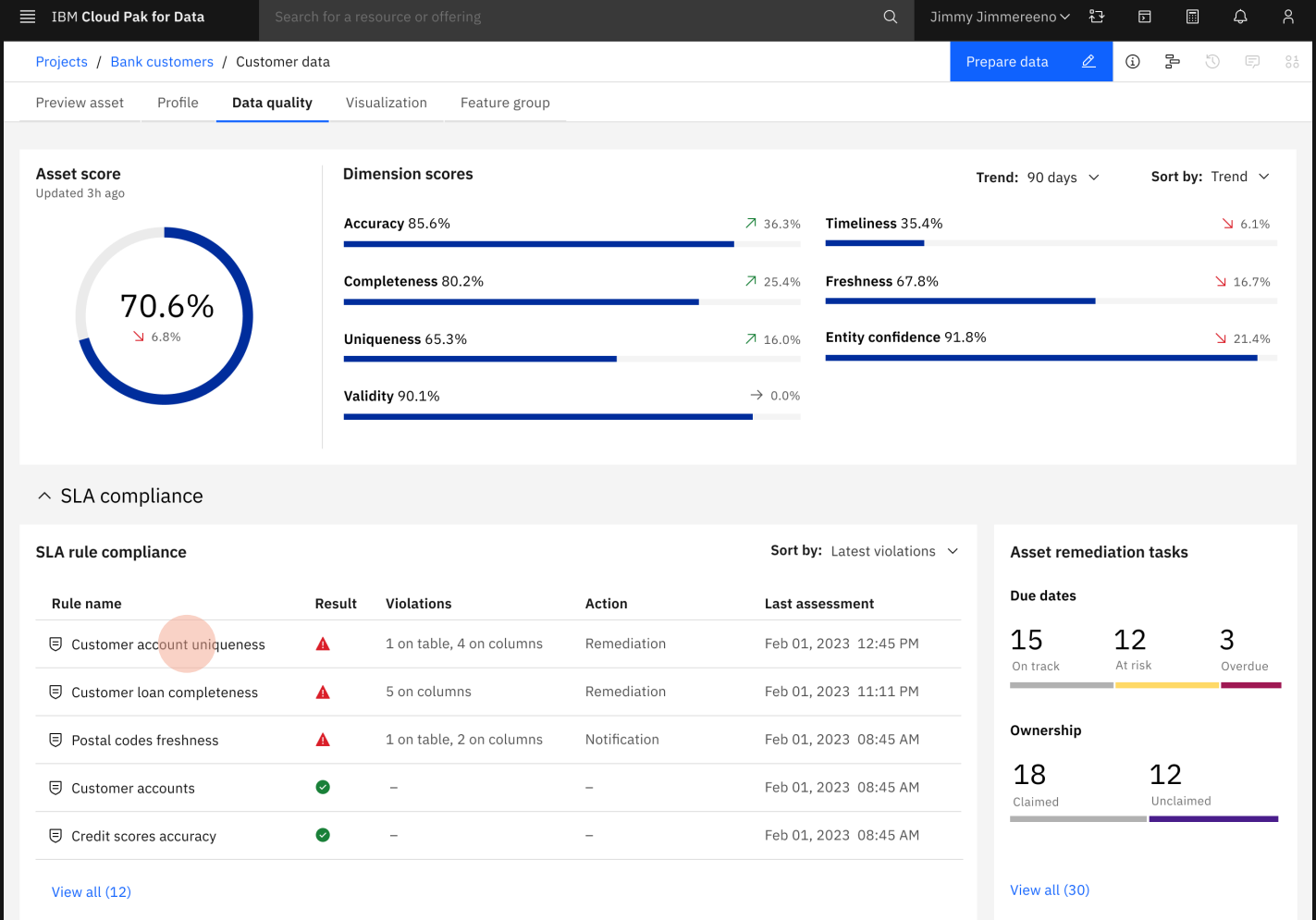
Data users can identify data quality metrics and trends from the centralized data quality tab found directly inside the data asset.



