Informatica

19 Nov, 2024

How to Get Started with DQ?

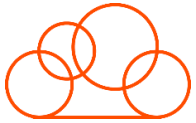
- Zoey Husband, Principal Consultant, EMEA Advisory

Where data & AI come to LIFE

Housekeeping Tips

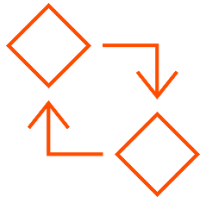
- Today's Webinar is scheduled for 1 hour
- The session will include a webcast and then your questions will be answered live at the end of the presentation
- All dial-in participants will be muted to enable the speakers to present without interruption
- Questions can be submitted to "All Panelists" via the Q&A option and we will respond at the end of the presentation
- The webinar is being recorded and will be available on our Success Portal - where you can download the slide deck for the presentation. The link to the recording will be emailed as well.
- Please take time to complete the post-webinar survey and provide your feedback and suggestions for upcoming topics.

Feature Rich Success Portal



Bootstrap trial and POC Customers

Enriched Customer Onboarding experience



Product Learning Paths and Weekly Expert Sessions

Informatica Concierge

Tailored training and content recommendations

More Information

Success Portal

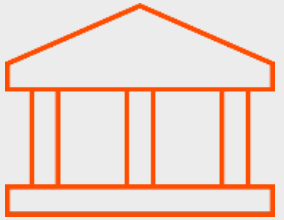
https://success.informatica.com

Communities & Support

https://network.informatica.com

Documentation

https://docs.informatica.com



University

https://www.informatica.com/in/services-and-training/informatica-university.html

Safe Harbor

The information being provided today is for informational purposes only. The development, release, and timing of any Informatica product or functionality described today remain at the sole discretion of Informatica and should not be relied upon in making a purchasing decision.

Statements made today are based on currently available information, which is subject to change. Such statements should not be relied upon as a representation, warranty or commitment to deliver specific products or functionality in the future.

Informatica

How to get started in DQ

Zoey Husband

Principal Consultant – Advisory Services EMEA

Where data & AI come to LIFE

Agenda

1. Introduction
2. What is Data Quality?
3. How to encourage engagement from the business?
4. DQ dimensions
5. Methods of defining initial DQ rules
6. Example rules
7. Roadmap for maturing the capability
8. Q&A

Informatica

"Data quality is the key to the success of Fannie Mae's mission: getting the right people into the right homes."
- IT Director Fannie Mae

Quality data is a prerequisite for quality predictive models.
Microsoft

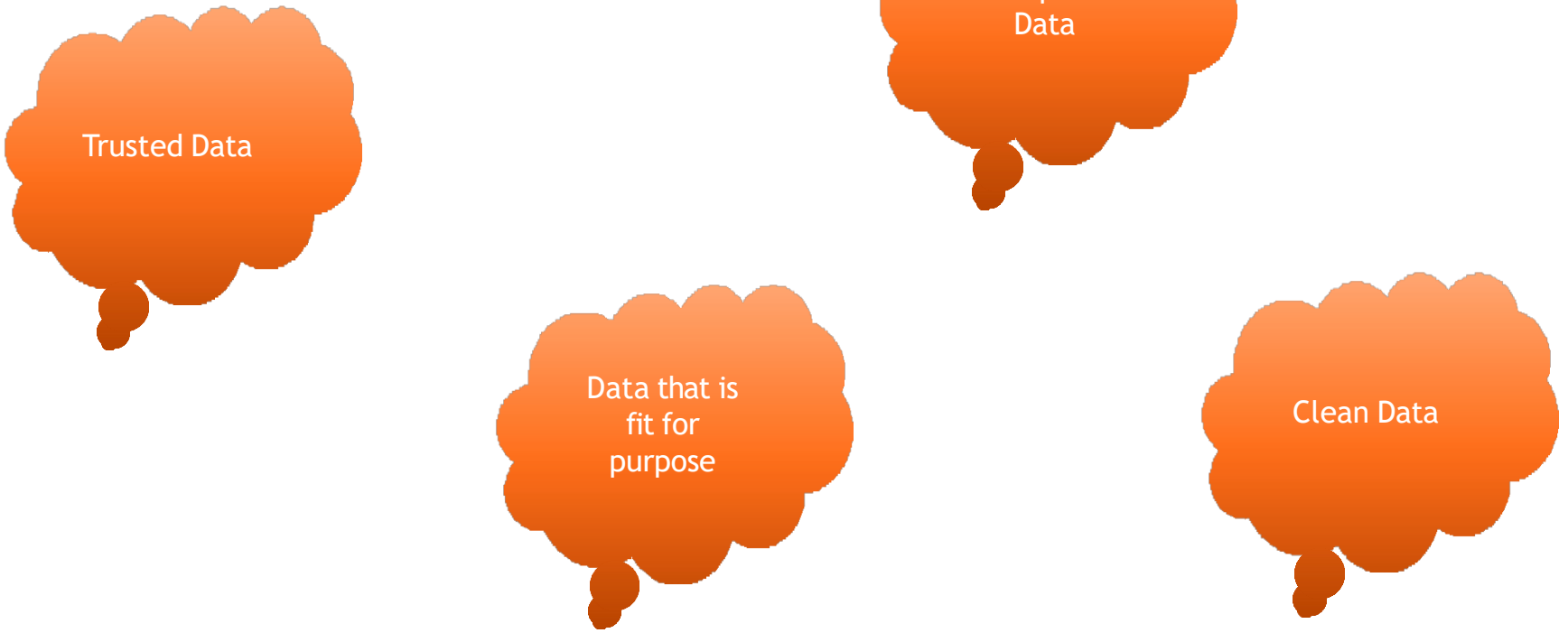
"Only 38 percent of decisionmakers have a high level of confidence in their customer insights."
Survey: KPMG

