Identifying and Managing Data Quality Requirements: A Design Science Study in the Field of Automated Driving - Data Package

Shameer Kumar Pradhan, Hans-Martin Heyn, Eric Knauss

1 Introduction

This document acts as a data package for the article titled ***Identifying and Managing Data Quality Requirements: A Design Science Study in the Field of Automated Driving***. It covers the following topics.

- Section 2 - Interview Guide
- Section 3 - Survey Questionnaires
- Section 4 - Challenge Score
- Section 5 - Focus Group Data
- Section 6 - Survey 2 Data

2 Interview Guide

2.1 Interview Standardized Consent Form

This section presents a standardized consent form presented to the interviews prior to the interviews in the first and the second cycles. Table 1 presents the fields of the form, the description of those fields, and sample data that could be filled in those fields.

Table 1: Interview Standardized Consent Form Template

Field	Description	Example
Interview number	The identifier for the order of the interview with a particular interviewee.	*1-1*
Interviewee	Name of the interviewee	*John Doe*
Position	Current position held by the interviewee	*Product Owner*
Team	The team that the interviewee belongs to in his/her organization	*Research*
Company	Name of the company in which the interviewee is employed at	*Veoneer Sweden*
Academic background	Academic degree held by the interviewee and the name of his/her field	*PhD in Electrical Engineering*
Experience	Number of years of work experience the interviewee has in his/her field	*15 years*
Date	Date on which the interview is conducted	*March 17, 2021*
Length (in minutes)	Number of minutes the interview was conducted	*80 minutes*
Consent for audio recording	Boolean value denoting if consent was given by the interviewee for audio recording of the interview session	*Yes*
Consent for publication	Boolean value denoting if consent was given by the interviewee for publication of data from the interview	*Yes*

2.2 Interview Questions

This section presents the questions of the interviews conducted in both the first and second cycles of the study. There are three versions of the Cycle 1 question set. Although there is no vast difference between the questions in these versions, there are slight modifications based on the feedback received during the interviews.

2.2.1 Cycle 1 Interview Question Set - Version 1

General Questions

1. What is your role in the company?
2. What does your team focus on?
3. What kind of system are you working on in your team?

Data Questions

4. What are the constituent components of the system that you are working on?
5. How are the volume, variety, and velocity of the data in the system you work on and how does it impact the quality of data?
6. What are the key data gathered in your field?
7. How are the data collected used in your field?
8. How does the data affect the behavior of the system you are working on?
9. What do you think are appropriate metrics for each of the data you mentioned in Q.7?
10. Why are the metrics you mentioned earlier are appropriate for the data?

Data Quality Procedure Questions

11. In your current context, how would you make data quality measurable?

12. What's the current procedure Veoneer follows to ensure data quality?
    (a) If yes, what are the drawbacks of the current procedure if there is any to ensure data quality?
13. (If they mentioned documentation in Q.12) How can documentation procedure be improved?
14. (If they do not mentioned documentation in Q.12) How are data quality requirements documented in Veoneer?
15. What could be the potential ways in which data quality assessment procedure can be improved?
16. What are some of the challenges that you have faced or you think exist in handling data and assessing the quality of it?

Data and Safety Questions

17. What is the relationship between safety and the data collected?

Wrap-up Questions

18. To summarize, what do you think is the most challenging aspect of handling data?
19. In the future, how do you envision the data quality requirements are dealt with in Veoneer?
20. Do you think we forgot to ask you about something that we should know about?

2.2.2 Cycle 1 Interview Question Set - Version 2

General Questions

1. What is your role in the company?
2. What does your team focus on?
3. What kind of system are you working on in your team? What are the goals of the system?
4. What are the constituent components of the system you are working?

Data Questions

5. Can you tell me what you understand by 'data' in your system? What is typical data in your field?
6. Among these data that you mentioned, what is the most important data in the field?
7. How would you set requirements on data?
8. Who is responsible for setting the data requirements?
9. In your opinion, how would you evaluate the importance of each person or group responsible to set the data requirements?
10. How would you characterize good data quality in your system? Can you give examples of it?
11. What would data with subpar quality look like in your system?
12. What are the typical metrics you apply to the data? Why are these the appropriate metrics for the data?
13. How does the data affect the behavior of the system you are working on?
14. In your current context, how do you make data quality measurable?
15. Do you have a procedure that you follow in your team to ensure data quality?
    (a) If yes, can you describe the procedure?
        i. How are data quality requirements elicited and set in your team?

ii. How do you segregate good and bad data?

(b) If yes, what are the drawbacks and challenges of the current procedure?

(c) If yes, how do you document the data quality requirements?

(d) If yes, what could be the potential ways in which data quality assessment procedure can be improved?

(e) If not, how would you make data quality measurable? Can you describe an ideal procedure to assess data quality in your system?

Data and Safety Questions

16. What is the relationship between the goal of your system and the data collected?

Wrap-up Questions

17. In the future, how do you envision the data quality requirements are dealt with in your team?

18. Do you think we forgot to ask you about something that we should know about?

2.2.3 Cycle 1 Interview Question Set - Version 3

General Questions

1. What is your role in the company?

2. What does your team focus on?

3. What kind of system are you working on in your team? What are the goals of the system?

4. What are the constituent components of the system you are working?

Data Questions

5. Can you tell me what you understand by 'data' in your system? What is typical data in your field?

6. Among these data that you mentioned, what is the most important data in the field?

7. How are data quality requirements elicited and set in your team?

8. Who is responsible for setting the data requirements?

9. In your opinion, how would you evaluate the importance of each person or group responsible to set the data requirements?

10. How would you characterize good data quality in your system? Can you give examples of it?

11. What would data with subpar quality look like in your system?

12. What are the typical metrics you apply to the data? Why are these the appropriate metrics for the data?

13. How does the data affect the behavior of the system you are working on?

14. In your current context, how do you make data quality measurable?

15. Do you have a procedure that you follow in your team to ensure data quality?

    (a) If yes, can you describe the procedure?

    (b) If yes, what are the drawbacks and challenges of the current procedure?

    (c) If yes, how do you document the data quality requirements?

    (d) If yes, what could be the potential ways in which data quality assessment procedure can be improved?

(e) If not, can you describe an ideal procedure to assess data quality in your system?

16. How do you segregate good and bad data?

Data and Safety Questions

17. What is the relationship between the goal of your system and the data collected?

Wrap-up Questions

18. In the future, how do you envision the data quality requirements are dealt with in your team?
19. Do you think we forgot to ask you about something that we should know about?

The questions asked in the interviews of the second cycle were more open-ended. The questions pertain to the potential solutions to the data quality challenges. The following is a general framing of the questions for individual potential solutions.

2.2.4 Cycle 2 Interview Question Set

Potential Solutions Questions

1. What is the information needed to determine the solution?
2. When should you decide to determine the information for this solution?
3. Is the solution useful for first-party or third-party data?
4. How to use this solution? What is the process to identify and solve this challenge?
5. What constraints do you see in practical implementation of this solution, if any?
6. Is this solution currently being used in your organization?
7. Does this solution actually help solve the challenge or not?
8. What is your opinion regarding the flow of the solution?
9. Is this just a theoretical solution or a practically applicable one as well?
10. If not, what can be other potential solutions for the challenge?
11. In your opinion, what kind of terms between the seller and the buyer help solve this challenge?
12. What kind of penalty and tolerance are applicable for this challenge?

3 Survey Questionnaires

Note: The options that the survey participants can select from are provided inside square brackets ([ ]) and are separated by a slash (/). For example, *[Yes/No]* means a survey participant can provide either a "Yes" or a "No" response.

3.1 Survey 1 Questionnaire

Survey 1 Questionnaire was sent during the first cycle. Below is the set of questions presented in the questionnaire form.

Note: Mandatory questions are denoted with an asterisk (*).

