



Kwaliteitsrapport

Bedrijfs/Personendata

Datum	13-04-2021
Document titel	Kwaliteitsrapport
Document versie	1.0
Auteur(s)	Data Quality Accelerator framework
Eigenaar	IntoData
Adres	Veldkant 33b 2550 Kontich

Inhoudstafel

Inhoudstafel	2
1 Data Quality: waarom, wat, hoe ?	3
2 Harmonisatie	4
3 Standaardisatie.....	5
3.1 Analyse BTW Nr	6
3.2 Analyse Bedrijfs telefoon.....	7
4 Deduplicatie	9
4.1 Match regels voor bedrijven.....	9
4.2 Match regels voor personen.....	9
5 Conclusie	10

De informatie in dit document, inclusief eventuele bijlage(n), is vertrouwelijk en uitsluitend bestemd voor de geadresseerde. De inhoud blijft intellectuele eigendom van IntoData NV. Niets uit dit document mag zonder voorafgaande schriftelijke toestemming van IntoData NV worden verveelvoudigd en/of openbaar gemaakt en/of doorgegeven aan derden. Niets uit dit document mag worden gereproduceerd ten behoeve van een onderneming, organisatie of instelling.

1 Data Quality: waarom, wat, hoe ?

Waarom?

Weten wie je klanten zijn is de kern van zakendoen. De klantgegevens worden heel vaak in verschillende applicaties bijgehouden, met alle gevolgen van dien.

De kwaliteit en standaardisatie van deze gegevens zijn daarom cruciaal.

Onjuiste of ontbrekende e-mailadressen kunnen een aanzienlijke impact hebben op marketingcampagnes. Goederen worden naar de verkeerde locaties verzonden.

Een verbeterde kwaliteit heeft een rechtstreekse impact op uw business en uw communicatie. Het zal bijgevolg de waarde van uw contacten met huidige en toekomstige klanten positief beïnvloeden.

Wat?

De marketing- of klantenservice teams kunnen meestal kiezen uit drie mogelijke manieren om contact op te nemen met de klanten: per post, e-mail of telefoon.

Deze contactgegevens zoals postadressen, e-mailadressen en telefoonnummers zijn dus sleutelkenmerken voor goeie communicatie tussen bedrijven en klanten.

De standaardisatie-inspanningen zijn vooral gericht op deze attributen.

Het is jouw taak om ervoor te zorgen dat elk van die kanalen zo volledig mogelijk én juist zijn. Het Data Quality Accelerator (DQA) framework helpt je hierbij.

Hoe?

Eerst en vooral hebben we in een **harmonisatie** stap jouw gegevens gemapt op ons generiek datamodel voor bedrijfsgegevens. Vervolgens doorloopt het DQA Framework voor het optimaliseren van de kwaliteit van de data twee grote proces stappen.

1. **Standaardisatie:**

De gegevens gaan eerst door een voorbereidende verwerking, zoals het verwijderen van ruis, standaard formaat voor email, telefoon nummers en btw-nummers, normaliseren van adressen, herkenning van de juridische vorm, enz.

2. **Deduplicatie:**

Deze opgeschoonde data wordt daarna behandeld door matching algoritmes.

Om het resultaat van dit proces inzichtelijk te maken worden voor de kritische attributen een aantal metrics voor en na de verwerking gecapteerd. Deze worden gevisualiseerd aan de hand van twee grafieken:

- I. De volledigheid, waarbij een onderscheid wordt gemaakt tussen geldig, ongeldig en leeg. De bedoeling dat het geldigheidsgraad stijgt.
- II. De duplicaatsgraad, waarbij een onderscheid wordt gemaakt tussen unieke en dubbele waarden
Het standaardisatieproces zorgt ervoor dat het aantal dubbels kan stijgen.