"If Your Data Is Bad, Your Machine Learning Tools Are Useless." - Thomas Redman, Harvard Business Review

Cross-Industry Cost of Poor Quality Data

	Symptoms	Impact
Sales/Marketing	▪ Low customer satisfaction ▪ Many address change requests ▪ No trust/agreement in reporting	▪ Customer churn ▪ Excessive mailing expense ▪ Opportunity cost of lost sales
Finance	▪ Budgets take forever to get "right" ▪ Significant budget discrepancies ▪ No trust/agreement in reporting	▪ Budget overruns ▪ Out-of-control expenses ▪ Slow decision making. Fines!
Supply Chain	▪ "Out of stock" situations ▪ Poor quality products ▪ No trust/agreement in reporting	▪ Reduced top-line revenue ▪ Customer satisfaction ▪ Supply chain inefficiencies
IT	▪ Large IT projects fail ▪ Low use of applications ▪ No trust/agreement in reporting	▪ IT investments wasted ▪ Productivity remains low ▪ Wasted bus. Analyst time
Business/Data Analyst Group	▪ More time preparing vs analyzing ▪ Poor timeliness of data	▪ Hidden & significant waste of expensive analyst time ▪ Competitors move faster

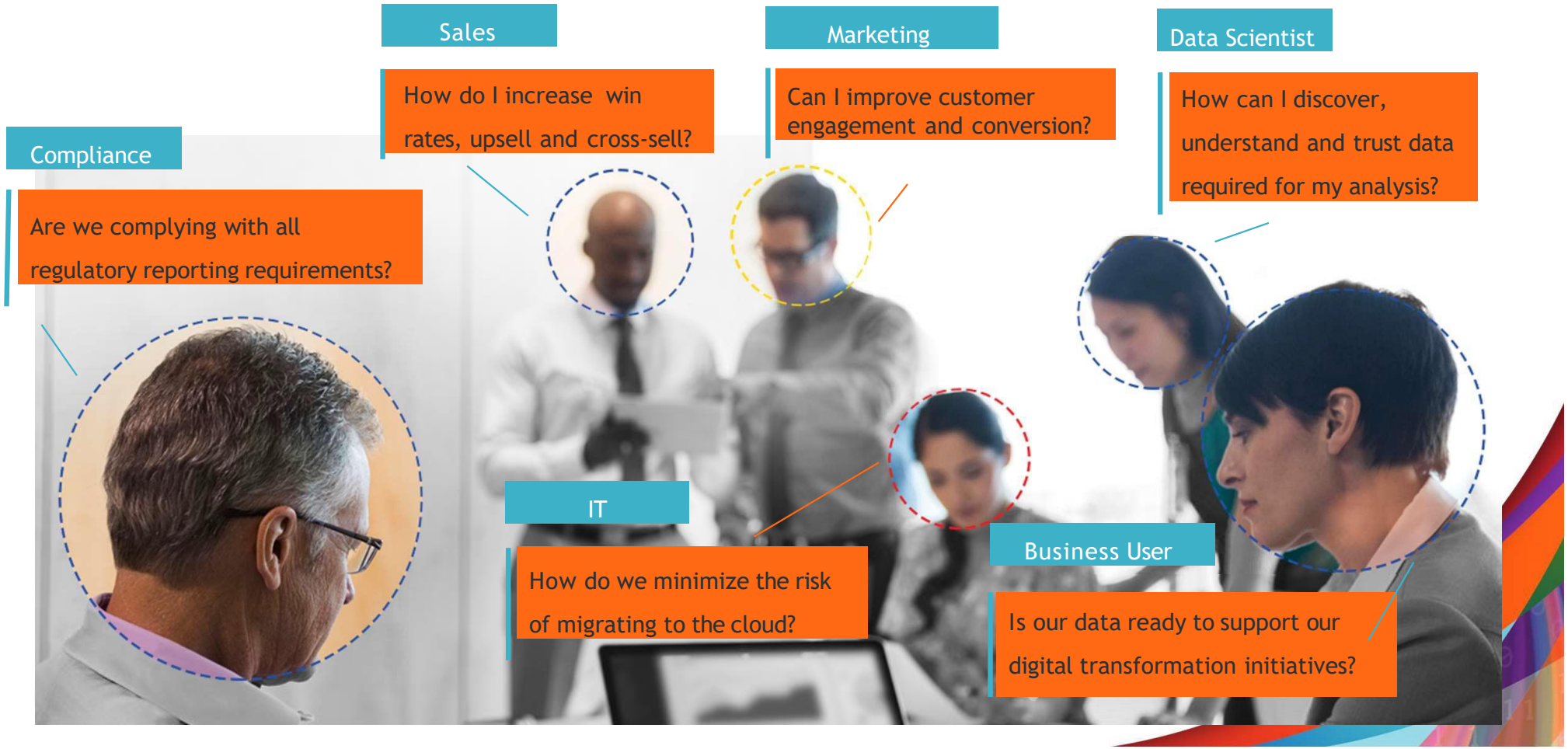
What is Data Quality?