1. Please rank the following 'Data source' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*
    - Wrongly-calibrated / defective sensor generates incorrect data

- Reliance on a single data source (e.g. only depending on a single radar to collect data, or depending on a single type of sensor)
- Fast increasing data overwhelms the ML algorithms that need to process the data
- Data dependent on external conditions can be of low quality sometimes
- New data types from various sources make data integration difficult
- Lack of variety in test environment causes AI to be poorly trained for situations that it has not been trained for
- Noise (unwanted data that is mixed with valuable data)

2. Please rank the following 'Data Availability' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*

    - Low labeled data volume (i.e. the amount of data that is labeled is lesser than the the amount of data that is unlabeled)
    - Data delay (i.e. there is a delay in transmission of data from the source to the destination, e.g. data store to function, sensor to sensor)
    - Data drop (e.g. dropping of a required attribute of data, dropping a chunk of data during transmission)
    - Incomplete data (e.g. missing attribute of data, missing chunk of data)

3. Please rank the following 'Data management' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*

    - Improper data transfer (e.g. mismanagement of the way data is transferred between source and destination)
    - Redundant data
    - Large volume of data makes it difficult to assess the quality of the data
    - Reliance on suppliers to raise error (i.e. there is no internal process to raise an error, and raising error is dependent on the suppliers doing so)
    - Data ownership (i.e. who is the legal owner of the data and can use it without any prior approval?)
    - Data acquisition (i.e. purchase of data from third-party and processes associated with it)
    - Imbalanced dataset (e.g. data from a single location or single weather type dominates other locations or weathers)
    - Regulatory compliance (i.e. what are the rules and regulations for handling data and assuring data quality)
    - Manual data collection
    - Manual data labeling

4. Please rank the following 'Trust in data' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*

    - Fake data (e.g. data is maliciously manipulated)
    - Incorrect labeling
    - Wrong metadata
    - Lack of good data from simulations (i.e. data from simulations are relatively of lesser quality than data from real-world collection)
    - Uncertain data quality identification

5. Please rank the following 'Data structure' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*
    - Data fragmentation (i.e. data required by a function is located in different places)
    - Outlier data (i.e. data is out of bounds of acceptable range)
    - Incompatible data formats
    - Unstructured data
    - Data mixup (e.g. rows and columns are mixed together)
6. Please rate these challenges (1 being the least pressing and 6 being the most pressing)*
    - 'Data source' challenges (related to question 1) [1/2/3/4/5/6]
    - 'Availability of data' challenges (related to question 2) [1/2/3/4/5/6]
    - 'Data management' challenges (related to question 3) [1/2/3/4/5/6]
    - 'Trust in data' challenges (related to question 4) [1/2/3/4/5/6]
    - 'Data structure' challenges (related to question 5) [1/2/3/4/5/6]

3.2 Survey 2 Questionnaire

Survey 1 Questionnaire was sent during the third cycle of the study. Below is the set of questions presented in the questionnaire form.

Note: Mandatory questions are denoted with an asterisk (*).

Artifact

In this section, we are trying to validate the structure of the artifacts.

1. The following is the template for 'List of challenges' artifact. This artifact attempts to present each of the data quality challenges identified using different methods.*

    Name - The name of the challenge (e.g., Data Delay)
    Sources - How did we identify this challenge? (e.g., interviews, literature review)
    Description - Description of the challenge (e.g., what is it? why is it a challenge?)
    Whether the challenge directly affects AI function - Boolean value stating whether this challenge directly affects AI functions, models
    Challenge score - Ranking of the challenge in terms of 'pressing'-ness (calculated using a formula developed during the thesis work)

    In your opinion, are these appropriate fields for the template for 'List of challenges' artifact?

    - Name [Yes/No]
    - Sources [Yes/No]
    - Description [Yes/No]
    - Whether the challenge directly affects AI function [Yes/No]
    - Challenge score [Yes/No]
2. For the above template for 'List of challenges' artifact, what other field/s do you think can be added so that the artifact provides more information than it does right now?
3. The following is the template for 'List of Data quality attributes' artifact. This artifact attempts to present a list of data quality attributes and whether they are affected by data quality challenges.*

    Name - The name of the data quality attribute (e.g., Accuracy)
    Sources - How did we identify this data quality attribute? (e.g., interviews, literature review)
    Definition - Definition of the data quality attribute (e.g., what is it?)
    Which challenges affect the data quality attribute? - A list of challenges that affect the data

quality attribute

In your opinion, are these appropriate fields for the template for 'List of Data quality attributes' artifact?

- Name [Yes/No]
- Sources [Yes/No]
- Definition [Yes/No]
- Which challenges affect the data quality attribute? [Yes/No]

4. For the above template for 'List of Data quality attributes' artifact, what other field/s do you think can be added so that the artifact provides more information than it does right now?

5. The following is the template for 'List of Data quality attribute metrics' artifact. This artifact attempts to present a list of data quality attribute metrics.

   Data quality attribute - The name of the data
   quality attribute (e.g., Accuracy)
   Metric - The name of the data quality attribute metric (e.g., Degree of accuracy)
   Formula - Formula of how the metric is calculated

   In your opinion, are these appropriate fields for the template for 'List of Data quality attribute metrics' artifact?*

   - Data quality attribute [Yes/No]
   - Metric [Yes/No]
   - Formula [Yes/No]

6. For the above template for 'List of Data quality attribute metrics' artifact, what other field/s do you think can be added so that the artifact provides more information than it does right now?

7. The following is the template for 'Potential solutions' artifact. This artifact attempts to present and explain a list of potential solutions for data quality challenges.

   Name - The name of the solution (e.g., Continuous data processing)
   Requirements specifications - The information that should specified before the implementation fo the solution (e.g., Determine an acceptable range of time for data arrival)
   Implementation details - How the solution is implemented?

   In your opinion, are these appropriate fields for the template for 'Potential solutions' artifact?*

   - Name [Yes/No]
   - Requirements specifications [Yes/No]
   - Implementation details [Yes/No]

8. For the above template for 'Potential solutions' artifact, what other field/s do you think can be added so that the artifact provides more information than it does right now?

Challenges Affecting AI models
In this section, we are trying to validate the structure of the artifacts.

9. Does the following challenges affect AI models directly or not?*

   - Data delay (i.e. there is a delay in transmission of data from the source to the destination, e.g. data store to function, sensor to sensor) [Yes/No]
   - Data drop (e.g. dropping of a required attribute of data, dropping a chunk of data during transmission) [Yes/No]

- Incomplete data (e.g. missing attribute of data, missing chunk of data) [Yes/No]
- Low labeled data volume (i.e. the amount of data that is labeled is lesser than the the amount of data that is unlabeled) [Yes/No]
- Data acquisition (i.e. purchase of data from third-party and processes associated with it) [Yes/No]
- Improper data transfer (e.g. mismanagement of the way data is transferred between source and destination) [Yes/No]
- Imbalanced dataset (e.g. data from a single location or single weather type dominates other locations or weathers) [Yes/No]
- Redundant data [Yes/No]
- Manual data collection [Yes/No]
- Manual data labeling [Yes/No]
- Expensive procedure [Yes/No]
- Reliance on suppliers to raise error (i.e. there is no internal process to raise an error, and raising error is dependent on the suppliers doing so) [Yes/No]
- Large volume of data makes it difficult to assess the quality of the data [Yes/No]
- Time consuming [Yes/No]
- Data ownership (i.e. who is the legal owner of the data and can use it without any prior approval?) [Yes/No]
- Wrongly-calibrated / defective sensor generates incorrect data [Yes/No]
- New data types from various sources make data integration difficult [Yes/No]
- Lack of variety in test environment causes AI to be poorly trained for situations that it has not been trained for [Yes/No]
- Data dependent on external conditions can be of low quality sometimes [Yes/No]
- Data fragmentation (i.e. data required by a function is located in different places) [Yes/No]
- Incompatible data formats [Yes/No]
- Outlier data (i.e. data is out of bounds of acceptable range) [Yes/No]
- Unstructured data [Yes/No]
- Noise (unwanted data that is mixed with valuable data) [Yes/No]
- Lack of good data from simulations (i.e. data from simulations are relatively of lesser quality than data from real-world collection) [Yes/No]
- Incorrect labeling [Yes/No]

Data Quality Challenges

The questions in this section attempt to understand the 'pressing'-ness of data quality challenges.

