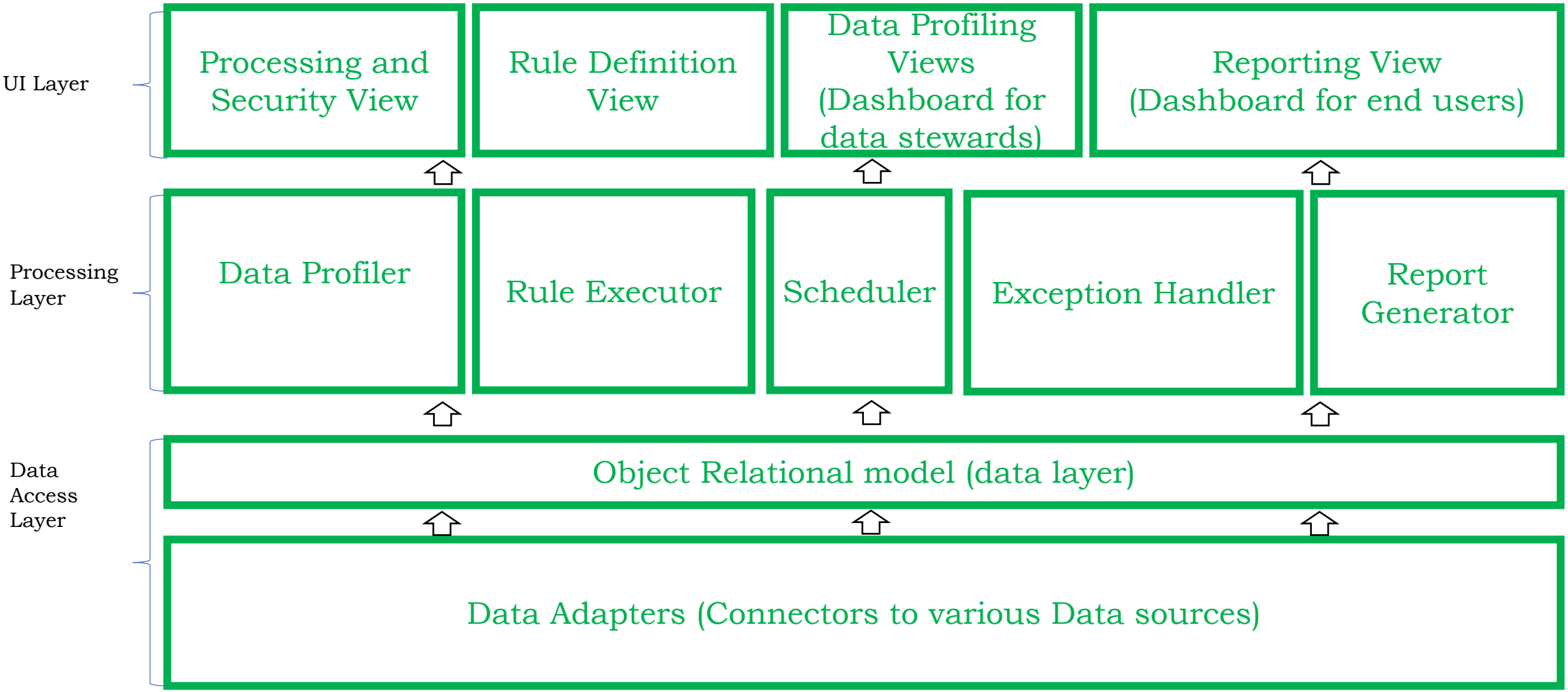
VRAIFUTURE

DATA QUALITY PRODUCT
Minimum Viable Product

Vinoth Raman | Amsterdam | 02 Apr 2019

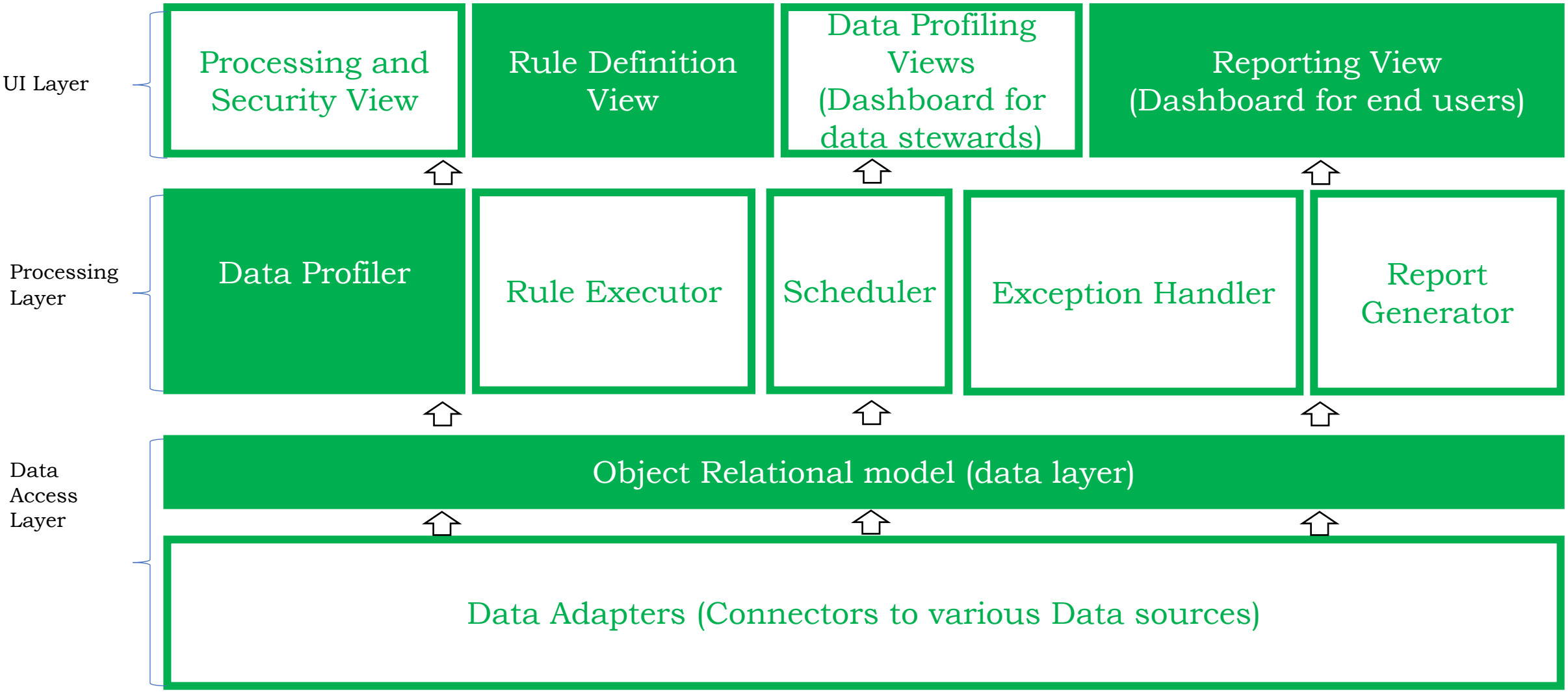
VRAIFUTURE

DQ Tool - Architecture



VRAIFUTURE

DQ Tool – Architecture – Components for MVP



What is the MVP?

MVP

- CSV file of flights will be used.
- This CSV file needs to be loaded into the data layer in our proprietary format. No data loss should occur. If there are issues with loading the entire file, the file should be rejected and all loaded data should be removed. Partial loading can be looked at at a later stage.(No copy of the data will be stored)
- Based on the columns DQ suggestions should be generated in the data profiler module abd should be available for the users via the rule definition view module.
    - All the columns are loaded with business names:
        - Ex: DAY_OF_WEEK is loaded as "Day of the week" . ORIGIN_AIRPORT stored as "Origination Airport" . In this file, the names are logical. But if some files have columns that are technical, we need to convert them into business savvy names.
    - Column type is identified (Date, Airline, Flight number, etc.)
    - Based on the column type (integer, range of values, dates, amount, etc.), we need to come up with the generic DQ rule suggestions.
        - Departure time is before the arrival time
        - Date (Day) should not be 32
        - Origination Airport and Destination Airport cannot be the same
- Based on the user defined rules, the checks are executed and dashboard is generated for the end user, which is visible in the Dashboard/Reporting View

VRAIFUTURE

DQ Rules

DQ rules

DATE

- Year should be 2015 or a range (if 99% of the date is 2015 – then we set it like this) : User can pick a range or a definitive year
- Month should be 1 to 12
- Day should be 1 to 31 (with the exceptions of 30, 28 and 29)

AIRLINE

- Should be Alpha and 2 characters.
- There should be an approved list of values (a separate file which we can load for this particular column). All values in the file will be compared against this and if the value is not in the approved list, then it is marked as suspect.

FLIGT NUMBER

- Numeric. Can be 2,3 and 4 characters.
- Combination of Airline flight number should be correct (Level 3 check). EV should have 4 digit flight number.

