Data warehousing

ANBC Prototype

Naam: Can Isgüzar (623974) – Lars Cornelissen (628451)  
Course: Data warehousing  
Docent: Frank Tempelman  
Versie: 1.0

Inhoud

[Bron analyse 3](#_Toc131616644)

[Data kwaliteit: 3](#_Toc131616645)

[Data kwantiteit: 3](#_Toc131616646)

[Conclusie 7](#_Toc131616647)

[KNMI Dataset: 8](#_Toc131616648)

[Functioneel ontwerp 10](#_Toc131616649)

[Te realiseren KPI’s: 10](#_Toc131616650)

[Bulb plot functionele hiërarchieën: 10](#_Toc131616651)

[Functionele invloed van de data analyse op de meetplannen: 10](#_Toc131616652)

[Schermontwerp dashboard: 11](#_Toc131616653)

[Onderbouwing Keuzes: 12](#_Toc131616654)

[Technische Ontwerp en implementatie 13](#_Toc131616655)

[Binnen dit hoofdstuk wordt er verdiept in het technische ontwerp. De processen binnen het maken van de data vault, het ETL proces en de sterschema’s worden uitgebreid beschreven. 13](#_Toc131616656)

[Architectuur 13](#_Toc131616657)

[Datavault 15](#_Toc131616658)

[Beschrijving van conceptueel model naar sterschema 16](#_Toc131616659)

[Sterschema 17](#_Toc131616660)

[ETL proces 18](#_Toc131616661)

[Ontsluiting 19](#_Toc131616662)

[Testen 21](#_Toc131616663)

[Waarde vergelijkingen: 21](#_Toc131616664)

[SQL resultaten op de brondata: 23](#_Toc131616665)

[Conclusie testen: 27](#_Toc131616666)

[APA 28](#_Toc131616667)

[Bijlagen 29](#_Toc131616668)

[Datavault ETL 29](#_Toc131616669)

[Datamart ETL 30](#_Toc131616670)

[Datamart.sql 31](#_Toc131616671)

[Datavault.sql 32](#_Toc131616672)

# Bron analyse

Er word een bronanalyse uitgevoerd op de aangeleverde ANB\_B database om de bron te bestuderen op data kwaliteit en kwantiteit.

## Data kwaliteit:

Wat wij constateren van de aangeleverde database zijn de volgende data kwaliteiten;  
De databasestructuur is een gestructureerde database. Dit betekent dat de database een relationele database is.

Hieruit kijken we naar de volgende kwaliteitsdimensies van de database.

**Accuratesse** : Er zijn twee versies van de database aan ons geleverd. De normale versie, waar alle data ongefilterd in is geregistreerd en de gefilterde dataset, waar de extreme absurde feiten eruit zijn gefilterd. De database die wij gaan gebruiken is de gefilterde database, ABC\_B. Hierdoor kunnen wij constateren dat de database zich voldoet aan de kwaliteit dimensie accuratesse.

**Volledigheid:** De database is aangeleverd vanuit het bedrijf. Zelf hebben wij de database bestudeerd en concluderen wij dat de feiten die we nodig hebben aanwezig zijn en voldoet aan de datakwaliteit dimensie Volledigheid. Bij sommige feiten zijn er NULL waardes aanwezig, maar dit zal ons proces niet belemmeren en zal de kwaliteit van de database niet omlaag halen.

**Tijd:**

* **Geldigheid:** De feiten zijn correct.
* **Veranderlijkheid:** De feiten veranderen met elke nieuwe consignment.
* **Tijdigheid:** De feiten worden gelijk geüpdatet.
* **Geschiedenis**: Er word geen geschiedenis bijgehouden binnen dit database.

**Consistentie:** De ongefilterde versie van de database was vervuild. Doordat wij een gefilterde versie ook hebben ontvangen, die niet vervuild zou moeten zijn, werken wij verder aan die versie van de database. Toch zien wij dat er geen goede consistentie zich bevindt binnen de tabellen, de benamingen verschillen per tabel. Waar bijvoorbeeld de ID van de customer opgeslagen wordt als customer binnen de tabel en bij een andere tabel wordt die ID opgeslagen als customer no. Dit geeft aan dat de data van de database vervuild is en wij daar extra rekening mee moeten houden om het overal consistent te houden.