ONGOING MONITORING

Monitoring data quality on asset level

They can also view the status of SLA compliance and corresponding remediation tasks.



ONGOING MONITORING

Monitoring data quality on asset level

Users can learn how to prevent future violations by investigating the violation details and severity, as well as the rule details and expectations.

The screenshot displays the IBM Cloud Pak for Data interface. The main content area shows the 'Customer accounts and balance SLA' assessment results for 'Customer data' as of Feb 01, 2023, 12:45 PM. The interface includes a navigation sidebar on the left and a top search bar. The main content is divided into two sections: 'Violations and checks' and 'Data quality checks'.

Customer accounts and balance SLA

Time of assessment: Feb 01, 2023 12:45 PM

[Run history](#) [Edit SLA rule](#)

Violations and checks

Violating element	Dimension	Expected score	Current score	Deviation	Rule target
Customer data Table	Completeness	≥ 90%	48%	42pp	Customer data Asset name
Account number Column	Uniqueness	≥ 98%	56%	42pp	Customer account Business term
Balances Column	Uniqueness	≥ 98%	84%	14pp	Customer balances Business term

Data quality checks

Name & logic	Type	Dimension	Focus & percentage of data with issues	Data checked & issues found	Sampling	Score	Last checked
Unique account number UniqueAccountNum	Data quality rule	Uniqueness	3 columns 41.8% of data	941 records 432 records	75%	58.2% ↓ 16%	Jan 11, 2023 07:08:09 PM
CustomerDataCheck UniqueAccountNum	Data quality rule	Uniqueness	3 columns 42% of data	941 records 432 records	75%	55% ↓ 16%	Jan 11, 2023 07:08:09 PM
Latest transaction Transaction	Data quality rule	Completeness	1 column 80% of data	941 records 889 records	75%	20% ↓ 16%	Jan 11, 2023 07:08:09 PM

ONGOING MONITORING

Monitoring data quality on asset level

They can also view a list of all existing platform wide active data quality checks analysing this asset.

The screenshot displays the IBM Cloud Pak for Data interface for monitoring data quality on an asset level. The breadcrumb path is 'Projects / Bank customers / Customer data'. The 'Data quality' tab is active, showing an 'Asset score' of 70.6% (updated 3h ago) with a 6.8% change. Below this, 'Dimension scores' are listed: Accuracy (85.6%), Completeness (80.2%), Uniqueness (65.3%), Validity (90.1%), Timeliness (35.4%), Freshness (67.8%), and Entity confidence (91.8%). Each score is accompanied by a progress bar and a trend indicator. The 'SLA compliance' section shows a table of rule violations, and the 'Asset remediation tasks' section shows due dates and ownership status.

Asset score
Updated 3h ago
70.6%
6.8%

Dimension scores
Trend: 90 days | Sort by: Trend

Dimension	Score	Trend
Accuracy	85.6%	36.3%
Completeness	80.2%	25.4%
Uniqueness	65.3%	16.0%
Validity	90.1%	0.0%
Timeliness	35.4%	6.1%
Freshness	67.8%	16.7%
Entity confidence	91.8%	21.4%

SLA compliance
Sort by: Latest violations

Rule name	Result	Violations	Action	Last assessment
Customer account uniqueness	⚠️	1 on table, 4 on columns	Remediation	Feb 01, 2023 12:45 PM
Customer loan completeness	⚠️	5 on columns	Remediation	Feb 01, 2023 11:11 PM
Postal codes freshness	⚠️	1 on table, 2 on columns	Notification	Feb 01, 2023 08:45 AM
Customer accounts	✅	-	-	Feb 01, 2023 08:45 AM
Credit scores accuracy	✅	-	-	Feb 01, 2023 08:45 AM

[View all \(12\)](#)

Asset remediation tasks

Due dates
15 On track | 12 At risk | 3 Overdue

Ownership
18 Claimed | 12 Unclaimed

[View all \(30\)](#)

ONGOING MONITORING

Monitoring data quality on asset level

They can also view a list of all existing platform wide active data quality checks analysing this asset.