10. Please rank the following 'Data Availability' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*

    - Data delay (i.e. there is a delay in transmission of data from the source to the destination, e.g. data store to function, sensor to sensor)
    - Data drop (e.g. dropping of a required attribute of data, dropping a chunk of data during transmission)
    - Incomplete data (e.g. missing attribute of data, missing chunk of data)
    - Low labeled data volume (i.e. the amount of data that is labeled is lesser than the the amount of data that is unlabeled)

11. Please rank the following 'Data management' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*

- Data acquisition (i.e. purchase of data from third-party and processes associated with it)
- Improper data transfer (e.g. mismanagement of the way data is transferred between source and destination)
- Imbalanced dataset (e.g. data from a single location or single weather type dominates other locations or weathers)
- Redundant data
- Manual data collection and labeling
- Expensive procedure
- Reliance on suppliers to raise error (i.e. there is no internal process to raise an error, and raising error is dependent on the suppliers doing so)
- Large volume of data makes it difficult to assess the quality of the data
- Time consuming
- Data ownership (i.e. who is the legal owner of the data and can use it without any prior approval?)

12. Please rank the following 'Data source' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)

    - Wrongly-calibrated / defective sensor generates incorrect data
    - New data types from various sources make data integration difficult
    - Lack of variety in test environment causes AI to be poorly trained for situations that it has not been trained for
    - Data dependent on external conditions can be of low quality sometimes

13. Please rank the following 'Data structure' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*

    - Data fragmentation (i.e. data required by a function is located in different places)
    - Incompatible data formats
    - Outlier data (i.e. data is out of bounds of acceptable range)
    - Unstructured data

14. Please rank the following 'Data Trust' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*

    - Noise (unwanted data that is mixed with valuable data)
    - Lack of good data from simulations (i.e. data from simulations are relatively of lesser quality than data from real-world collection)
    - Incorrect labeling

15. Please rate these challenge sets (1 being the least pressing and 6 being the most pressing)

    - 'Data Availability' challenges (related to question 1) [1/2/3/4/5/6]
    - 'Data Management' challenges (related to question 2) [1/2/3/4/5/6]
    - 'Data Source' challenges (related to question 3)) [1/2/3/4/5/6]
    - 'Data Structure' challenges (related to question 4) [1/2/3/4/5/6]
    - 'Data Trust' challenges (related to question 5) [1/2/3/4/5/6]

Table 2 provides a list of data quality attributes that were presented in the Survey 2 questionnaire for validation of the association between data quality challenges and attributes. It also provides definitions in the survey questionnaire for respective data quality attributes.

Table 2: List of Data Quality Attributes and Their Definitions Provided in the Survey 2 Questionnaire

Data Quality Attribute	Definition of the Attribute Provided in the Survey Questionnaire
Accessibility	The extent to which data are available or easily and quickly retrievable
Access security	The extent to which access to data can be restricted and hence kept secure
Accuracy	The extent to which data are correct, reliable, and certified free of error
Auditability	It means that auditors can fairly evaluate data accuracy and integrity within rational time and manpower limits during the data use phase
Availability	The degree to which data can be consulted or retrieved by data consumers or processes
Completeness	Refers to whether all required data is present
Compliance	The degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use
Confidentiality	A property of data indicating the extent to which their unauthorised disclosure could be prejudicial or harmful to the interest of the source or other relevant parties
Consistency	Measures whether or not data is equivalent across systems or location of storage
Contact	Individual or organisational contact points for the data or metadata, including information on how to reach the contact points
Correctness	Every set of data stored represents a real world situation
Cost effectiveness	The extent to which the cost of collecting appropriate data is reasonable
Credibility	The extent to which data are trusted or highly regarded in terms of their source or content
Currency	The measure of whether data values are the most up-to-date version of the information
Ease of operation	The extent to which data are easily managed and manipulated (i.e., updated, moved, aggregated, reproduced, customized)
Efficiency	The degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use
Fitness	It has two-level requirements: 1) the amount of accessed data used by users and 2) the degree to which the data produced matches users' needs in the aspects of indicator definition, elements, classification, etc.
Flexibility	The extent to which data are expandable, adaptable, and easily applied to other needs
Frequency of dissemination	The time interval at which the statistics are disseminated over a given time period
Institutional mandate	Law, set of rules or other formal set of instructions assigning responsibility as well as the authority to an organisation for the collection, processing, and dissemination of statistics
Interpretability	The extent to which data are in an appropriate language and units and the data definitions are clear
Latency	The time between when the data was created and when it was made available for use
Lineage	Lineage measures whether factual documentation exists about where data came from, how it was transformed, where it went and end-to-end graphical illustration

Table 2: List of Data Quality Attributes and Their Definitions Provided in the Survey 2 Questionnaire

Data Quality Attribute	Definition of the Attribute Provided in the Survey Questionnaire
Portability	The degree to which data has attributes that enable it to be installed, replaced or moved from one system to another (while) preserving the existing quality in a specific context of use
Objectivity	The extent to which data are unbiased (unprejudiced) and impartial
Reasonability	Asks whether a data pattern meets expectations
Release policy	Rules for disseminating statistical data to all interested parties
Relevance	The extent to which data are applicable and helpful for the task at hand
Reliability	Reliability of the data, defined as the closeness of the initial estimated value to the subsequent estimated value
Representational consistency	The extent to which data are always presented in the same format and are compatible with previous data
Structure	It refers to the level of difficulty in transforming semi-structured or unstructured data to structured data through technology
Timeliness	Length of time between data availability and the event or phenomenon the data describe
Traceability	The extent to which data are well documented, verifiable, and easily attributed to a source
Understandability	The degree to which data has attributes that enable it to be read and interpreted by users, and are expressed in (an) appropriate languages, symbols and units in a specific context of use
Uniqueness	No entity exists more than once within the dataset
Usability	Is it understandable, simple, relevant, accessible, maintainable and at the right level of precision?
Usefulness	Extent to which information is applicable and helpful for the task at hand
Validity	Refers to whether data values are consistent with a defined domain of values
Variety	The extent to which data are available from several differing data sources

Data Availability Challenges - Data Quality Attributes

The goal of the questions in this section attempt is to understand if specific data availability challenges affect data quality attributes or not.

16. Does Data Delay (i.e. there is a delay in transmission of data from the source to the destination, e.g. data store to function, sensor to sensor) affect any of the following data quality attributes? (For e.g. read the question as in "is completeness of data affected by data delay?")

    - Accessibility [Yes/No]
    - Availability [Yes/No]
    - Completeness [Yes/No]
    - Currency [Yes/No]
    - Efficiency [Yes/No]
    - Latency [Yes/No]
    - Portability [Yes/No]
    - Timeliness [Yes/No]
    - Usefulness [Yes/No]

17. Does Data Drop (e.g. dropping of a required attribute of data, dropping a chunk of data during transmission) affect any of the following data quality attributes? (For e.g. read the question as in "is accuracy of data affected by data drop?")

    - Accessibility [Yes/No]

- Accuracy [Yes/No]
- Availability [Yes/No]
- Completeness [Yes/No]
- Consistency [Yes/No]
- Currency [Yes/No]
- Efficiency [Yes/No]
- Fitness [Yes/No]
- Flexibility [Yes/No]
- Objectivity [Yes/No]
- Portability [Yes/No]
- Reasonability [Yes/No]
- Reliability [Yes/No]
- Timeliness [Yes/No]
- Usefulness [Yes/No]

18. Does Incomplete Data (e.g. missing attribute of data, missing chunk of data) affect any of the following data quality attributes? (For e.g. read the question as in "is accuracy of data affected by incomplete data?")