ORIGIN / DESTINATION:

- Should be Alpha and 3 characters.
- There should be an approved list of values (a separate file which we can load for this particular column). All values in the file will be compared against this and if the value is not in the approved list, then it is marked as suspect.
- Origin and Destination cannot be the same (Level 3 check)

DQ Rules

DQ rules

DEPARTURE / ARRIVAL:

- ”DEPARTURE DELAY” should be the difference between “SCHEDULED DEPARTURE” and “DEPARTURE TIME” .
- If ”DEPARTURE DELAY” is more than 20, we need to flag that as analysis needed.
- ”WHEELS OFF” time cannot be before “DEPARTURE TIME”

- The above combination applies to ARRIVAL as well.

- For all the above columns, basic checks like numeric should be available.

OTHERS

- If A flight is canceled, then there should not be a data for departure time and arrival time
- If a flight is diverted, then there should be details on different arrival time, destination, etc.
- If there is a delay, then the total of reasons (airline, security, airsystem, late aircraft, weather) should be equal so all delays are explained.

Object relational model (data layer)

FR's

The csv sheets should be processed. Compressed and be ready to be consumed by Data profiler module.

NFR's

Usability – The module should be consistent with every file load and reload

Reliability – No data should be lost while loading

Performance – High speed is required.

Scalability – At this moment, the csv file with tens/hundreds of thousands of rows should be consumed if necessary in a short time.

If for whatever reason, if the whole file is not consumed, then we should have the option to roll back or clear the entre content so that the file can be uploaded again.

UXR's

The module should be available as Click and Go option. User should be able to point to the csv file's location and the system should do validation and accept/reject the file processing.

Clear feedback should be given if the file was processed correctly or not.

Data Profiler

FR's

The CSV file should be analyzed.
- The columns should be identified, type, possible business names based on the data.
- AUTO DQ rules should be generated based on the column and data set and should be available in the rule definition view.
- Anomalies should be detected.
- Exceptions should be detected.

Similar files should be identified and past rules should be re-used or updated based on the data set.

NFR's

Usability –
Reliability – Machine learning models should be highly accurate

Performance – High speed is required.

Scalability – At this moment, the csv file with tens/hundreds of thousands of rows should be consumed if necessary in a short time.

UXR's

Rule Definition View

FR's

Business user should be able to define DQ rules on a highly intuitive basis.
Many rules are already either automatically available or suggested based on the analysis of data profiler based on the input data.

NFR's

Usability – Should be very intuitive and a business user should be able to navigate without much issues.

Reliability – The module should be highly reliable in capturing the rules as they are the basis for the DQ checks.

Performance – System should perform well even if many concurrent users are defining or viewing the rules.

Scalability – When the DQ rules keep on increasing, smart categorization should be used so that the system is still very user friendly.

UXR's

UX experience should be very intuitive. So that all sizes of companies with their talent be able to use the tool

Reporting View / Dashboard for end users

FR's

Dashboards are the end results. Should be standard, consistent, easily interpretable and actionable.
Different dashboards for different levels of management/users need to be created.

NFR's

Usability – Ease of use. Should not be very complex.

Reliability – Should be highly reliable. Accuracy should be high so that people can take decisions and actions based on this.

Performance – Highly performant for many users and also for the interactions.

Scalability – When the data volume increases, the dashboards should scale and also when the user numbers increase.

UXR's

UX experience should be very intuitive. All kind of users should be longing to use the dashboard to understand the data quality

DQ Dashboards : Operational

DQ dashboards are not only for information but should be actionable.

Airline	Timeliness	Completeness	Validity	Accuracy	Uniqueness	Consistency
AA	100%	100%	90%	90%	80%	70%
DL	100%	100%	100%	100%	90%	80%
EV	100%	100%	100%	100%	100%	80%
MQ	100%	100%	100%	100%	80%	80%
US	100%	100%	100%	100%	60%	80%

DQ report on airline level CDEs.

This is an operational dashboard for giving feedback to airlines about the data quality issues.
Assumption: Airlines are sending their respective data.

Airlines	Jan 2015	Feb 2015	Mar 2015	Apr 2015	May 2015	Jun 2015
AA	100%	100%	90%	90%	80%	70%
DL	100%	100%	100%	100%	90%	80%
EV	100%	100%	100%	100%	100%	80%
MQ	100%	100%	100%	100%	80%	80%
US	100%	100%	100%	100%	60%	80%

DQ report on airline level CDEs.