## Data kwantiteit:

De volgende aspecten van de geleverde database worden geanalyseerd en beschreven in de tabel hieronder:

|  |  |
| --- | --- |
| Tabel | Analyse |
| CONSIGNMENT | Aantal consignments: 11501 |
| PARCEL\_OF\_CONSIGNMENT | Aantal parcels: 362 |
| CONSIGNMENT | In welke tijdsperiode speelt dit zich af: 2004-12-31 tot 2006-01-01 |
| EMPLOYEE | Hoeveel medewerkers zijn er: 9 |
| CUSTOMER | Aantal klanten: 999 |
| DISTRICT | Aantal districten: 7 |
| STATUS | Aantal vormen van status: 5 |

**CONSIGNMENT:**   
Consignment staat voor opdrachten. Binnen de CONSIGNMENT table, tref je de volgende feiten:

* **CONSIGNMENT\_NR:** Het identieke identificerende nummer van de consignment.
* **CUSTOMER:** ID nummer van de klant binnen het bedrijf.
* **PICKUPDISTRICT:** Het district waar het consignment zich bevindt.
* **ADRESSEEFIRSTNAME:** Voornaam van de geadresseerde.
* **ADRESSEELASTNAME:** Achternaam van de geadresseerde.
* **ADRESSEE\_COMPANY:** Naam van het bedrijf van de geadresseerde.
* **DELIVERYDISTRRICT:** Het district waar de consignment wordt geleverd.
* **SCHEDULED\_PICKUP\_TIME:** De besproken ophaaltijd. Deze tijd is afgesproken met de klant en de werknemer.
* **PICKUP\_TIME:** De werkelijke ophaaltijd.
* **DELIVERY-TYPE:** De type levering. Is het een directe levering of een dispatch. Dispatch betekent dat het eerst langs de sorteercentrum langs gaat.
* **CONSIGNMENTSTATUS:** Status van de levering. Is de levering geleverd of niet.
* **EXPRESS\_DELIVERY:** Is de consignment een express levering of niet. Dit wordt binair aangeven door een 1 of een nul. Een 1 betekent ja, een nul betekent nee.
* **EXP\_DURATION:** De tijd in minuten van hoelang de express levering heeft geduurd.

Dit tabel wordt onze feittabel, hiermee kunnen wij onze meeste KPI’s mee beantwoorden.

**PARCEL\_OF\_CONSIGNMENT:**  
Parcel of consignments is een link table tussen de parcel en de consignment. Dit table registreert de tijden en gewichten van de parcels van een consignment.

* **PARCEL\_NUMBER:** De nummer/id. van de parcel.
* **CONSIGNMENT**: De nummer/id. van de consignment/opdracht.
* **BIKER:** De nummer/id. van een biker/werknemer.
* **WEIGHT\_OF\_PARCEL**: Is het gewicht van de parcel/pakket.
* **SIGNATURE**: Is een lege tabel, heeft NULwaardes.
* **DELIVERY\_TIME:** Tijd van leveren.

**EMPLOYEE:**  
De Employee table behoudt data van de medewerkers binnen het bedrijf. Dit zijn zowel de centrale medewerkers en de fietskoeriers.

* **EMPLOYEE\_NR**: ID nummer van de werknemer binnen het bedrijf.
* **BSN:** De Burgerservicenummer van de werknemer.
* **EXPRESS:** Een binaire indicatie of de werknemer geschikt is voor een expres levering.
* **MAXHOUR:** De maximum uren die de werknemer mag werken.
* **CELLPHONENO:** De mobiele nummer van de werknemer.
* **HOMEPHONENO:** Het huis telefoonnummer van de werknemer.
* **BIRTHDATE:**  De geboortedatum van de werknemer.
* **GENDER:**  het geslacht van de werknemer.
* **FIRSTNAME:**  De voornaam van de werknemer.
* **LASTNAME:** De achternaam van de werknemer.
* **PASSWORD:**  De wachtwoord van de werknemer.

Met deze data kunnen we de KPI’s beantwoorden waar we de informatie van bepaalde koeriers nodig hebben zoals hun employee nr. om die te kunnen linken met de consignments en of de medewerker geschikt is voor een expres levering.