What is Data Quality?

Informatica

How to encourage engagement from the business?



How to encourage engagement from the business?

Understand your why

- Be clear on how your DQ aims relate back to business strategy
- Choose an area where:
  - value can be quantified
  - there are willing stakeholders

Inclusion

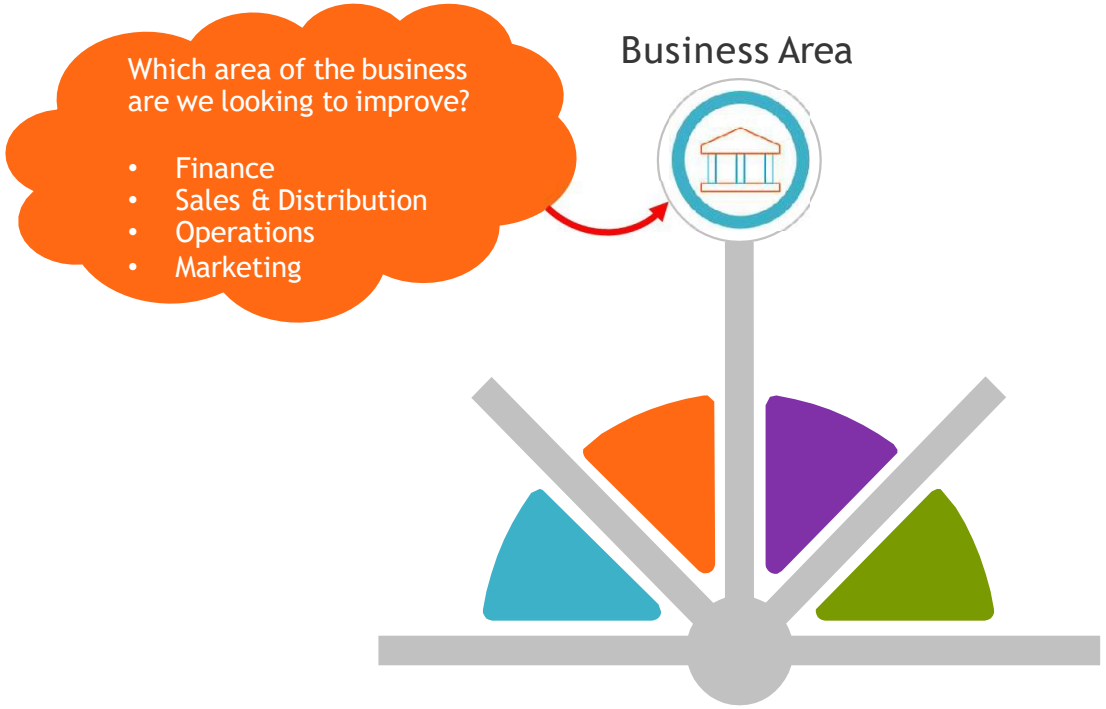
- The business SME know their data, you need their input

Quick Wins

- Choose an area where:
  - you have the ability to do something about any issues found
  - the outcome has a solid “so what” that the business cares about