    - Accuracy [Yes/No]
    - Availability [Yes/No]
    - Completeness [Yes/No]
    - Consistency [Yes/No]
    - Correctness [Yes/No]
    - Credibility [Yes/No]
    - Currency [Yes/No]
    - Efficiency [Yes/No]
    - Fitness [Yes/No]
    - Flexibility [Yes/No]
    - Objectivity [Yes/No]
    - Reasonability [Yes/No]
    - Reliability [Yes/No]
    - Understandability [Yes/No]
    - Usability [Yes/No]
    - Usefulness [Yes/No]

19. Does Low labeled data volume (i.e. the amount of data that is labeled is lesser than the the amount of data that is unlabeled) affect any of the following data quality attributes? (For e.g. read the question as in "is accuracy of data affected by low labeled data volume?")

    - Accuracy [Yes/No]
    - Availability [Yes/No]
    - Correctness [Yes/No]
    - Fitness [Yes/No]
    - Objectivity [Yes/No]
    - Usability [Yes/No]
    - Usefulness [Yes/No]

- Validity [Yes/No]

Data Management Challenges - Data Quality Attributes

The goal of the questions in this section attempt is to understand if specific data management challenges affect data quality attributes or not.

20. Does Data acquisition (i.e. purchase of data from third-party and processes associated with it) affect any of the following data quality attributes? (For e.g. read the question as in "is accessibility of data affected by data acquisition?")

    - Accessibility [Yes/No]
    - Availability [Yes/No]
    - Cost effectiveness [Yes/No]
    - Ease of operation [Yes/No]
    - Lineage [Yes/No]
    - Traceability [Yes/No]

21. Does Data ownership (i.e. who is the legal owner of the data and can use it without any prior approval?) affect any of the following data quality attributes? (For e.g. read the question as in "is accessibility of data affected by data ownership?")

    - Accessibility [Yes/No]
    - Auditability [Yes/No]
    - Compliance [Yes/No]
    - Confidentiality [Yes/No]
    - Ease of operation [Yes/No]
    - Lineage [Yes/No]
    - Traceability [Yes/No]

22. Does Imbalanced dataset (e.g. data from a single location or single weather type dominates other locations or weathers) affect any of the following data quality attributes? (For e.g. read the question as in "is correctness of data affected by imbalanced dataset?")

    - Correctness [Yes/No]
    - Efficiency [Yes/No]
    - Fitness [Yes/No]
    - Usability [Yes/No]
    - Usefulness [Yes/No]

23. Does Redundant data affect any of the following data quality attributes? (For e.g. read the question as in "is accuracy of data affected by redundant data?")

    - Accuracy [Yes/No]
    - Objectivity [Yes/No]
    - Uniqueness [Yes/No]
    - Usability [Yes/No]

24. Does Improper data transfer (e.g. mismanagement of the way data is transferred between source and destination) affect any of the following data quality attributes? (For e.g. read the question as in "is completeness of data affected by improper data transfer?")

    - Completeness [Yes/No]
    - Consistency [Yes/No]
    - Correctness [Yes/No]

- Currency [Yes/No]
- Ease of operation [Yes/No]
- Portability [Yes/No]
- Reliability [Yes/No]

25. Does Manual Data Collection affect any of the following data quality attributes? (For e.g. read the question as in "is accessibility of data affected by manual data collection?")

    - Accessibility [Yes/No]
    - Cost effectiveness [Yes/No]
    - Ease of operation [Yes/No]
    - Timeliness [Yes/No]

26. Does Manual Data Labeling affect any of the following data quality attributes? (For e.g. read the question as in "is timeliness of data affected by manual data labeling?")

    - Cost effectiveness [Yes/No]
    - Ease of operation [Yes/No]
    - Timeliness [Yes/No]

27. Does Regulatory Compliance affect any of the following data quality attributes? (For e.g. read the question as in "is confidentiality of data affected by regulatory compliance?")

    - Access security [Yes/No]
    - Compliance [Yes/No]
    - Contact [Yes/No]
    - Confidentiality [Yes/No]
    - Frequency of dissemination [Yes/No]
    - Institutional mandate [Yes/No]
    - Lineage [Yes/No]
    - Portability [Yes/No]
    - Release policy [Yes/No]
    - Traceability [Yes/No]

Data Source Challenges - Data Quality Attributes

The goal of the questions in this section attempt is to understand if specific data source challenges affect data quality attributes or not.

28. Does New data types from various sources affect any of the following data quality attributes? (For e.g. read the question as in "is relevance of data affected by new data types from various sources?")

    - Relevance [Yes/No]

29. Does Data dependent on external conditions affect any of the following data quality attributes? (For e.g. read the question as in "is accuracy of data affected by data dependent on external conditions?")

    - Accuracy [Yes/No]
    - Accessibility [Yes/No]
    - Correctness [Yes/No]

Data Structure Challenges - Data Quality Attributes

The goal of the questions in this section attempt is to understand if specific data structure challenges affect data quality attributes or not.

30. Does Incompatible data formats affect any of the following data quality attributes? (For e.g. read the question as in "is consistency of data affected by incompatible data formats challenge?")

    - Consistency [Yes/No]
    - Interpretability [Yes/No]
    - Validity [Yes/No]

31. Does Outlier data (i.e. data is out of bounds of acceptable range) affect any of the following data quality attributes? (For e.g. read the question as in "is consistency of data affected by outlier data challenge?")

    - Accuracy [Yes/No]
    - Correctness [Yes/No]
    - Credibility [Yes/No]
    - Efficiency [Yes/No]
    - Fitness [Yes/No]
    - Objectivity [Yes/No]

32. Does Unstructured data (i.e. data is out of bounds of acceptable range) affect any of the following data quality attributes? (For e.g. read the question as in "is consistency of data affected by outlier data challenge?")

    - Credibility [Yes/No]
    - Efficiency [Yes/No]
    - Representational consistency [Yes/No]
    - Structure [Yes/No]
    - Usability [Yes/No]
    - Validity [Yes/No]

Data Trust Challenges - Data Quality Attributes

The goal of the questions in this section attempt is to understand if specific data trust challenges affect data quality attributes or not.

33. Does Lack of good data from simulations (i.e. data from simulations are relatively of lesser quality than data from real-world collection) affect any of the following data quality attributes? (For e.g. read the question as in "is accuracy of data affected by lack of good data from simulations?")

    - Accuracy [Yes/No]
    - Credibility [Yes/No]
    - Fitness [Yes/No]
    - Objectivity [Yes/No]
    - Usefulness [Yes/No]
    - Variety [Yes/No]

34. Does Incorrect labeling affect any of the following data quality attributes? (For e.g. read the question as in "is accuracy of data affected by incorrect labeling?")

    - Accuracy [Yes/No]
    - Correctness [Yes/No]
    - Credibility [Yes/No]
    - Efficiency [Yes/No]
    - Fitness [Yes/No]
    - Objectivity [Yes/No]

- Reliability [Yes/No]
- Usability [Yes/No]
- Usefulness [Yes/No]
- Validity [Yes/No]

35. Does Noise (unwanted data that is mixed with valuable data) affect any of the following data quality attributes? (For e.g. read the question as in "is accuracy of data affected by noise?")

    - Accuracy [Yes/No]
    - Correctness [Yes/No]
    - Fitness [Yes/No]
    - Objectivity [Yes/No]
    - Usefulness [Yes/No]

4 Challenge Score

4.1 Challenge Score

The computation of the *Challenge Score* is based on the response from the survey conducted to rank the challenges. The survey contained two types of questions. One type of question asked the participants to provide a value of significance based on a Likert scale to five sets of challenges. Another type of question asked to rank individual challenges inside the five sets of challenges.

As there are two types of responses to two types of questions, they needed to be combined in some manner for both of them to be helpful. *Challenge Score* combines both types of responses in one final value. The value they provide for the comprehensive sets of challenges is recorded for each respondent. The highest-ranked challenge in a challenge set is given the highest numerical value corresponding to the number of challenges in that challenge set. Decreasing numerical values are assigned to remaining challenges in the particular challenge set. E.g., if there are four challenges in a challenge set, the highest-ranked challenge is given a value of 4, second-highest ranked is given a value of 3, and so on.