The screenshot displays the IBM Cloud Pak for Data interface for monitoring data quality on an asset level. The breadcrumb navigation shows 'Projects / Bank customers / Customer data'. The main content area is divided into several sections:

- Asset score:** A donut chart showing an overall score of 70.6%, updated 3 hours ago. A small red arrow indicates a 6.8% change.
- Dimension scores:** A table of scores for various dimensions, each with a progress bar and a trend indicator (up, down, or flat).
- SLA compliance:** A table listing rule compliance with columns for Rule name, Result, Violations, Action, and Last assessment.
- Asset remediation tasks:** A summary of due dates and ownership, including counts for 'On track', 'At risk', 'Overdue', 'Claimed', and 'Unclaimed'.

At the bottom, there are tabs for 'Checks' and 'Columns'.

Dimension	Score	Trend
Accuracy	85.6%	36.3% ↑
Completeness	80.2%	25.4% ↑
Uniqueness	65.3%	16.0% ↑
Validity	90.1%	0.0% →
Timeliness	35.4%	6.1% ↓
Freshness	67.8%	16.7% ↓
Entity confidence	91.8%	21.4% ↓

Rule name	Result	Violations	Action	Last assessment
Customer account uniqueness	⚠	1 on table, 4 on columns	Remediation	Feb 01, 2023 12:45 PM
Customer loan completeness	⚠	5 on columns	Remediation	Feb 01, 2023 11:11 PM
Postal codes freshness	⚠	1 on table, 2 on columns	Notification	Feb 01, 2023 08:45 AM
Customer accounts	✅	-	-	Feb 01, 2023 08:45 AM
Credit scores accuracy	✅	-	-	Feb 01, 2023 08:45 AM

Category	Count
On track	15
At risk	12
Overdue	3
Claimed	18
Unclaimed	12

ONGOING MONITORING

Monitoring data quality on asset level

They can also view a list of all existing platform wide active data quality checks analysing this asset.

^ SLA compliance

SLA rule compliance Sort by: Latest violations ▾

Rule name	Result	Violations	Action	Last assessment
Customer account uniqueness	▲	1 on table, 4 on columns	Remediation	Feb 01, 2023 12:45 PM
Customer loan completeness	▲	5 on columns	Remediation	Feb 01, 2023 11:11 PM
Postal codes freshness	▲	1 on table, 2 on columns	Notification	Feb 01, 2023 08:45 AM
Customer accounts	●	-	-	Feb 01, 2023 08:45 AM
Credit scores accuracy	●	-	-	Feb 01, 2023 08:45 AM

[View all \(12\)](#)

Asset remediation tasks

Due dates

15 On track 12 At risk 3 Overdue

Ownership

18 Claimed 12 Unclaimed

[View all \(30\)](#)

^ Data quality checks

Checks Columns

Find data quality checks by name Create data quality check ▾

Name & logic ⓘ	Type	Dimension	Focus & percentage of data with issues	Data checked & issues found	Sampling	Score	Contributes to overall score	Last checked
Unique account number Validaccount	Data quality rule	Uniqueness	1 column 88.6% of data	941 records 1237 issues	200 records Interval	11.4% ↓ 16.1%	●	Jan 11, 2023 07:08:09 PM
Valid_OperationalCountry ValidCountry	Data quality rule	Validity	3 columns 45.1% of data	941 records 432 issues	1000 records Random	54.9% ↓ 16.1%	●	Jan 11, 2023 07:08:09 PM
Check_personal_data ValidAge	Data quality rule	Validity	3 columns 18.0% of data	941 records 225 issues	1000 records Random	82.0% ↓ 16.2%	●	Jan 11, 2023 07:08:09 PM
Potential matches	Matching	Entity confidence	Table 8.2% of entities	941 entities 889 issues	-	91.8% → 0.0%	●	Jan 11, 2023 07:08:09 PM
Suspect values	Profiling	Validity	144 Columns 88.1% of data	941 records 837 issues	50% of records Random	11.9% ↓ 16.1%	●	Jan 11, 2023 07:08:09 PM
Duplicated values	Profiling	Uniqueness	144 Columns 83.5% of data	941 records 837 issues	-	16.5% ↑ 16.2%	●	Jan 11, 2023 07:08:09 PM