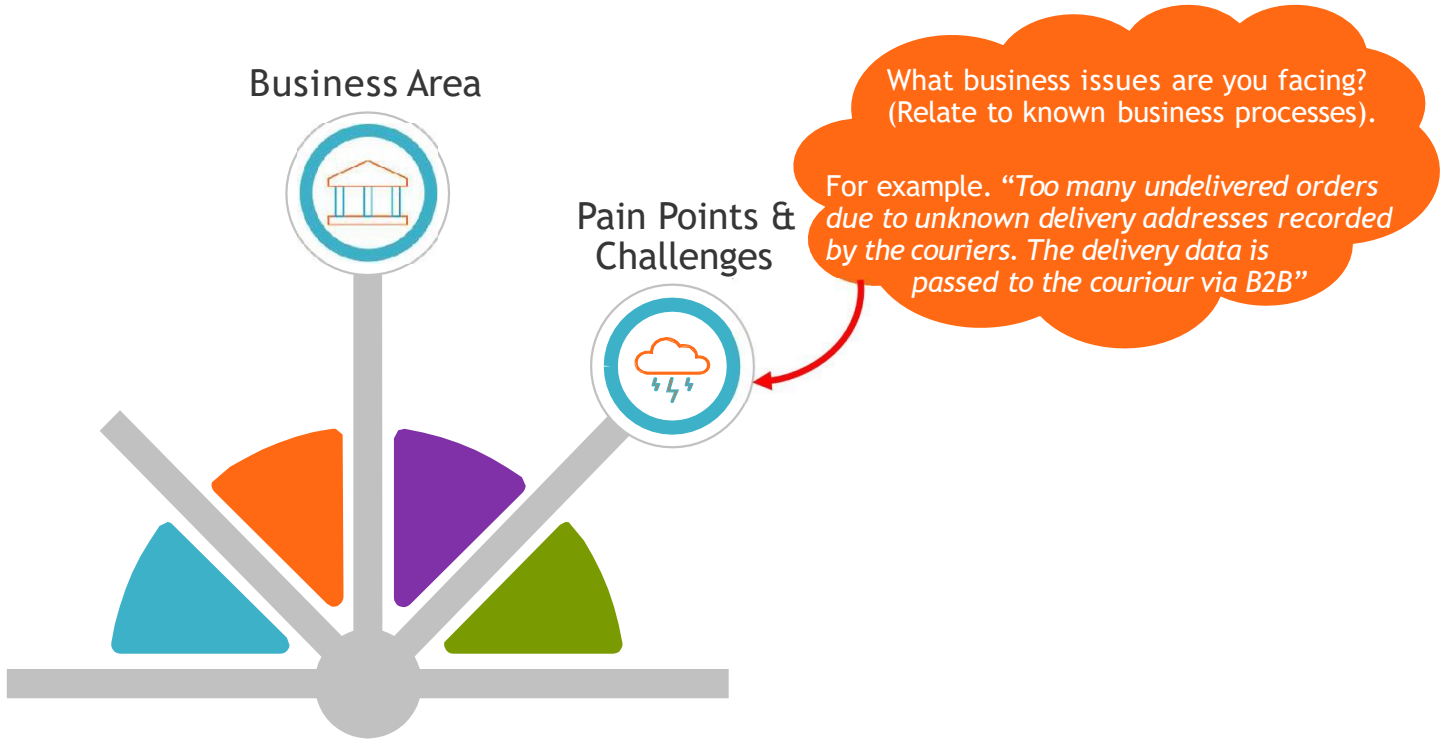
How to encourage engagement from the business?

Creating the Business Case



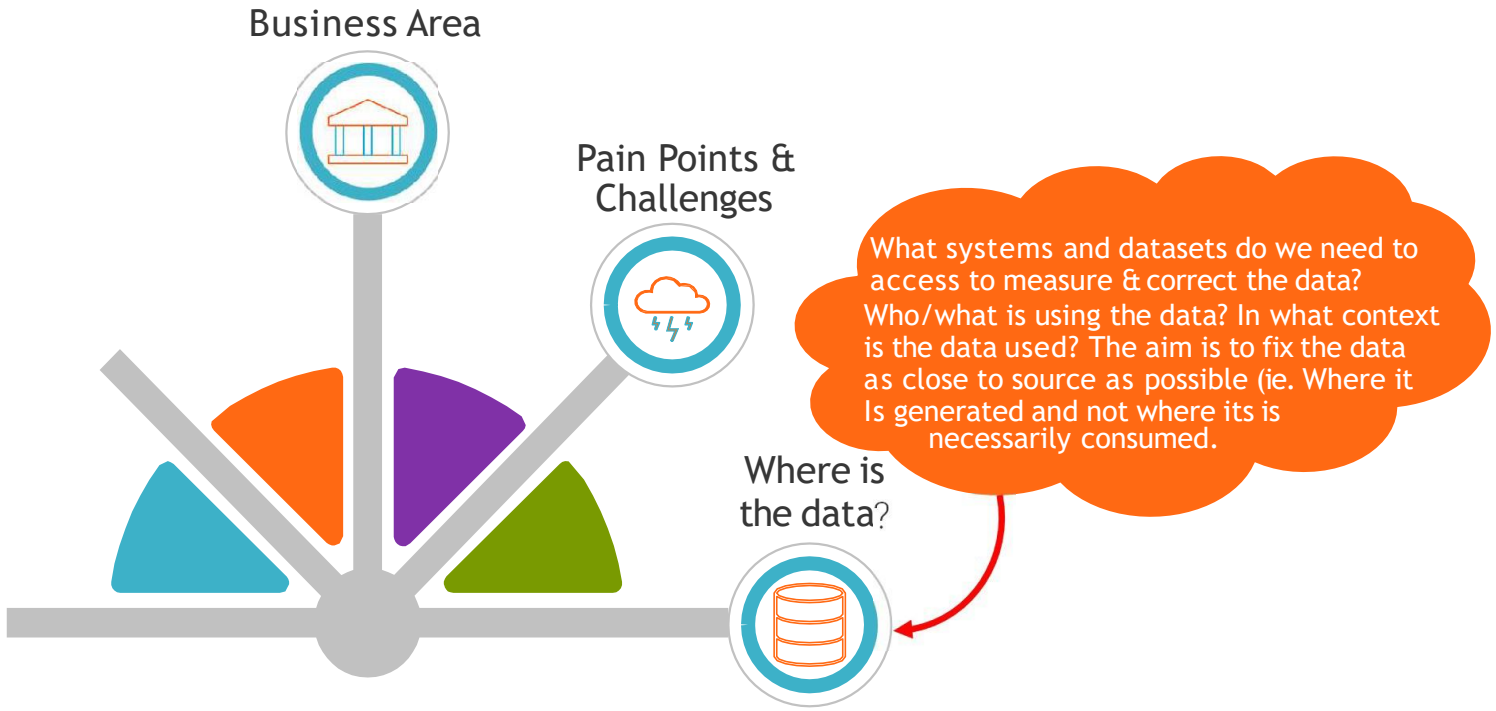
How to encourage engagement from the business?