For each challenge, the assigned numerical value is multiplied by the value given by that particular participant for the challenge set of that particular challenge, which is done for all participants and challenges. The product values calculated for all participants for individual challenges are summed. The final *Challenge Score* is calculated by dividing this sum by the total number of challenges in the particular challenge set and further by dividing the result by the total number of participants, which is done to normalize the final value.

4.2 Survey 1

Table 3 depicts the ranking of data quality challenges given by participants in Survey 1 during the first cycle. Here, *S1-S6* are the six survey participants, *Total Score* is the sum of the product of rankings, and *Challenge Score* is the final normalized *Challenge Score*. The data is presented in descending order of the *Challenge Score*.

Note: Rows with gray background are the challenges removed from the final version of the artifact.

Table 3: Survey 1 - Challenge Score

Rank	Challenge Set	Challenge	S1	S2	S3	S4	S5	S6	Total Score	Challenge Score
1	Data Availability	Low Labeled Data Volume	4	1	4	4	4	4	104	4.333
2	Data Source	Lack of Variety in Test Environment	6	5	6	7	6	7	159	3.786
3	Data Availability	Incomplete Data	3	4	2	3	2	3	80	3.333
4	Data Source	Data Dependent on External Conditions	4	7	7	5	7	2	136	3.238
5	Data Management	Manual Data Labeling	10	3	8	10	9	9	193	3.217
6	Data Management	Imbalanced Dataset	7	10	1	9	10	10	178	2.967
7	Data Availability	Data Drop	2	3	3	1	3	2	68	2.833
8	Data Trust	Incorrect Labeling	5	1	4	5	5	4	80	2.667
9	Data Source	Wrongly-Calibrated / Defective Sensors	3	4	2	6	4	6	111	2.643
10	Data Trust	Uncertain Data Quality Identification	4	5	5	2	4	2	78	2.600
11	Data Management	Reliance on Suppliers to Raise Error	6	5	10	5	7	7	155	2.583
12	Data Management	Manual Data Collection	9	2	7	6	8	4	150	2.500
13	Data Source	Noise	5	6	4	3	1	5	103	2.452
14	Data Source	New Data Type	7	3	5	2	5	1	101	2.405
15	Data Management	Large Volume of Data	8	7	5	7	3	8	135	2.250
16	Data Trust	Wrong Metadata	2	4	3	4	2	3	61	2.033
17	Data Management	Regulatory Compliance	4	8	6	2	6	2	112	1.867
18	Data Structure	Unstructured Data	5	5	5	3	2	5	55	1.833
19	Data Management	Data Ownership	5	9	9	3	2	1	109	1.817

Table 3: Survey 1 - Challenge Score

Rank	Challenge Set	Challenge	S1	S2	S3	S4	S5	S6	Total Score	Challenge Score
20	Data Trust	Lack of Good Data from Simulation	3	3	2	3	3	1	54	1.800
21	Data Management	Improper Data Transfer	1	4	4	8	4	6	100	1.667
22	Data Availability	Data Delay	1	2	1	2	1	1	38	1.583
23	Data Source	Reliance on Single Data Source	1	1	3	4	3	4	66	1.571
24	Data Structure	Data Mix-up	1	1	3	1	5	3	47	1.567
25	Data Structure	Outlier Data	2	3	1	4	4	1	43	1.433
26	Data Trust	Fake Data	1	2	1	1	1	5	42	1.400
27	Data Management	Redundant Data	2	6	3	1	5	5	82	1.367
28	Data Structure	Incompatible Data Formats	4	4	4	4	1	2	40	1.333
28	Data Structure	Data Fragmentation	3	2	2	2	3	4	40	1.333
30	Data Source	Fast Increasing Data	2	2	1	1	2	3	52	1.238
31	Data Management	Data Acquisition	3	1	2	4	1	3	51	0.850

Note: In Table 3, *Expensive Procedure* and *Time Consuming* challenges are not included as they were identified during the second cycle.

Table 4 presents the values of Likert scale selected for each challenge set by the survey participants. Here, *S1-S6* are the six survey participants. The data is presented in alphabetical order of the challenge set.

Table 4: Survey 1 - Ranking of Challenge Sets

Challenge Set	S1	S2	S3	S4	S5	S6
Data Availability	6	4	5	5	6	3
Data Management	4	3	4	4	6	2
Data Source	5	5	2	3	6	5
Data Structure	2	2	3	1	6	1
Data Trust	3	5	1	2	6	4

4.3 Survey 2

Table 5 presents the ranking of data quality challenges given by participants in Survey 2 during the third cycle. Here, *S7-S10* are the four survey participants, *Total Score* is the sum of the product of rankings, and *Challenge Score* is the final normalized *Challenge Score*. The data is presented in descending order of the *Challenge Score*.

Note: Rows with gray background are the challenges added during the second cycle of the study.

Table 5: Survey 2 - Challenge Score

Rank	Challenge Set	Challenge	S7	S8	S9	S10	Total Score	Challenge Score
1	Data Availability	Low Labeled Data Volume	3	4	4	4	60	3.750
1	Data Trust	Incorrect Labeling	3	3	2	3	45	3.750
3	Data Source	Wrongly-Calibrated / Defective Sensors	4	2	4	3	58	3.625
3	Data Source	Lack of Variety in Test Environment	3	4	3	4	58	3.625
5	Data Availability	Incomplete Data	4	3	3	3	52	3.250
6	Data Management	Imbalanced Dataset	8	6	9	10	121	3.025
7	Data Source	Noise	2	1	3	1	35	2.917
8	Data Management	Large Volume of Data	4	8	10	4	106	2.650
9	Data Structure	Outlier Data	4	3	2	3	40	2.500
10	Data Structure	Incompatible Data Formats	3	2	4	1	39	2.438
11	Data Management	Manual Data Collection and Labeling	6	7	7	5	97	2.425
12	Data Management	Data Ownership	2	10	6	9	96	2.400
13	Data Management	Time Consuming	9	9	2	6	94	2.350
14	Data Management	Reliance on Suppliers to Raise Error	7	5	5	8	89	2.225
15	Data Availability	Data Drop	2	2	2	2	32	2.000
15	Data Structure	Data Fragmentation	1	4	1	4	32	2.000
17	Data Management	Improper Data Transfer	5	3	8	3	78	1.950
18	Data Source	Data Dependent on External Conditions	1	3	2	2	31	1.937
19	Data Management	Expensive Procedure	10	2	3	7	77	1.925
20	Data Trust	Lack of Good Data from Simulation	1	2	1	2	22	1.833
21	Data Structure	Unstructured Data	2	1	3	2	29	1.813
22	Data Source	New Data Type	2	1	1	1	23	1.438
23	Data Management	Data Acquisition	1	4	4	2	44	1.100
24	Data Availability	Data Delay	1	1	1	1	16	1.000
25	Data Management	Redundant Data	3	1	1	1	23	0.575

Table 5: Survey 2 - Challenge Score

Rank	Challenge Set	Challenge	S7	S8	S9	S10	Total Score	Challenge Score
26	Data Management	Regulatory Compliance						

Note:

(1) Due to limitations on the number of options provided by the survey tool used (Microsoft Forms), *Manual Data Collection* and *Manual Data Labeling* challenges were combined into a single challenge named *Manual Data Collection and Labeling* for ranking. They are still regarded as separate challenges in the *List of Challenges* artifact component.

(2) Due to technical error, *Regulatory Compliance* was not included in the second cycle survey. Hence, the calculation of *Challenge Score* ranking disregards it. It is only for calculation of the *Challenge Score*; the challenge is still included in the *List of Challenges* artifact component.

Table 6 presents the values of the Likert scale selected for each challenge set by the survey participants. Here, *S7-S10* are the four survey participants. The data is presented in alphabetical order of the challenge set.