ONGOING MONITORING

Monitoring data quality on asset level

Or switch to a column overview where all columns are displayed with an overall score and a breakdown on data quality dimensions.

SLA compliance

SLA rule compliance Sort by: Latest violations ▾

Rule name	Result	Violations	Action	Last assessment
Customer account uniqueness	⚠	1 on table, 4 on columns	Remediation	Feb 01, 2023 12:45 PM
Customer loan completeness	⚠	5 on columns	Remediation	Feb 01, 2023 11:11 PM
Postal codes freshness	⚠	1 on table, 2 on columns	Notification	Feb 01, 2023 08:45 AM
Customer accounts	✅	-	-	Feb 01, 2023 08:45 AM
Credit scores accuracy	✅	-	-	Feb 01, 2023 08:45 AM

[View all \(12\)](#)

Asset remediation tasks

Due dates

15 On track 12 At risk 3 Overdue

Ownership

18 Claimed 12 Unclaimed

[View all \(30\)](#)

Column overview

Checks Columns

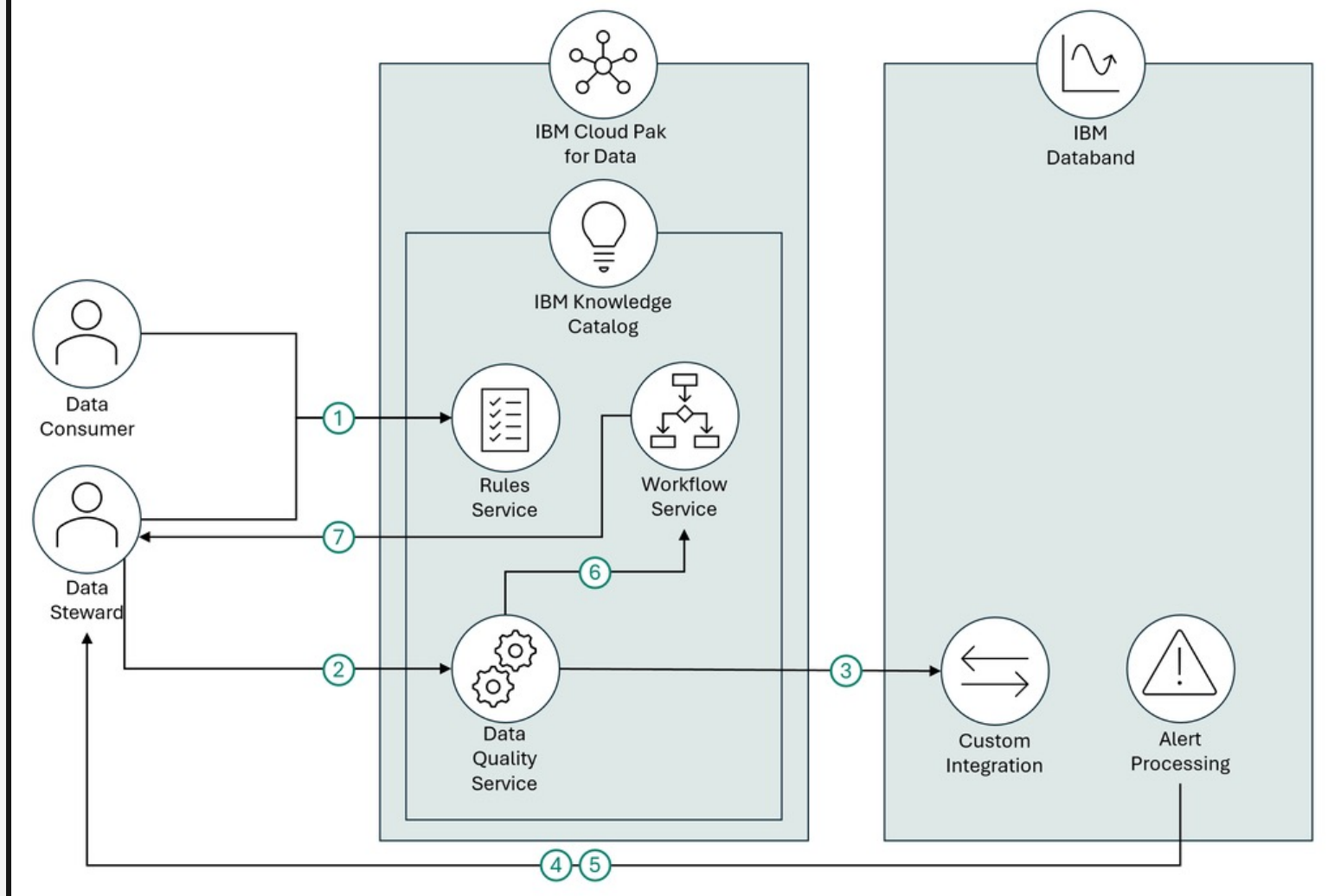
Find columns ☰

Column	Score	Accuracy Dimension	Completeness Dimension	Uniqueness Dimension	Validity Dimension	Timeliness Dimension	Freshness Dimension	Checks	Contributes to overall score	Last checked
Age Number	73%	23%	19%	55%	77%	-	-	23	🟢	Jan 11, 2023 07:08:09 PM
Job Varchar	23%	23%	55%	21%	55%	-	-	23	🟢	Jan 11, 2023 07:08:09 PM
Marital Varchar	73%	23%	23%	61%	55%	-	-	23	🟢	Jan 11, 2023 07:08:09 PM
Housing Varchar	23%	23%	55%	87%	55%	-	-	23	🔴	Jan 11, 2023 07:08:09 PM
Loan Varchar	23%	23%	55%	55%	55%	-	-	23	🟢	Jan 11, 2023 07:08:09 PM
Year Number	23%	23%	55%	55%	55%	-	-	23	🔴	Mon 1 Jan 2023 07:08:09 PST

ONGOING MONITORING

Monitoring data quality pipeline using IBM Databand

1. Data Steward and Data Consumer create agreed data quality service level agreements (DQSLAs)