**CUSTOMER:**  
Binnen de customer tabel worden de gegevens van de klanten opgeslagen.

* **CUSTOMERNO:** ID nummer van de klant binnen het bedrijf.
* **COMPANY:** De naam van het bedrijf van de klant.
* **CONTACTPHONENO:**  De telefoonnummer van de klant.
* **CONTACT\_CELLPHONENO:** De mobiele telefoonnummer van de klant.
* **GENERALPHONENO:** De generale telefoonnummer van de klant.
* **FAXNO:** De fax nummer van de klant. Deze waarde is NULL.
* **CONTACTFIRSTNAME:** Voornaam van de klant.
* **CONTACTLASTNAME:** Achternaam van de klant.
* **DIV:** Division**.**

De customer tabel is niet voor ons belang voor het beantwoorden van onze kpi’s.

**DISTRICT:**De district table behoudt de namen van de wijken en steden waar ANBC kan leveren.

* **DISTRICTNAME:** Naam van het district.
* **TOWN:**  Naam van de stad.

Met dit tabel kunnen wij de kpi beantwoorden waarmee we kunnen meten hoe vaak er in een week overgeplaatst wordt.

**STATUS:**De status geeft aan wat de huidige stand/status van een parcel/consignment is.

* **STATUSNAME:**  Status, naamgeving, dispatch of directe levering.

De status van een pakket kan ons helpen bij het bekijken van hoeveel parcels er meegenomen worden naar het distributiecentra en misschien vervolgens daar overgebracht worden.

**DELIVERY\_PRICE:**Binnen dit table zijn de indicaties voor de prijsgeving geregistreerd. Aan de hand hiervan worden de prijzen van de parcels op basis van gewicht bepaald en of het expres levering is of niet.

* **FROMWEIGHT**: Minste gewicht.
* **TOWEIGHT**: Maximaal hoogste gewicht.
* **EXPRESS**: Is het een expres levering, wordt binair aangegeven.
* **PRICE**: De prijs van de levering.

Met dit tabel kunnen we de kpi voor de omzet bepalen en beantwoorden.

**DELIVERY\_TYPE:**  
De delivery type table geeft de informatie weer of de parcel/consignment direct geleverd is/wordt en of het eerst langs het sorteercentrum moet.

* **DISPATCH\_DIRECT:** Type leveren of het een directe levering is of gaat het eerst langs het sorteercentrum.

Dit tabel heeft voor ons geen belang bij het beantwoorden van onze kpi’s.

**WEEKLY\_DISTRICT\_DIVISION:**  
Is een table die bijhoudt welke employee welke week in welke stad heeft bezorgd.

* **WEEK:** Het nummer van de week.
* **EMPLOYEE:** Het nummer van de koerier.
* **DISTRICT**: De naam van de stad.

Dit tabel heeft voor ons geen belang voor het beantwoorden van onze kpi’s.

**Opvallers:**Data niet consistent over de tabellen;

* Consignment.Customer & Customer.CustomerNO
* Consignment.ConsignmentNr & PARCEL\_OF\_CONSIGNMENT.Consignment
* PARCEL\_OF\_CONSIGNMENT.Biker & Employee.EMPLOYEE\_NR
* Consignment.EXPRESS\_DELIVERY & Delivery\_Price.Express
* EMPLOYEE.Express niet consistente benaming.

### Conclusie

De datakwaliteit van dit database is slordig en inconsistent. Alhoewel de accuratesse wel aanwezig is op 2 features na. De inconsistentie zorgde voor verwarring tijdens de analyse. Sommige benamingen van features moeten aangepast worden naar duidelijke benamingen. Wat ons ook opviel is dat de ingevoerde data gegenereerd is. Dit herkennen we aan populaties zoals parcel 1 en customer 1, customer 2 en customer 3.

De tabellen die voor ons van belang zullen zijn, zijn de tabellen CONSIGNMENT, PARCEL\_OF\_CONSIGNMENT, EMPLOYEE, DISTRICT, STATUS en DELIVERY\_PRICE.

De statussen van bezorgingen en de status van het weer zullen een rol spelen op het dashboard. Daarbij zal er onderscheid gemaakt worden tussen een express- en een normale levering. Zelfde geldt dat ook voor goed en slecht weer.

Sommige tabellen hebben features gevuld met NULwaarden. Deze kunnen wij straks eruit filteren in de data vault, maar kiezen er zelf voor om het erin te laten. De reden hiervoor is uitbreidbaarheid, wij weten dat het bedrijf met een bon systeem werkt, en dat de consignment pas betaald is wanneer de bon is overhandigd aan de klant. Deze overhandiging kan geregistreerd worden in de vorm van signature. Het bedrijf zei tegen ons dat ze hier al mee werkten en daarom gaan wij het er niet uit filteren. Zelfde geldt ook voor het Faxnummer, er kunnen sommige klanten nog fax gebruiken. De optie laten behouden wij voor het bedrijf.