Creating the Business Case



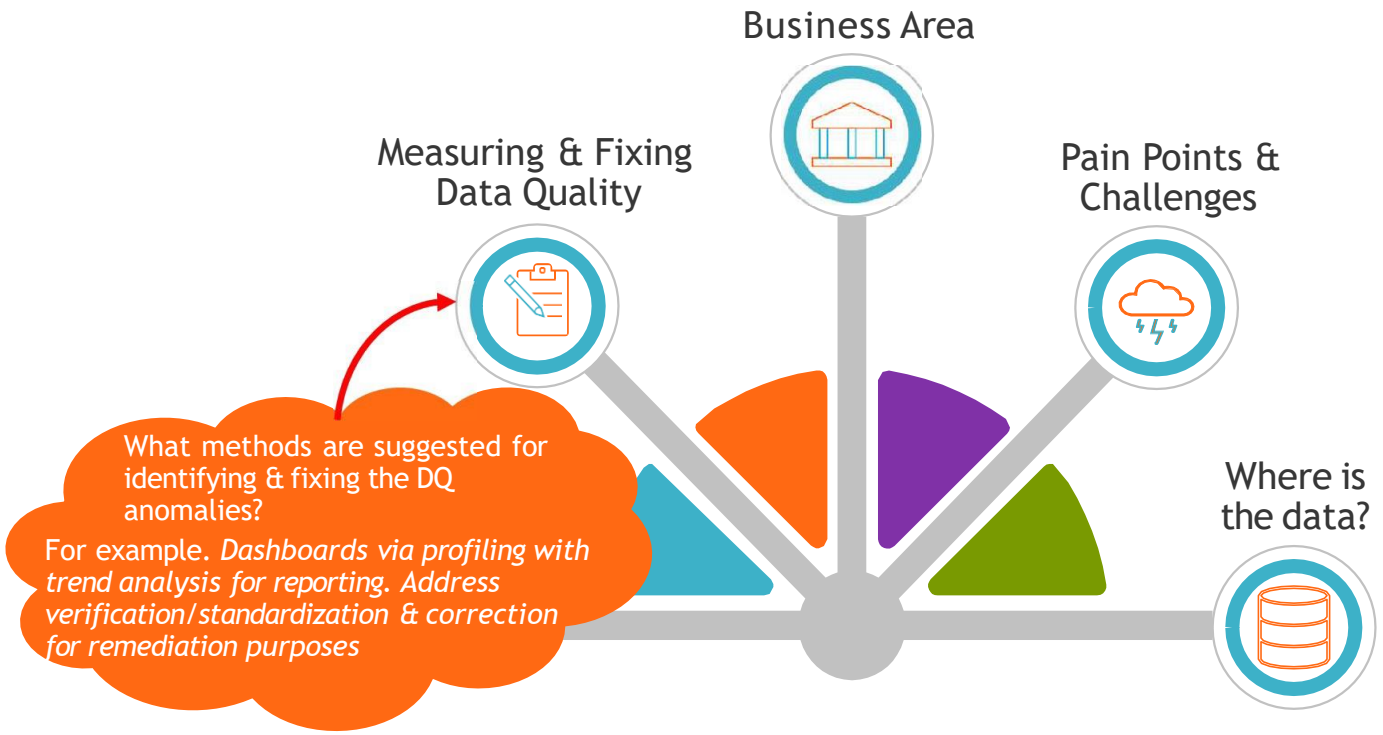
How to encourage engagement from the business?