2 Harmonisatie

Het aantal verwerkte records: 283

Volgende attributen uit het generiek party-model zijn leeg en bevatten geen data:

- Alternatieve Naam (nick_name)
- Ondernemings Nr (enterprise_nr)
- Iban (iban)
- Bic (bic)
- Voornaam (first_name)
- Familiennaam (last_name)
- E-mailadres (email)
- Persoonlijk telefoonnummer (phone)
- GSM-nummer (mobile)
- Geboortedatum (birth_day)
- Geslacht (gender)
- Nationaliteit (nationality)
- Aanspreking (title)

3 Standaardisatie

Volgende attributen zijn verwerkt door het standaardisatie proces :

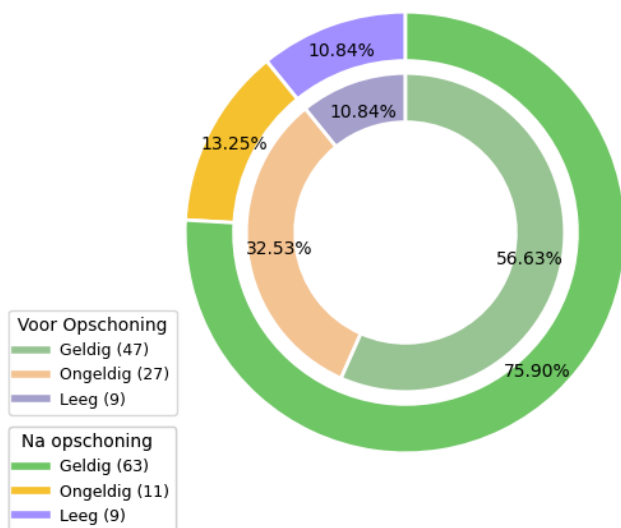
- BTW-nummer (vat)
- Bedrijfs telefoon (company_phone)
- Bedrijfs e-mailadres (company_email)
- Website (website)
- Vennootschapsvorm (legal_type)
- Taal (language_cd)
- Land (country)
- Adres (addressline1, addressline2)
- Straat (street)
- Huisnummer (house_nr)
- Bus (box)
- Postcode (zip)
- Gemeente (municipality)

Aandachtspunten bij het interpreteren van de figuren

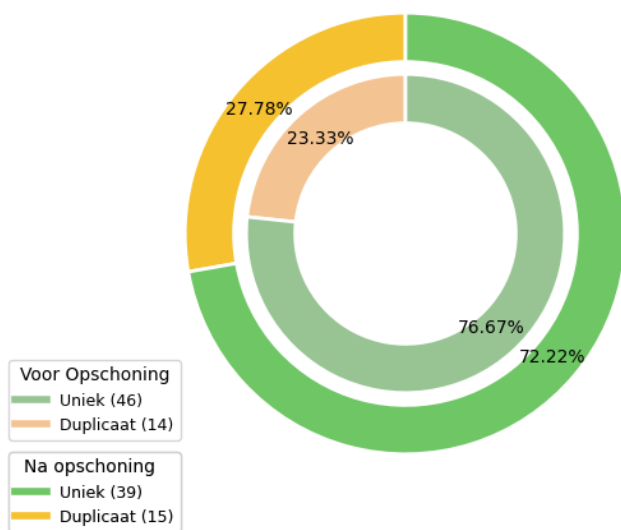
- De binnen - cirkel bevat data VOOR het cleansen
- De buiten - cirkel bevat data NA het cleansen
- De eerste figuur visualiseert de geldigheid van de data
- De tweede figuur visualiseert de potentiële dubbels in de data