This is the same as above but analyzing the DQ trend

DQ Dashboards : Management

DQ dashboards are not only for information but should be actionable.

WORK IN PROGRESS

Appendix

Data Quality Tool - Requirements

DQ Tool

Measurement

Ability to capture Critical Data Elements (CDE)
Tool should come up with automatic relevant checks for CDE.
Data quality can be measured at any point of the data lifecycle but the optimal place is at the source systems.
The critical task of DQ tool is to measure the data quality along the lines of data quality dimensions
Interface status report
Service level agreement breach report

Reporting

Various data quality dashboards needed to be created for various roles such as data owner, data user, management, etc.
Various reports need to be created for various user groups

Configuration

Business users should be able to add rules. Categorization is based on the data quality dimensions
New categories can be added

Integration

DQ tool should be system agnostic.
Should be able to integrate into another system or work next to it.
DQ tool should be able to handle all kind of inputs (flat files, XML, real time data, etc.)

Data Quality Tool - Requirements

DQ Tool

Remediation

Reconciliation report
Duplicates report

Data Quality Rules

Data quality rules can be captured in any of the system of the data lifecycle. In a local system as well as in a global system. A global system should be aware of DQ rules of local where needed.

Data quality agreements should be used as a source for the data quality rules. Data quality measurements of source and destination should reconcile with accepted deviations. Ability to read the data quality agreement and transform them into data quality rules with machine learning will be a great addition.

Data Quality Reporting

DQ Reporting

Appropriateness and Timeliness

Critical data elements report and their impact on business
Interface status report
Service level agreement breach report

Completeness

% complete report for Critical data elements and other key data attributes

Validity

Review of data elements
% correctness against the expected values.

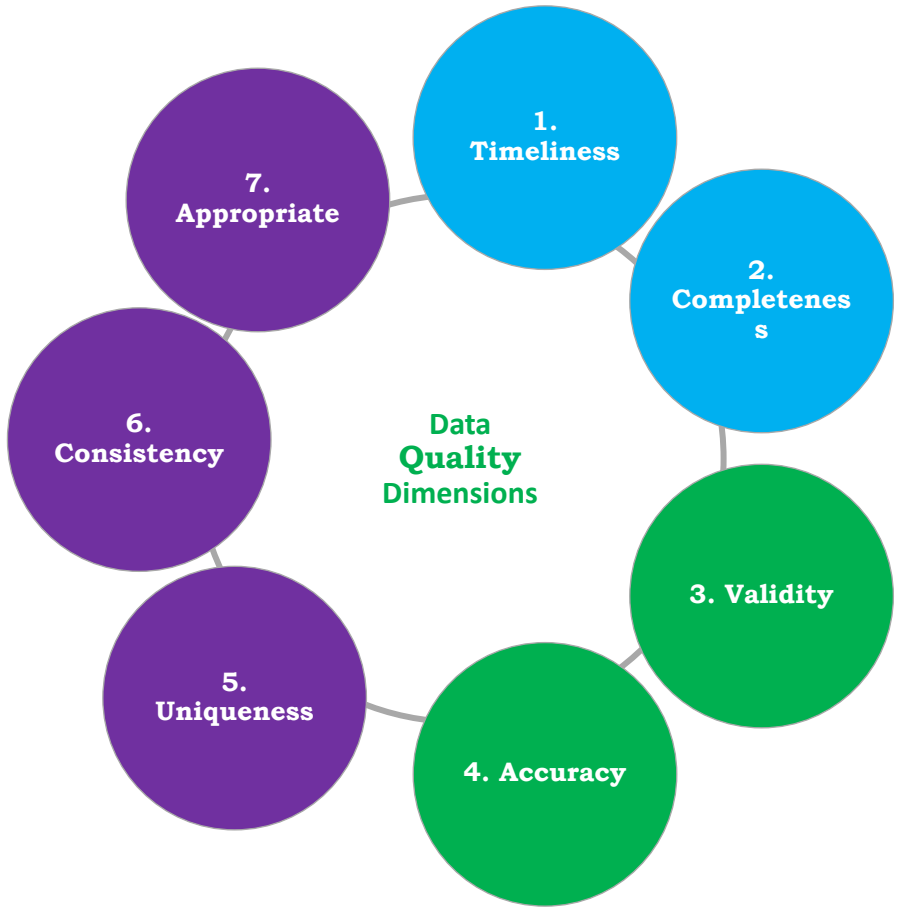
Accuracy

Dependency report (CDE is depending on what and the issues over there)
Correlation report (Correlation between the CDEs and the issues related to them)
Trend analysis report (Accuracy with respect to previous period)