2. Data Steward schedules automated data quality process
3. Each data quality process run for an asset is reported and appears as a pipeline run in IBM Databand
4. For each pipeline run, a critical upstream failure (unexpected decrease of contributing data quality scores over time) triggers a proactive notification (slack or email) of the responsible data stewards for remediation
5. For each pipeline, a run frequency anomaly triggers a notification (slack or email) of the responsible data steward
6. If the asset violates the DQSLA, a reactive workflow is triggered
7. The data steward will see tasks created by the workflow in the task inbox



Unified data quality approach as an [automated process](#)



UNDERSTAND DATA

Identify & classify critical data elements

Prepare governance artifacts and auto-assign business terms and classifications to data in Metadata enrichment.



ANALYSE DATA

Run data quality checks on data

Create and run source agnostic analysis to get data quality scores based on dimensions.



DEFINE QUALITY EXPECTATIONS

Data quality governance with SLA rules

Define thresholds as acceptable criteria and decide what automated action to take in case of violations.



ONGOING MONITORING

Monitor data quality & SLA compliance

Review data quality scores based on dimensions and SLA compliance. Investigate cause of issues and see failed records.

Thank you
Your questions ...

If you like to learn more about our solutions, we'd be happy to connect with you via email or phone, just let us know via this form

Link: <https://www.ibm.com/account/reg/us-en-signup?formid=urx-32526>