Table 6: Survey 2 - Ranking of Challenge Sets

Challenge Set	S7	S8	S9	S10
Data Availability	4	1	6	5
Data Management	4	4	5	2
Data Source	6	3	4	4
Data Structure	3	4	5	2
Data Trust	6	2	6	3

5 Focus Group Data

5.1 Challenge Ranking

Tables 7, 8, 9, 10, and 11 present the data quality challenges and the number of participants who selected respective challenges as their preferred ranking (e.g., the number of items a particular challenge is selected as the 1st (i.e., top-most pressing), 2nd, and so on).

Table 7: Focus Group - *Data Availability* Challenges Ranking

Challenge	1st	2nd	3rd	4th
Data Delay	0	1	1	3
Data Drop	0	1	3	1
Incomplete Data	3	2	0	0
Low Labeled Data Volume	2	1	1	1

Table 8: Focus Group - *Data Management* Challenges Ranking

Challenge	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**	**8th**	**9th**	**10th**	**11th**	**12th**
Data Acquisition	0	0	0	0	3	0	0	1	0	0	0	0
Data Ownership	0	0	0	1	0	2	0	0	0	0	0	0
Expensive Procedure	1	0	1	1	0	0	1	0	0	1	0	0
Imbalanced Dataset	2	0	1	0	0	0	1	1	0	0	0	0
Improper Data Transfer	0	0	0	0	0	0	0	0	2	0	0	1
Large Volume of Data	0	2	0	1	1	0	0	0	0	0	0	0
Manual Data Collection	1	0	0	1	0	0	0	0	0	0	1	1
Manual data Labeling	1	2	0	0	0	0	1	0	0	0	1	0
Redundant Data	0	0	0	0	0	0	0	0	0	1	1	1
Regulatory Compliance	0	1	2	0	0	1	0	0	0	1	0	0
Reliance on Suppliers to Raise Error	0	0	0	1	0	1	0	0	1	0	0	0
Time Consuming	0	0	1	0	1	0	1	1	0	0	0	0

Table 9: Focus Group - *Data Source* Challenges Ranking

Challenge	1st	2nd	3rd	4th
Data Dependent on External Conditions	1	2	1	0
Lack of Variety in Test Environment	3	1	0	0
New Data Types	0	1	3	0
Wrongly-calibrated / Defective Sensor	0	0	0	4

Table 10: Focus Group - *Data Structure* Challenges Ranking

Challenge	1st	2nd	3rd	4th
Data Fragmentation	0	0	2	2
Incompatible Data Formats	0	3	0	1
Outlier Data	2	0	1	1
Unstructured Data	2	1	1	0

Table 11: Focus Group - *Data Trust* Challenges Ranking

Challenge	1st	2nd	3rd
Incorrect Labeling	3	0	1
Lack of Good Data from Simulations	1	3	0
Noise	0	1	3

Table 12 presents the number of times the challenge sets are given a particular value on a Likert scale. Here, *F1-F6* are the six focus group participants.

Table 12: Focus Group - Challenge Set Ranking

Challenge Set	F1	F2	F3	F4	F5	F6	Weighted Average
Data Availability	0	1	1	1	1	0	3.50
Data Management	0	1	0	1	2	0	4.00
Data Source	0	0	2	1	0	1	4.00
Data Structure	0	1	3	0	0	0	2.75
Data Trust	0	0	0	0	1	3	5.75

5.2 Data Quality Challenge - Attribute Association

Tables between Table 13 and Table 27 present the validation results from the focus group session. It includes the number of participants who selected *"yes, the challenge affects the attribute"* and *"no, the challenge does not affect the attribute."*

Data Availability Challenges

Table 13: *Data Delay* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Accessibility	2	2	0.5
Availability	1	3	0.25
Completeness	0	4	0
Currency	4	0	1
Efficiency	3	1	0.75
Latency	4	0	1
Portability	0	4	0

Table 13: *Data Delay* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Timeliness	4	0	1
Usefulness	2	2	0.5

Table 14: *Data Drop* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Accessibility	3	2	0.6
Accuracy	4	1	0.8
Availability	5	0	1
Completeness	4	1	0.8
Consistency	4	1	0.8
Currency	2	3	0.4
Efficiency	3	2	0.6
Fitness	2	3	0.4
Flexibility	1	4	0.2
Objectivity	1	4	0.2
Portability	1	4	0.2
Reasonability	2	3	0.4
Reliability	4	1	0.8
Timeliness	3	2	0.6
Usefulness	2	3	0.4

Table 15: *Incomplete Data* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Accuracy	5	0	1
Availability	3	2	0.6
Completeness	5	0	1
Consistency	3	2	0.6
Correctness	3	2	0.6
Credibility	5	0	1
Currency	0	5	0
Efficiency	1	4	0.2
Fitness	4	1	0.8
Flexibility	3	2	0.6
Objectivity	1	4	0.2
Reasonability	4	1	0.8
Reliability	4	1	0.8
Understandability	4	1	0.8
Usability	5	0	1
Usefulness	4	1	0.8

Table 16: *Low Labeled Data Volume* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Accuracy	4	1	0.8
Availability	1	4	0.2

Table 16: *Low Labeled Data Volume* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Correctness	3	2	0.6
Fitness	4	1	0.8
Objectivity	3	2	0.6
Usability	5	0	1
Usefulness	5	0	1
Validity	5	0	1

Data Management Challenges

Table 17: *Data Acquisition* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Accessibility	4	1	0.8
Availability	5	0	1
Cost Effectiveness	5	0	1
Ease of Operation	2	3	0.4
Lineage	5	0	1
Traceability	4	1	0.8

Table 18: *Imbalanced Dataset* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Correctness	5	0	1
Efficiency	1	4	0.2
Fitness	5	0	1
Usability	5	0	1
Usefulness	5	0	1

Table 19: *Improper Data Transfer* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Completeness	3	2	0.6
Consistency	3	2	0.6
Correctness	3	2	0.6
Currency	2	3	0.4
Ease of Operation	4	1	0.8
Portability	4	1	0.8
Reliability	4	1	0.8

Table 20: *Manual Data Collection* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Accessibility	1	4	0.2
Cost Effectiveness	5	0	1
Ease of Operation	3	2	0.6

Table 20: *Manual Data Collection* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Timeliness	1	4	0.2

Table 21: *Manual Data Labeling* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Cost Effectiveness	5	0	1
Ease of Operation	4	1	0.8
Timeliness	4	1	0.8

Table 22: *Redundant Data* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Accuracy	2	3	0.4
Objectivity	1	4	0.2
Usability	3	2	0.6
Usefulness	3	2	0.6

Data Source Challenges

Table 23: *Data Dependent on External Conditions* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Accessibility	5	0	1
Accuracy	3	2	0.6
Correctness	3	2	0.6

Data Structure Challenges

Table 24: *Outlier Data* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Accuracy	2	3	0.4
Correctness	0	5	0
Credibility	1	4	0.2
Efficiency	1	4	0.2
Fitness	1	4	0.2
Objectivity	1	4	0.2

Data Trust Challenges

Table 25: *Incorrect Labeling* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Accuracy	5	0	1
Correctness	5	0	1

Table 25: *Incorrect Labeling* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Credibility	5	0	1
Efficiency	1	4	0.2
Fitness	5	0	1
Objectivity	3	2	0.6
Reliability	5	0	1
Usability	5	0	1
Usefulness	5	0	1
Validity	5	0	1

Table 26: *Lack of Good Data from Simulations* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Accuracy	4	1	0.8
Credibility	3	2	0.6
Fitness	4	1	0.8
Objectivity	4	1	0.8
Usefulness	4	1	0.8
Variety	2	3	0.4

Table 27: *Noise* Challenge and Attributes Associated with it

Attribute	Yes	No	Weighted Average
Accuracy	5	0	1
Correctness	3	2	0.6
Fitness	3	2	0.6
Objectivity	1	4	0.2
Usefulness	3	2	0.6

6 Survey 2 Data

Here, *S7-S10* are the four survey participants.