Creating the Business Case



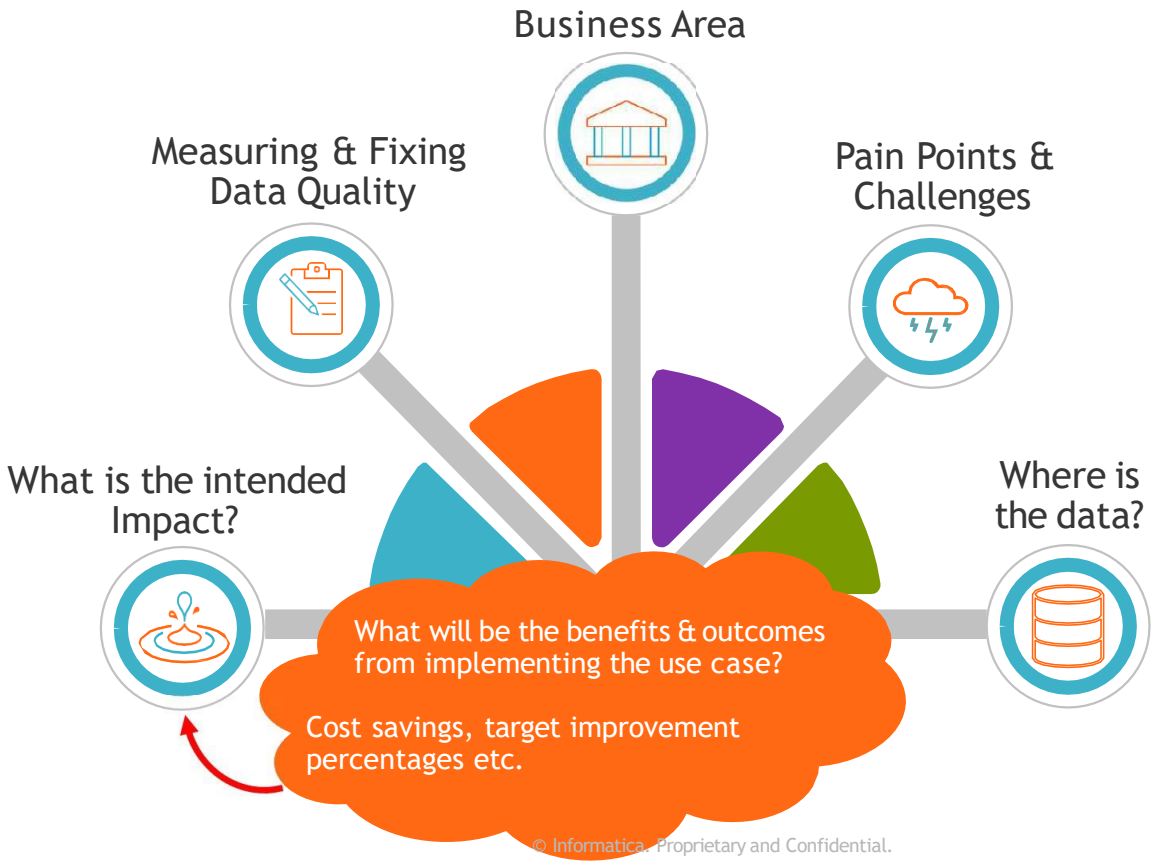
How to encourage engagement from the business?

Creating the Business Case



How to encourage engagement from the business?

Creating the Business Case



An Example Business Case for DQ

Business area(s) affected: Shipping & Distribution (Operations)

Existing pain points & challenges to overcome: The organization is estimating that 15% of all customer deliveries are being noted as "undelivered ". The NDR (Not Delivered Report) indicates that 9% of these are due to inaccurate addresses on the delivery data passed to our courier partner. It is estimated that these undelivered orders due to inaccurate address data is currently costing the company £1.72M per year in lost revenue & increased operational costs. The challenge to overcome is to increase the accuracy of our delivery address data and reduce the number of undelivered orders.

Systems, Datasets & Data elements to use in the DQ process: The sales orders and distribution system. This system feeds the B2B process that transmits order/delivery details to the appropriate couriers. This system contains several tables that contain addresses (customer address, billing address, delivery address etc). For the purpose, of this use case the focus must be on the CURRENT addresses specifically used for delivery info.

What business processes and/or people consume the data and what is the data used for?: The IT-driven B2B process that transmits order/delivery details to the appropriate couriers

How will anomalies in the data be identified & fixed?: Use a certified address verifier via a managed data quality process to identify missing or inaccurate address data. The address verifier will standardize & correct the data with an accuracy score. The corrected addresses with an accuracy above 95% threshold will be automatically loaded back into the S&D system. The data below the threshold will be reviewed and manually approved/rejected/corrected by data stewards. The manually corrected data will then be loaded back to the S&D system. Dashboards will measure and identify the "bad data" and the reporting process for this DQ initiative will also provide trend analysis such that improvements and impact of the corrections can be measured over time.

Expected business benefit: Increases the accuracy of the delivery address data will reduce the undelivered orders due to data errors from 9% down to <2%. Cost benefit will be between £1M-£2M per year

An Example Business Case for DQ

Business area(s) affected: Shipping & Distribution (Operations)

Existing pain points & challenges to overcome: The organization is estimating that 15% of all customer deliveries are being noted as "undelivered ". The NDR (Not Delivered Report) indicates that 9% of these are due to inaccurate addresses on the delivery data passed to our courier partner. It is estimated that these undelivered orders due to inaccurate address data is currently costing the company £1.72M per year in lost revenue & increased operational costs. The challenge to overcome is to increase the accuracy of our delivery address data and reduce the number of undelivered orders.