# KNMI Dataset:

Om de live data van de KNMI op te halen, is er een API call gedaan naar Opendata API van het KNMI. Deze call ziet er als volgt uit:

<https://www.daggegevens.knmi.nl/klimatologie/uurgegevens?fmt=json&start=2004123100&end=2006011600&stns=275:375>

De call bestaat uit 4 onderdelen:

* Fmt: Json. JSON wordt gebruikt om beter de gegevens te kunnen lezen met python
* Start: De start datum van de opgevraagde data. De dataset loopt van 2004-12-31 23:45:00 dus we hebben alle data vanaf 2004-12-31 00:00 gebruikt
* End: De einddatum van de opgevraagde data. De dataset loop top 2006-01-15 19:37:00 dus we hebben alle data vanaf 2006-01-16 00:00 gebruikt.
* Stns: De weerstations die opgevraagd worden. In Arnhem en Nijmegen zijn er geen KNMI weerstation dus hebben wij de dichtstbijzijnde gebruikt (Deelen (275) en Volkel (375))

De volgende station ID’s zijn opgehaald:

Deelen 275: Dit is de station id in de buurt van Arnhem. Hiermee kunnen wij de weer data van Arnhem ophalen.

Volkel 375: Dit is de station id in de buurt van Nijmegen. Hiermee kunnen wij de weer data van Nijmegen ophalen.

Binnen deze data kan het volgende afgelezen worden:

**STN LON(east) LAT(north) ALT(m) NAME**

275 5.873 52.056 48.20 Deelen = Arnhem

375 5.707 51.659 22.00 Volkel = Nijmegen

**DD**  : Windrichting (in graden) gemiddeld over de laatste 10 minuten van het afgelopen uur (360=noord; 90=oost; 180=zuid; 270=west; 0=windstil 990=veranderlijk. Zie http://www.knmi.nl/kennis-en-datacentrum/achtergrond/klimatologische-brochures-en-boeken / Mean wind direction (in degrees) during the 10-minute period preceding the time of observation (360=north; 90=east; 180=south; 270=west; 0=calm 990=variable)

**FH** : Uurgemiddelde windsnelheid (in 0.1 m/s). Zie http://www.knmi.nl/kennis-en-datacentrum/achtergrond/klimatologische-brochures-en-boeken / Hourly mean wind speed (in 0.1 m/s)

**FF** : Windsnelheid (in 0.1 m/s) gemiddeld over de laatste 10 minuten van het afgelopen uur / Mean wind speed (in 0.1 m/s) during the 10-minute period preceding the time of observation

**FX** : Hoogste windstoot (in 0.1 m/s) over het afgelopen uurvak / Maximum wind gust (in 0.1 m/s) during the hourly division

**T** : Temperatuur (in 0.1 graden Celsius) op 1.50 m hoogte tijdens de waarneming / Temperature (in 0.1 degrees Celsius) at 1.50 m at the time of observation

**T10N** : Minimumtemperatuur (in 0.1 graden Celsius) op 10 cm hoogte in de afgelopen 6 uur / Minimum temperature (in 0.1 degrees Celsius) at 0.1 m in the preceding 6-hour period

**TD** : Dauwpuntstemperatuur (in 0.1 graden Celsius) op 1.50 m hoogte tijdens de waarneming / Dew point temperature (in 0.1 degrees Celsius) at 1.50 m at the time of observation

**SQ** : Duur van de zonneschijn (in 0.1 uren) per uurvak; berekend uit globale straling (-1 for <0.05 uur) / Sunshine duration (in 0.1 hour) during the hourly division; calculated from global radiation (-1 for <0.05 hour)

**Q** : Globale straling (in J/cm2) per uurvak / Global radiation (in J/cm2) during the hourly division

**DR** : Duur van de neerslag (in 0.1 uur) per uurvak / Precipitation duration (in 0.1 hour) during the hourly division

**RH** : Uursom van de neerslag (in 0.1 mm) (-1 voor <0.05 mm) / Hourly precipitation amount (in 0.1 mm) (-1 for <0.05 mm)

**P** : Luchtdruk (in 0.1 hPa) herleid naar zeeniveau; tijdens de waarneming / Air pressure (in 0.1 hPa) reduced to mean sea level; at the time of observation