6.1 Template Fields

Table 28: Template Fields Validation Result for *List of Challenges* Artifact Component

Field	S7	S8	S9	S10	No. of "Yes"	No. of "No"	Weighted Average
Name	Yes	Yes	Yes	Yes	4	0	1
Sources	Yes	Yes	Yes	No	3	1	0.75
Description	Yes	Yes	Yes	Yes	4	0	1
Whether the challenge directly affects AI function	Yes	Yes	Yes	Yes	4	0	1
Challenge Score	Yes	Yes	Yes	Yes	4	0	1

Table 29: Template Fields Validation Result for *List of Data Quality Attributes* Artifact Component

Field	S7	S8	S9	S10	No. of "Yes"	No. of "No"	Weighted Average
Name	Yes	Yes	Yes	Yes	4	0	1
Sources	No	Yes	Yes	No	2	2	0.5
Definition	Yes	Yes	Yes	Yes	4	0	1
Which challenges affect the data quality attribute?	Yes	Yes	Yes	Yes	4	0	1

Table 30: Template Fields Validation Result for *List of Data Quality Attribute Metrics* Artifact Component

Field	S7	S8	S9	S10	No. of "Yes"	No. of "No"	Weighted Average
Data quality attribute	Yes	Yes	Yes	Yes	4	0	1
Metric	Yes	Yes	Yes	Yes	4	0	1
Formula	Yes	Yes	Yes	Yes	4	0	1

Table 31: Template Fields Validation Result for *Potential Solutions* Artifact Component

Field	S7	S8	S9	S10	No. of "Yes"	No. of "No"	Weighted Average
Name	Yes	Yes	Yes	Yes	4	0	1
Requirements specifications	Yes	Yes	Yes	Yes	4	0	1
Implementation details	Yes	Yes	Yes	Yes	4	0	1

6.2 Survey Result - Challenges Directly Affecting AI Models

Table 32: List of Challenges Directly Affecting AI Models

Challenge	S7	S8	S9	S10	No. of "Yes"	No. of "No"	Weighted Average
Data Acquisition	Yes	Yes	No	No	2	2	0.5
Data Delay	No	Yes	No	No	1	3	0.25
Data Dependent on External Conditions	Yes	Yes	Yes	Yes	4	0	1
Data Drop	No	Yes	Yes	Yes	3	1	0.75
Data Ownership	No	Yes	No	No	1	3	0.25
Expensive Procedure	No	No	No	No	0	4	0
Fragmented Data	No	No	No	Yes	1	3	0.25
Imbalanced Dataset	Yes	Yes	Yes	Yes	4	0	1
Improper Data Transfer	No	Yes	Yes	Yes	3	1	0.75

Table 32: List of Challenges Directly Affecting AI Models

Challenge	S7	S8	S9	S10	No. of "Yes"	No. of "No"	Weighted Average
Incompatible Data Formats	No	No	Yes	No	1	3	0.25
Incomplete Data	No	Yes	Yes	Yes	3	1	0.75
Incorrect Labeling	Yes	Yes	Yes	Yes	4	0	1
Lack of Good Data from Simulations	Yes	Yes	Yes	Yes	4	0	1
Lack of Variety in Test Environment	Yes	Yes	Yes	Yes	4	0	1
Large Volume of Data	No	Yes	Yes	No	2	2	0.5
Low Labeled Data Volume	Yes	Yes	Yes	Yes	4	0	1
Manual Data Collection	Yes	Yes	No	No	2	2	0.5
Manual Data Labeling	Yes	Yes	No	No	2	2	0.5
New Data Types	Yes	Yes	No	No	2	2	0.5
Noise	Yes	No	Yes	No	2	2	0.5
Outlier Data	Yes	No	Yes	No	2	2	0.5
Redundant Data	Yes	No	Yes	No	2	2	0.5
Reliance on Suppliers to Raise Error	No	Yes	No	No	1	3	0.25
Time Consuming	No	Yes	No	No	1	3	0.25
Unstructured Data	No	No	Yes	No	1	3	0.25
Wrongly-Calibrated / Defective Sensor	Yes	Yes	Yes	Yes	4	0	1

6.3 Data Quality Challenge - Data Quality Attribute Association Survey Results

Table 33: List of Data Quality Challenge - Attribute Association Survey Validation Results

Challenge	Attribute	S7	S8	S9	S10	No. of "Yes"	No. of "No"	Weighted Average
Data Acquisition	Accessibility	Yes		Yes	No	2	1	0.66
Data Acquisition	Availability	Yes		Yes	No	2	1	0.66
Data Acquisition	Cost effectiveness	Yes		No	Yes	2	1	0.66
Data Acquisition	Ease of operation	Yes		No	No	1	2	0.33
Data Acquisition	Lineage	Yes		Yes	Yes	3	0	1
Data Acquisition	Traceability	Yes		Yes	Yes	3	0	1
Data Delay	Accessibility	Yes	No	Yes	No	2	2	0.5
Data Delay	Availability	Yes	No	Yes	No	2	2	0.5
Data Delay	Completeness	No	No	Yes	No	1	3	0.25
Data Delay	Currency	No	Yes	Yes	Yes	3	1	0.75
Data Delay	Efficiency	No	No	Yes	Yes	2	2	0.5
Data Delay	Latency	Yes	Yes	Yes	Yes	4	0	1
Data Delay	Portability	No	No	No	No	0	4	0
Data Delay	Timeliness	Yes	Yes	Yes	No	3	1	0.75
Data Delay	Usefulness	Yes	Yes	Yes	No	3	1	0.75
Data Dependent on External Conditions	Accessibility	Yes		Yes	Yes	3	0	1
Data Dependent on External Conditions	Accuracy	No		No	No	0	3	0
Data Dependent on External Conditions	Correctness	No		No	Yes	1	2	0.33
Data Drop	Accessibility	Yes	No	Yes	No	2	2	0.5
Data Drop	Accuracy	Yes	Yes	Yes	Yes	4	0	1
Data Drop	Availability	Yes	Yes	Yes	No	3	1	0.75
Data Drop	Completeness	Yes	Yes	Yes	Yes	4	0	1

Table 33: List of Data Quality Challenge - Attribute Association Survey Validation Results

Challenge	Attribute	S7	S8	S9	S10	No. of "Yes"	No. of "No"	Weighted Average
Data Drop	Consistency	Yes	Yes	Yes	Yes	4	0	1
Data Drop	Currency	Yes	No	No	No	1	3	0.25
Data Drop	Efficiency	Yes	No	Yes	No	2	2	0.5
Data Drop	Fitness	No	Yes	Yes	No	2	2	0.5
Data Drop	Flexibility	No	No	Yes	No	1	3	0.25
Data Drop	Objectivity	No	Yes	Yes	Yes	3	1	0.75
Data Drop	Portability	No	No	Yes		1	2	0.33
Data Drop	Reasonability	No	Yes	Yes	No	2	2	0.5
Data Drop	Reliability	Yes	Yes	Yes	Yes	4	0	1
Data Drop	Timeliness	Yes	No	No	No	1	3	0.25
Data Drop	Usefulness	Yes	Yes	Yes	No	3	1	0.75
Data Ownership	Accessibility	Yes		Yes	Yes	3	0	1
Data Ownership	Auditability	Yes		No	No	1	2	0.33
Data Ownership	Compliance	Yes		Yes	Yes	3	0	1
Data Ownership	Confidentiality	Yes		Yes	No	2	1	0.66
Data Ownership	Ease of operation	Yes		No	Yes	2	1	0.66
Data Ownership	Lineage	Yes		Yes	No	2	1	0.66
Data Ownership	Traceability	Yes		Yes	No	2	1	0.66
Imbalanced Dataset	Correctness	Yes		No	Yes	2	1	0.66
Imbalanced Dataset	Efficiency	No		No	No	0	3	0
Imbalanced Dataset	Fitness	Yes		No	Yes	2	1	0.66
Imbalanced Dataset	Usability	Yes		Yes	Yes	3	0	1
Imbalanced Dataset	Usefulness	Yes		Yes	Yes	3	0	1
Improper Data Transfer	Completeness	Yes		Yes	Yes	3	0	1
Improper Data Transfer	Consistency	Yes		Yes	No	2	1	0.66