3.1 Analyse BTW Nr

I. Geldigheidsgraad

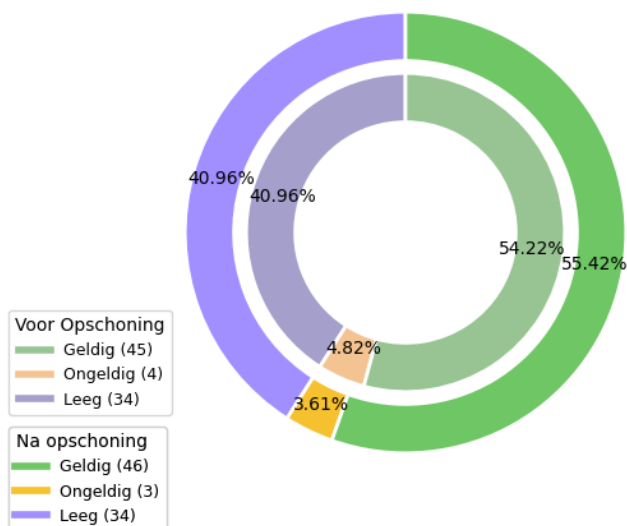


II. Duplicaatsgraad

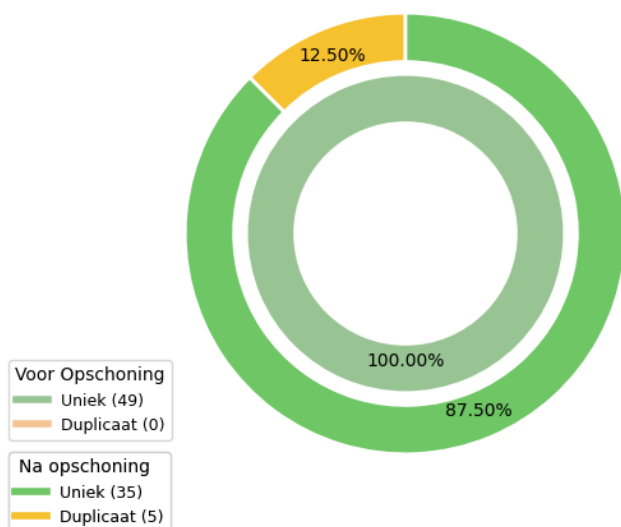


3.2 Analyse Bedrijfs telefoon

I. Geldigheidsgraad



II. Duplicaatsgraad



{ Nog meer analyses }

4 Deduplicatie

De aangeleverde dataset bevat potentieel 114 dubbels. Deze zijn terug te brengen tot 39 golden records.

4.1 Match regels voor bedrijven

De toegepaste matching regels zijn:

Regel1:

Match Kolom	Algoritme	Offset
1. Company Name	jaro	0.10
2. VAT	levenshtein	0.0

Regel2:

Match Kolom	Algoritme	Offset
1. Company Name	jaro	0.10
2. address	levenshtein	0.0

4.2 Match regels voor personen

De toegepaste matching regels zijn:

Regel3:

Match Kolom	Algoritme	Offset
1. Email	levenshtein	0.0

Regel4:

Match Kolom	Algoritme	Offset
1. Last Name	jaro	0.10
2. address	levenshtein	0.0

5 Conclusie

Het Data Quality Accelerator (DQA) framework heeft de door jouw geleverde dataset met klantinformatie verwerkt. In totaal hebben we 283 klantenrecords verwerkt.

Tijdens de "standaardisatie"-stap, heeft onze DQ-engine 1308 standaardisatie acties uitgevoerd. In de volgende tabel ziet u een overzicht van het aantal wijzigingen gegroepeerd per type actie:

Algemene cleaning stappen	642
Opsplitsen adreslijnen in individuele componenten	237
Opzoeken Belgische gemeente aan de hand van de postcode	35
Uitbreiden straatnaam	28
Opsplitsen company_name en legal_type	67
Verrijking adressen België	123
Verrijking adressen Nederland	8
Verrijking latitude & longitude	168

Na het opschoonproces zien we dat er nog steeds 248 klantgerelateerde velden als "ongeldig" worden gescoord. Daarom raden we aan om extra aandacht te besteden aan de bronapplicaties en, waar mogelijk, extra validatieregels toe te voegen voor de volgende velden:

BTW-nummer	37
Bedrijfs telefoon	11
Bedrijfs e-mailadres	28
Website	11
Straat	73
Huisnr	5
Postcode	83

In een tweede stap zorgt het "ontdubbelings"-proces voor het toepassing van de matching-regels op de door ons gestandaardiseerde data. Zonder een gedetailleerde analyse van de data en de mogelijke matching-regels & algoritmes, detecteert onze DQA-engine 114 potentiële dubbels: dat is **40.28** procent van de aangeleverde dataset.

IntoData ondersteunt u graag bij het nemen van de volgende stappen om de kwaliteit van uw data te verbeteren. Indien u de set gestandaardiseerde gegevens wenst te ontvangen, met vermelding van de duplicaten, neem dan contact met ons op.