**VV** : Horizontaal zicht tijdens de waarneming (0=minder dan 100m; 1=100-200m; 2=200-300m;...; 49=4900-5000m; 50=5-6km; 56=6-7km; 57=7-8km; ...; 79=29-30km; 80=30-35km; 81=35-40km;...; 89=meer dan 70km) / Horizontal visibility at the time of observation (0=less than 100m; 1=100-200m; 2=200-300m;...; 49=4900-5000m; 50=5-6km; 56=6-7km; 57=7-8km; ...; 79=29-30km; 80=30-35km; 81=35-40km;...; 89=more than 70km)

**N** : Bewolking (bedekkingsgraad van de bovenlucht in achtsten); tijdens de waarneming (9=bovenlucht onzichtbaar) / Cloud cover (in octants); at the time of observation (9=sky invisible)

**U** : Relatieve vochtigheid (in procenten) op 1.50 m hoogte tijdens de waarneming / Relative atmospheric humidity (in percents) at 1.50 m at the time of observation

**WW** : Weercode (00-99); visueel(WW) of automatisch(WaWa) waargenomen; voor het actuele weer of het weer in het afgelopen uur. Zie http://bibliotheek.knmi.nl/scholierenpdf/weercodes\_Nederland / Present weather code (00-99); description for the hourly division.

**IX** : Weercode indicator voor de wijze van waarnemen op een bemand of automatisch station (1=bemand gebruikmakend van code uit visuele waarnemingen; 2;3=bemand en weggelaten (geen belangrijk weersverschijnsel; geen gegevens); 4=automatisch en opgenomen (gebruikmakend van code uit visuele waarnemingen); 5;6=automatisch en weggelaten (geen belangrijk weersverschijnsel; geen gegevens); 7=automatisch gebruikmakend van code uit automatische waarnemingen) / Indicator present weather code (1=manned and recorded (using code from visual observations); 2;3=manned and omitted (no significant weather phenomenon to report; not available); 4=automatically recorded (using code from visual observations); 5;6=automatically omitted (no significant weather phenomenon to report; not available); 7=automatically set (using code from automated observations)

**M** : Mist 0=niet voorgekomen; 1=wel voorgekomen in het voorgaande uur en/of tijdens de waarneming / Fog 0=no occurrence; 1=occurred during the preceding hour and/or at the time of observation

**R** : Regen 0=niet voorgekomen; 1=wel voorgekomen in het voorgaande uur en/of tijdens de waarneming / Rainfall 0=no occurrence; 1=occurred during the preceding hour and/or at the time of observation

**S** : Sneeuw 0=niet voorgekomen; 1=wel voorgekomen in het voorgaande uur en/of tijdens de waarneming / Snow 0=no occurrence; 1=occurred during the preceding hour and/or at the time of observation

**O** : Onweer 0=niet voorgekomen; 1=wel voorgekomen in het voorgaande uur en/of tijdens de waarneming / Thunder 0=no occurrence; 1=occurred during the preceding hour and/or at the time of observation

**Y** : IJsvorming 0=niet voorgekomen; 1=wel voorgekomen in het voorgaande uur en/of tijdens de waarneming / Ice formation 0=no occurrence; 1=occurred during the preceding hour and/or at the time of observation

# Functioneel ontwerp

Binnen het Functioneel ontwerp wordt de blauwdruk van het uiteindelijke dashboard gerealiseerd. De KPI’s die gerealiseerd gaan worden, worden nogmaals kort beschreven, daarbij wordt er een hiërarchische bulbplot weergegevens, daarbij word er een functionele invloed van de data analyse op de meetplannen beschreven en als volgt een wireframe van de te realiseren dashboard.

## Te realiseren KPI’s:

De volgende KPI’s uit het meet actieplan zullen gerealiseerd worden tot een dashboard:

* Hoelang duurt het bezorgen van een levering door een nieuwe koerier T.o.v. een ervaren koerier.

Zal gerealiseerd worden tot een: Staafgrafiek.

* Hoeveel procent van de express leveringen worden niet op tijd bezorgd.

Zal gerealiseerd worden tot een: Donut grafiek.

* Hoeveel consignments worden er op slechte weersomstandigheden dagen bezorgd T.o.v. normale dagen.

Zal gerealiseerd worden tot een: Histogram.

* Hoeveel consignments worden er gemiddeld in een week overgeplaatst.

Zal gerealiseerd worden tot een: Histogram.

* Zijn er genoeg koeriers voor de hoeveelheid parcels op een dag.