Table 33: List of Data Quality Challenge - Attribute Association Survey Validation Results

Challenge	Attribute	S7	S8	S9	S10	No. of "Yes"	No. of "No"	Weighted Average
Improper Data Transfer	Correctness	Yes		Yes	No	2	1	0.66
Improper Data Transfer	Currency	Yes		Yes	Yes	3	0	1
Improper Data Transfer	Ease of operation	Yes		Yes	No	2	1	0.66
Improper Data Transfer	Portability	Yes		Yes	No	2	1	0.66
Improper Data Transfer	Reliability	Yes		Yes	No	2	1	0.66
Incompatible Data Formats	Consistency	Yes		Yes	No	2	1	0.66
Incompatible Data Formats	Interpretability	Yes		Yes	Yes	3	0	1
Incompatible Data Formats	Validity	Yes		Yes	No	2	1	0.66
Incomplete Data	Accuracy	Yes	Yes	Yes	Yes	4	0	1
Incomplete Data	Availability	Yes	Yes	Yes	No	3	1	0.75
Incomplete Data	Completeness	Yes	Yes	Yes	Yes	4	0	1
Incomplete Data	Consistency	Yes	Yes	Yes	Yes	4	0	1
Incomplete Data	Correctness	Yes	Yes	Yes	Yes	4	0	1
Incomplete Data	Credibility	Yes	Yes	Yes	Yes	4	0	1
Incomplete Data	Currency	Yes	Yes	Yes	No	3	1	0.75
Incomplete Data	Efficiency	Yes	Yes	Yes	No	3	1	0.75
Incomplete Data	Fitness	Yes	Yes	Yes	Yes	4	0	1
Incomplete Data	Flexibility	No	No	No	No	0	4	0
Incomplete Data	Objectivity	Yes	Yes	Yes	Yes	4	0	1
Incomplete Data	Reasonability	No	Yes	Yes	No	2	2	0.5
Incomplete Data	Reliability	Yes	Yes	Yes	Yes	4	0	1

Table 33: List of Data Quality Challenge - Attribute Association Survey Validation Results

Challenge	Attribute	S7	S8	S9	S10	No. of "Yes"	No. of "No"	Weighted Average
Incomplete Data	Understandability	No	Yes	Yes	Yes	3	1	0.75
Incomplete Data	Usability	Yes	Yes	No	No	2	2	0.5
Incomplete Data	Usefulness	No	Yes	Yes	No	2	2	0.5
Incorrect Labeling	Accuracy	Yes		Yes	Yes	3	0	1
Incorrect Labeling	Correctness	Yes		Yes	Yes	3	0	1
Incorrect Labeling	Credibility	Yes		Yes	Yes	3	0	1
Incorrect Labeling	Efficiency	Yes		Yes	No	2	1	0.66
Incorrect Labeling	Fitness	Yes		Yes	Yes	3	0	1
Incorrect Labeling	Objectivity	Yes		Yes	Yes	3	0	1
Incorrect Labeling	Reliability	Yes		Yes	Yes	3	0	1
Incorrect Labeling	Usability	Yes		Yes	Yes	3	0	1
Incorrect Labeling	Usefulness	Yes		Yes	Yes	3	0	1
Incorrect Labeling	Validity	Yes		Yes	Yes	3	0	1
Lack of Good Data from Simulations	Accuracy	Yes		Yes	No	2	1	0.66
Lack of Good Data from Simulations	Credibility	Yes		Yes	Yes	3	0	1
Lack of Good Data from Simulations	Fitness	Yes		Yes	No	2	1	0.66
Lack of Good Data from Simulations	Objectivity	Yes		Yes	Yes	3	0	1
Lack of Good Data from Simulations	Usefulness	Yes		Yes	No	2	1	0.66
Lack of Good Data from Simulations	Variety	Yes		Yes	Yes	3	0	1
Low Labeled Data Volume	Accuracy	Yes	Yes	Yes	Yes	4	0	1

Table 33: List of Data Quality Challenge - Attribute Association Survey Validation Results

Challenge	Attribute	S7	S8	S9	S10	No. of "Yes"	No. of "No"	Weighted Average
Low Labeled Data Volume	Availability	No	Yes	No	No	1	3	0.25
Low Labeled Data Volume	Correctness	No	Yes	Yes	Yes	3	1	0.75
Low Labeled Data Volume	Fitness	Yes	Yes	Yes	Yes	4	0	1
Low Labeled Data Volume	Objectivity	No	Yes	Yes	Yes	3	1	0.75
Low Labeled Data Volume	Usability	No	Yes	Yes	Yes	3	1	0.75
Low Labeled Data Volume	Usefulness	No	Yes	Yes	Yes	3	1	0.75
Low Labeled Data Volume	Validity	No	Yes	Yes	Yes	3	1	0.75
Manual Data Collection	Accessibility	Yes		No	Yes	2	1	0.66
Manual Data Collection	Cost effectiveness	Yes		No	Yes	2	1	0.66
Manual Data Collection	Ease of operation	Yes		No	Yes	2	1	0.66
Manual Data Collection	Timeliness	Yes		No	Yes	2	1	0.66
New Data Types from Various Sources	Relevance	No		Yes	No	1	2	0.33
Noise	Accuracy	Yes		Yes	No	2	1	0.66
Noise	Correctness	Yes		Yes	No	2	1	0.66
Noise	Fitness	Yes		Yes	No	2	1	0.66
Noise	Objectivity	Yes		Yes	No	2	1	0.66

Table 33: List of Data Quality Challenge - Attribute Association Survey Validation Results

Challenge	Attribute	S7	S8	S9	S10	No. of "Yes"	No. of "No"	Weighted Average
Noise	Usefulness	Yes		Yes	No	2	1	0.66
Outlier Data	Accuracy	Yes		Yes	No	2	1	0.66
Outlier Data	Correctness	Yes		No	No	1	2	0.33
Outlier Data	Credibility	Yes		Yes	Yes	3	0	1
Outlier Data	Efficiency	Yes		No	No	1	2	0.33
Outlier Data	Fitness	Yes		Yes	Yes	3	0	1
Outlier Data	Objectivity	Yes		No	No	1	2	0.33
Redundant Data	Accuracy	Yes		No	No	1	2	0.33
Redundant Data	Objectivity	Yes		No	No	1	2	0.33
Redundant Data	Uniqueness	Yes		Yes	Yes	3	0	1
Redundant Data	Usability	Yes		No	No	1	2	0.33
Regulatory Compliance	Access security	Yes		Yes	No	2	1	0.66
Regulatory Compliance	Compliance	Yes		Yes	Yes	3	0	1
Regulatory Compliance	Confidentiality	Yes		Yes	No	2	1	0.66
Regulatory Compliance	Contact	Yes		Yes	No	2	1	0.66
Regulatory Compliance	Frequency of dissemination	Yes		Yes	No	2	1	0.66
Regulatory Compliance	Institutional mandate	Yes		Yes	Yes	3	0	1
Regulatory Compliance	Lineage	Yes		Yes	No	2	1	0.66
Regulatory Compliance	Portability	Yes		Yes	No	2	1	0.66

Table 33: List of Data Quality Challenge - Attribute Association Survey Validation Results

Challenge	Attribute	S7	S8	S9	S10	No. of "Yes"	No. of "No"	Weighted Average
Regulatory Compliance	Release policy	Yes		Yes	Yes	3	0	0
Regulatory Compliance	Traceability	Yes		Yes	No	2	1	0.66
Unstructured Data	Credibility	No		No	No	0	3	0
Unstructured Data	Efficiency	Yes		No	Yes	2	1	0.66
Unstructured Data	Representational consistency	Yes		Yes	Yes	3	0	1
Unstructured Data	Structure	Yes		Yes	No	2	1	0.66
Unstructured Data	Usability	Yes		No	No	1	2	0.33
Unstructured Data	Validity	Yes		No	No	1	2	0.33