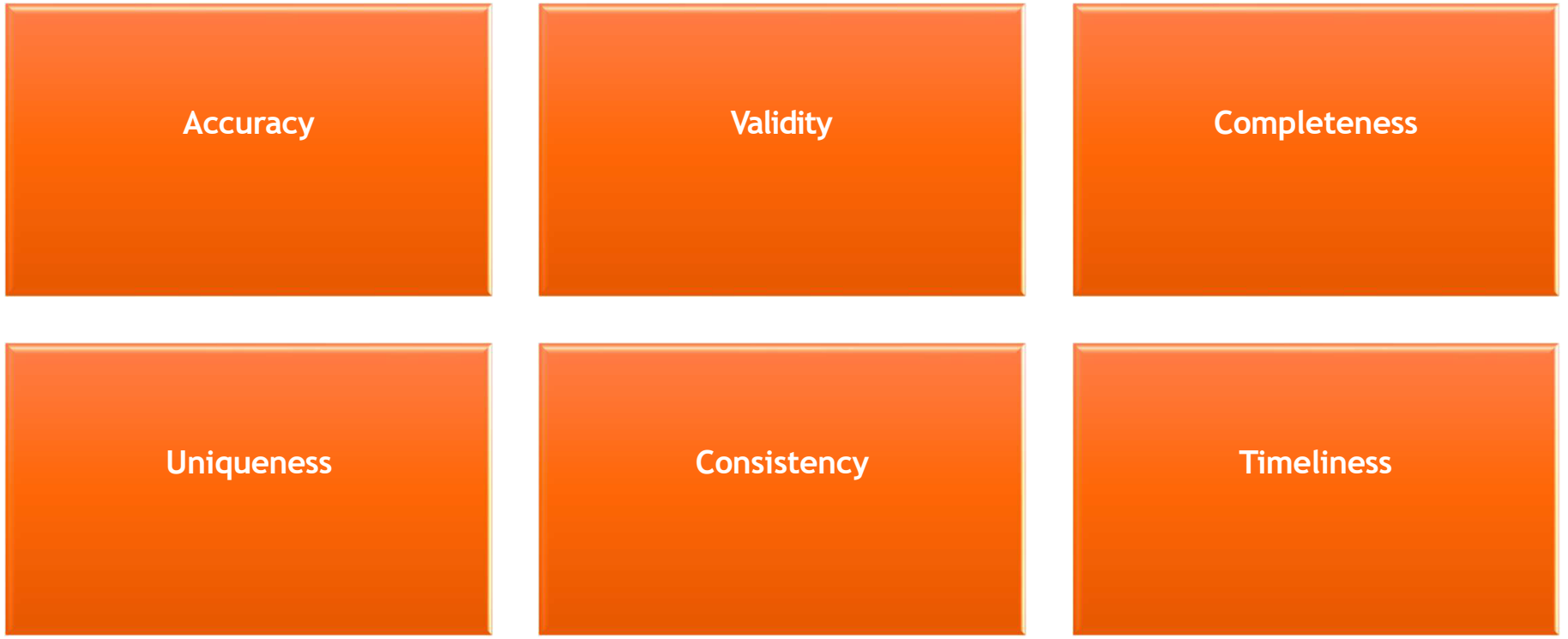
Systems, Datasets & Data elements to use in the DQ process: The sales orders and distribution system. This system feeds the B2B process that transmits order/delivery details to the appropriate couriers. This system contains several tables that contain addresses (customer address, billing address, delivery address etc). For the purpose, of this use case the focus must be on the CURRENT addresses specifically used for delivery info.

What business processes and/or people consume the data and what is the data used for?: The IT-driven B2B process that transmits order/delivery details to the appropriate couriers

How will anomalies in the data be identified & fixed?: Use a certified address verifier via a managed data quality process to identify missing or inaccurate address data. The address verifier will standardize & correct the data with an accuracy score. The corrected addresses with an accuracy above 95% threshold will be automatically loaded back into the S&D system. The data below the threshold will be reviewed and manually approved/rejected/corrected by data stewards. The manually corrected data will then be loaded back to the S&D system. Dashboards will measure and identify the "bad data" and the reporting process for this DQ initiative will also provide trend analysis such that improvements and impact of the corrections can be measured over time.

Expected business benefit: Increases the accuracy of the delivery address data will reduce the undelivered orders due to data errors from 9% down to <2%. Cost benefit will be between £1M-£2M per year

DQ dimensions



DQ dimensions

Accuracy

- Data is accurate when it reflects reality.
- For example, this can refer to correct names, addresses or represent factual and up to date data.
- Accuracy can be challenging to measure due to the need in many cases to have an benchmark to measure against.

Validity

- Validity is defined as the extent to which the data conforms to the expected format, type, and range.
- For example, an email address must have an '@' symbol; postcodes are valid if they appear in the Royal Mail postcode list; month should be between one and twelve.
- Validity and accuracy often incorrectly used interchangeably. Data can be valid but inaccurate.

Completeness

- Completeness can be measured from 2 perspectives:

1. The extent to which a field is populated
2. The extent to which a group of fields are populated to make up a full record ie name and address data.

- Completeness is not the same as accuracy as a full data set may still have incorrect values.

DQ dimensions

Uniqueness	Consistency	Timeliness
• Uniqueness is the measure of volume of duplication. • Uniqueness requirements will vary across fields, ie you would not expect a town field to be unique, but you would expect and ID field to be.	• Consistency measures the extent to which data does not conflict across different datasets.	• Timeliness measures the extent to which data is available when expected.