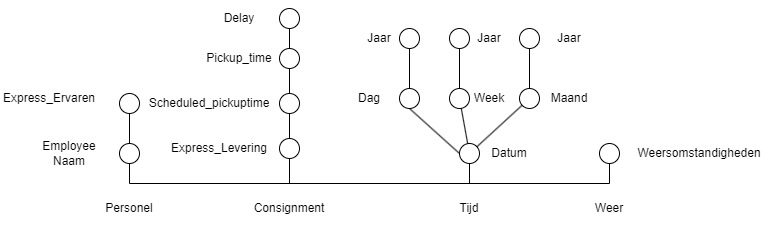
Zal gerealiseerd worden tot een: Staafgrafiek

* Wat is het verschil in omzet in omstandigheden van goed en slecht weer per activiteit.

Zal gerealiseerd worden tot een: Wafelgrafiek en divergerende grafiek.

De focus van deze KPI's ligt vooral binnen het proces van afhalen en bezorgen. Wij willen dit proces binnen het bedrijf ANBC meten en in kaart brengen in de vorm van een dashboard.  
Voor uitgebreidere beschrijvingen van de KPI’s zie het meetactieplan document in de bijlagen.(*Zie document Data ware housing Casus - Can Isguzar(623974) – Lars Cornelissen(628451).docx).*

## Bulb plot functionele hiërarchieën:



Figuur 1: bulb plot

## Functionele invloed van de data analyse op de meetplannen:

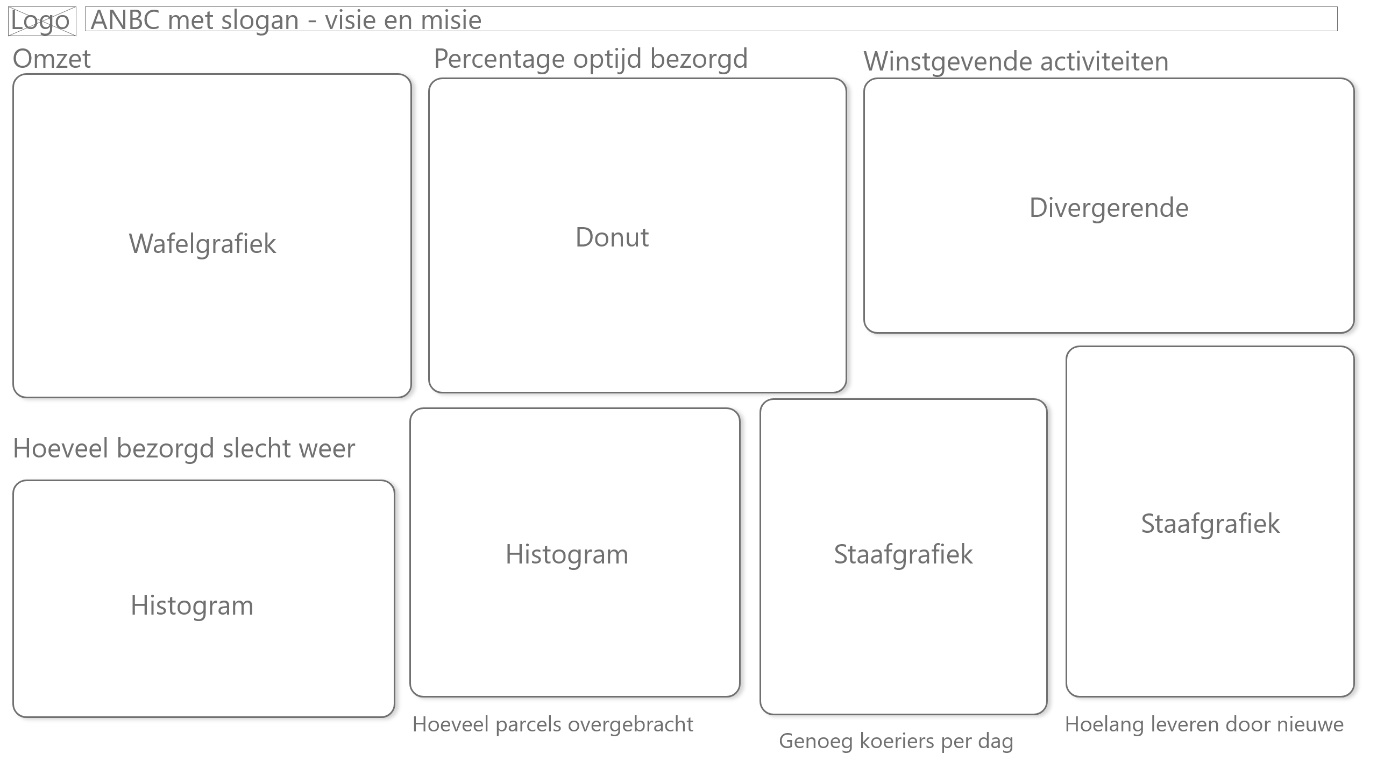
Door de inconsistentie van de data, dat is gebleken uit de data analyse, moeten naamgevingen van features binnen de tabellen aangepast worden zodat het consistent wordt in de gehele database.