Uniqueness and Consistency

Duplication suspect report
Reconciliation report

Data Quality Dimensions for MVP



	Data Dimensions	Examples of Checks
Pre	✓ **Timeliness**	• Timeliness checks
Pre	✓ **Completeness**	• Completeness checks • Comprehensiveness checks • Coverage
Controls	✓ **Validity**	• Integrity checks • Optionality checks
Controls	✓ **Accuracy**	• Internal Consistency checks • Correctness checks • Plausibility checks
Post	✓ **Uniqueness**	• Deduplication
Post	**6. Consistency**	• Reconciliation
Post	**6. Appropriate**	• Data is appropriate for the business purpose

Pre Check | Run time Check | Post Check

Data Quality Dimensions for MVP - Details

DQ Dimensions

Timeliness
Whether data is delivered timely as outlined in the data delivery agreement and is available when needed

Completeness
Whether data set is complete or not. All the Critical data elements (key data attributes) are delivered

Accuracy
Data that is delivered is accurate and unambiguous

Integrity
Only allowed values are provided for the key data attributes

Consistency
Data is consistent if measured against time and related key data attributes

Appropriateness (Validity, Usefulness)
Data that is delivered is appropriate for the data usage – “Fit for the purpose”

Comprehensiveness
Dependent, atomic data attributes are delivered as well

Adaptability
Data is delivered at atomic level so that it is easy to adapt the derived data.

Data Quality Dimensions – Additional ones

Dimension	Description
Accuracy	Refers to the degree that data correctly represents the “real-life” entities they model
Completeness	Certain attributes always have assigned values in a data set. All appropriate rows in a dataset are present
Consistency	Refers to ensuring that data values in one data set are consistent with values in another data set
Currency	Refers to the degree to which information is current with the world that it models
Precision	Refers to the level of detail of the data element
Privacy	Refers to the need for access control and usage monitoring
Reasonableness	To consider consistency expectations relevant within specific operational contexts
Referential Integrity	Is the condition that exists when all intended references from data in one column of a table to data in another column of the same or different table is valid
Timeliness	Refers to the time expectation for accessibility and availability of information
Uniqueness	States that no entity exists more than once within the data set
Validity	Refers to whether data instances are stored, exchanged, or presented in a format that is consistent with the domain of values, as well as consistent with other similar attribute values

DQ Checks - Examples

Timeliness:

Customer data should be available before making a marketing campaign
Data for monthly reporting should be available before the 2nd working day of the next month

Completeness:

All the CDE and other mandatory attributes are having values. Ex: Customer name, address, email and telephone
The delivered key attributes should be the same at a given point of time with respect to the source system.

Validity:

Invalid format (Amount field contains more than required decimal places)
Data should be one of the pre-defined field (Country code, zip codes)
If wedding anniversary field is filled, marriage date cannot be empty
Negative shoe size

Accuracy:

The customers are classified correctly (prime , high potential, etc.)
If there is a discount price , then it should be less than the original sale price
The wedding anniversary date should be after the marriage date
Prime customers should be paying a recurring fee

Uniqueness:

There are two shoes with same barcode in two different shops of Netherlands
There are two people in the database with same date of birth, place, etc. with one spelling mistake (can be true but needs to be reviewed)

Consistency:

The address of a customer is received from two source systems and they both are different. This is not consistent.
The inventory of a particular shoe is consolidated across Netherlands and checked against delivery and it should match.

DQ Measurement – Data dimension complexities

Level	DQ Dimension	Examples
-1	Appropriateness	Data fit for the purpose
0	Timeliness	Data delivered on time
1	Completeness	All critical data elements are delivered
2	Integrity	Data delivered with right format, list of values, range of values
3	Accuracy, Consistency	Correlated data is correct. Consistent value over time
4	……	……

- Level 1 and 2 can be automated with tools to a great extent
- Level 3 suspects need further analysis with experts prior to deciding if it is of data quality issues or the difference is genuine
- Accuracy is covered partly by Level 2 checks and partly by Level 3 checks. Still accuracy cannot be checked by rules and the data owner always needs to ensure the accuracy of data

Why levels?

As you go up in the levels, the complexity of data quality rules and controls increase. If everything is working well, most probably all the level -1, 0 , 1 and 2 checks should pass 100% (or the minimum required quality requirements) and level 3 checks should pass around 99% with 1% suspect data in it. It is also easy then for the management to escalate by monitoring the dashboards in different levels.