Methods of defining initial DQ rules

Define Rules 1st

Taking this approach will normally require some rework as there will things that come up in the data you had not expected

Profile Data 1st

This route stops wasted effort from defining rules that are not required, but profiling alone has limitations on quantifying impact of an issue

Hybrid

Having some basic rules when starting profiling work will accelerate the process and allow you to get to a useful rule set in the shortest amount of time

Example rules

Accuracy

- The post code for a company is correct against company name in Companies House data

Validity

- Field X must not 01-01-0001

Completeness

- Field X must not be null or blank

Uniqueness

- Content of field X must not appear more than once

Consistency

- Field X in dataset A must = Field X in dataset B
- must not be null or blank

Timeliness

- Last Updated Date in table A must be = to yesterday's date

Defining a Remediation Strategy

Human Task (100% Manual)

Data Stewards responsible for fixing the data are identified and assigned.
Define SLA's (Time to fix).
Notifications setup.
Workflow driven.
Exception records, & DQ dashboards generated by the Data Quality System maybe used to accelerate the process for manually fixing the data

Fully Automated

The Data quality System cleanses/standardizes data by running rules against that enrich data with cleansed/standardized values.
The DQ system connects to, and automatically updates the source system with the cleansed values. This is a fully audited process

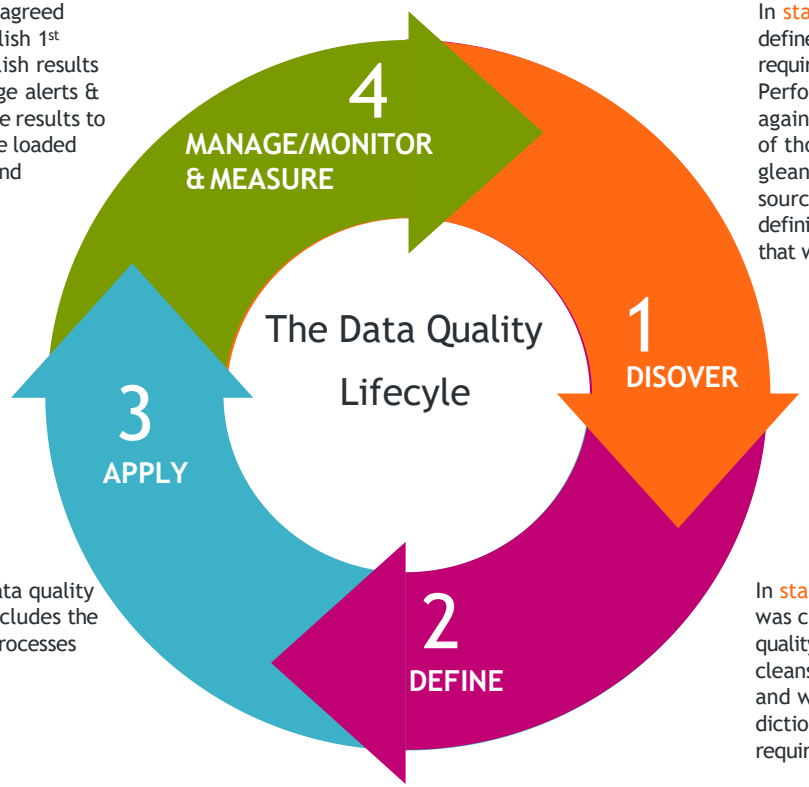
Semi Automated

The data quality system cleanses/standardizes by running rules against the data that enriches it with cleansed/standardized values.
Threshold driven configuration for automated cleansing & manual review process for data stewards to check & correct generated enrichments. This is a fully audited process.

Informatica

The Data Quality Curation Points – IDMC

In stage 4, Implement the DQ rules against the agreed sources, via profiles. Schedule the profiles, publish 1st results & monitor trend over multiple runs. Publish results as metrics to CDGC scorecards. Setup & manage alerts & notifications Profile cleansed values & compare results to actual (simulate the impact if these values were loaded into source). The manage phase also defines and implements the process for exception handling



In stage 1, work with the business or project sponsor to understand, define and document any initial DQ metrics & targets that are required. Identify and document sources for discovery profiling. Perform initial discovery profiling & analyse results. Compare results against pre-defined rules and validate feasibility of implementation of those rule. Propose & recommend new rules from insights gleaned from the discovery profile analysis. Identify potential data sources for dictionary content that might be used in any rule definitions. Finally document an agreed set of sources (tables/files) that will be used in the profiling for Define & Apply stages.

In stage 3, Develop & test the CDI mappings, data quality rules & dictionaries specified in stage 2. This includes the creation of any deduplication & consolidation processes that may be required

In stage 2, use the documented results from the analysis that was created in step 1 to create a definitive set of re-usable data quality rules (rule specifications) for both measurement and cleansing. Define which rules can be applied directly to profiles, and which rules require CDI mappings. Define and document any dictionaries which the specified rules will use. Create the required CDI mapping specs.

Informatica

Roadmap for maturing the capability

Phase 1 - Pilot

- Identify Pilot use case
- Define & test rules
- Share results/remediate
- Communicate success

Phase 2 - Initial Go Live/BAU

- Capture possible follow-on use cases
- Balanced scorecard published
- Stewardship & ownership
- Ongoing monitoring
- Create a community

Phase 3 - BAU Expansion

- Formalised polices & processes
- Published enterprise standards
- Prioritise follow-on use cases

Phase 4 scale

- Implementation of the follow-on use cases
- Formalised process on how the business can raise use cases for consideration
- Alignment with related DG and DM initiatives

Additional Info

If you would like any additional information, please contact your CSM or reach out to me at zhusband@infomatica.com

Informatica

Q&A

Where data & AI come to LIFE