Verder weten we dat de database gefilterd is van rotzooi data maar wij zijn toch achter gekomen vanuit de analyse dat er sommige features aanwezig zijn met NULwaarden. Deze zullen geen effect hebben op het uiteindelijk te realiseren dashboard.

De KPI “*hoeveel consignments worden er gemiddeld in een week overgeplaatst*” moeten we aanscherpen. Want er is binnen de data geen regelrechte aanwijzing dat een pakket overgeplaatst moet worden. Dit kunnen we beter aanscherpen naar, hoeveel consignments worden er direct geleverd t.o.v. die gebracht worden naar het distributiecentra.

## Schermontwerp dashboard:

Hier ziet u de indeling van het dashboard concept vormgegeven in een wire frame.



Figuur 2: Schermontwerp dashboard



Figuur 3: LOGO ANBC

Dit is de huisstijl die wij hebben gecreëerd voor het ANBC dashboard.

Alle KPI’s worden in 1 opzicht opgenomen, waarvan de grafieken klikbaar zullen worden die de data filteren. Dit doen we om zo veel mogelijk ruimte in beslag te nemen om alles duidelijk op 1 scherm te kunnen krijgen.

*Disclaimer: Dit is en blijft een concept, het uiteindelijke product kan afwijken van het concept.*

## Onderbouwing Keuzes:

De volgende keuzes zijn er genomen tijdens het maken van de bulb plot en het schermontwerp.

Wij hebben voor de 4 dimensies gekozen, personeel, consignment, tijd en weer, omdat deze dimensies nodig zijn om onze Kpi’s te kunnen beantwoorden. Uit deze 4 dimensies is de consignment dimensie het feit tabel. Van elke dimensie hebben we maar 1 attribuut nodig die we weer kunnen doorlussen met de andere dimensie tabellen, waardoor we onze benodigde informatie eruit kunnen halen en zo de KPI's kunnen beantwoorden. De tijdsdimensie tabel is opgesplitst van datum in dag, week en maand. Die weer alle 3 aan het jaar gebonden zijn. Hier is ervoor gekozen om de dag te kunnen linken aan het KNMI weer data om zo de weersomstandigheden voor die dag op te kunnen halen om zo de KPI's mee te kunnen beantwoorden.

Bij het schermontwerp is er de keuze gemaakt één doorlopend verhaal te vertellen waarin de belangrijkste onderwerpen als eerste aan bod komen. Doordat het hoofddoel van het bedrijf is de omzet te vergroten , komt de omzet in een jaaroverzicht als eerste aanbod. Vervolgens worden de percentages van geleverde producten weergegeven met daar achteraan welke activiteiten daarin het meeste verdienen. De volgende rij verdiept zich in die processen, waar het goed en fout gaat in vormen van staaf en histogrammen.

Wij hebben er ook een huisstijl bij gemaakt, om het dashboard echt tot het bedrijf te laten horen. Hiervoor hebben wij een logo gemaakt voor ANBC. Dit hadden ze namelijk niet. Met daarbij passende groene energie, representatieve kleuren en een leuke passende slogan. Vervolgens hebben we de kleuren voor het dashboard ook in kaart gebracht. We wilden de kleuren van het logo terugbrengen in het dashboard.

De grafieken zijn klikbaar, waar de true false waardes de data over het gehele dashboard filteren. Als de gebruiker bijvoorbeeld in de donut grafiek klikt op de express bezorgers, dan wordt het hele dashboard gefilterd op alleen express leveringen. Zelfde geldt ook voor slechte weersomstandigheden.

# Technische Ontwerp en implementatie

## Binnen dit hoofdstuk wordt er verdiept in het technische ontwerp. De processen binnen het maken van de data vault, het ETL proces en de sterschema’s worden uitgebreid beschreven.

## Architectuur

Om de architectuur van de Datawarehouse te realiseren moeten wij denken aan de lagen, de bronsystemen en welke ETL proces/processen er nodig zijn. Daarbij ook de beschrijving van de tooling die we nodig zullen hebben om de gegevens te ontsluiten.

**Lagen:**  
De lagen representeren de stappen die gevolgd gaan worden van data source naar data marts. Hierin is de data source de aangeleverde originele database vanuit het bedrijf ANBC. Binnen de staging area wordt het ETL proces uitgevoerd, hierin worden de gegevens uit verschillende bronnen gehaald, getransformeerd in een geschikt formaat en vervolgens geladen in de doelbestemming, in dit geval het datawarehouse in de storage layer. Vervolgens wordt er binnen die storage layer van de data warehouse, data marts gemaakt. Sterschema’s. Met deze sterschema’s kunnen we de benodigde informatie filteren en tonen in de presentatielaag. Wat weer leesbaar is voor de onwetende. (*Zie bron:* (Making Sense of Data Warehouse Architecture, 2018)).

* Source layer
* Staging Layer
* Storage Layer
* Presentation Layer

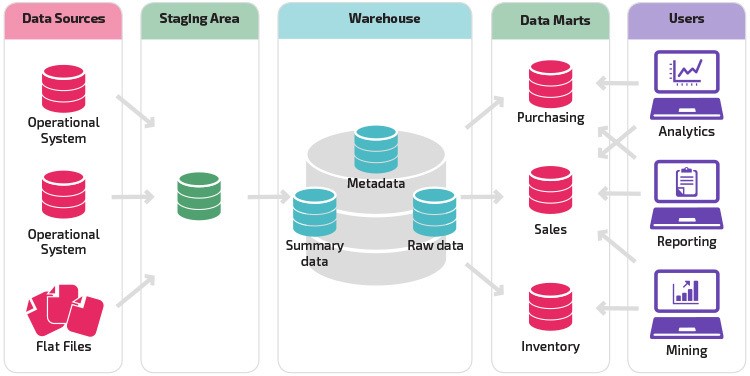


Figure 1: Layers

**Bron systemen:**  
Voor het gehele proces worden de volgende bron systemen gebruikt;

* ANBC\_B Database
* KNMI Open Database

**ANBC\_B data:**  
Het ANBC\_B data is afkomstig vanuit school. Dit is een gefilterde versie van het originele database. Hierin zijn alle extreme en onzinnige data eruit gefilterd.

**KNMI data:**  
Vervolgens hebben wij de KNMI data voor het weer nodig om de KPI’s rondom slecht weer mee te kunnen beantwoorden. Hiervoor hebben wij de volgende features gekozen om die over te nemen in ons datawarehouse.

* FH: Uurgemiddelde windsnelheid
* T: Temperatuur
* RH: Uursom neerslag
* P: Luchtdruk
* VV: Horizontaal zicht
* N: Bewolking
* U: Relatieve vochtigheid
* M: Mist Ja/Nee
* R: Regen Ja/Nee
* S: Sneeuw Ja/Nee
* O: Onweer Ja/Nee
* Y: ijsvorming Ja/Nee

**ETL processen:**

Dit is een proces waarin de gegevens uit verschillende bronnen worden gehaald, getransformeerd in een geschikt formaat en vervolgens geladen in de doelbestemming. Voor het uitvoeren van dit systeem gaan wij gebruik maken van python.

Er zijn twee ETL processen nodig:

1. Van de source-database (ABC\_B) naar de datavault (ANBC).
2. Van de datavault (ANBC) naar de datamart/sterschema (ANBC\_STER).

**Tooling:**De volgende tooling worden gebruikt voor het uitsluiten van de data.

**Power-BI**: Is verplicht vanuit school om te gebruiken.

**Tableau:** Voor school moeten we de uitsluiting van de data nogmaals realiseren binnen een andere omgeving. Wij kiezen ervoor om dit binnen Tableau te doen. Omdat dit programma in de lijst zit van meest gebruikte tools binnen dashboard software volgens de volgende bron. (*zie bron:* (10 Best Dashboard Software and Tools for 2023, 2023)*.*

## Afbeelding met diagram Automatisch gegenereerde beschrijvingDatavault

Figure 2: Datavault PDM

**Keuzes toelichting:**Tijdens het opstellen van de datavault zijn de volgende keuzes gemaakt en worden hier onderbouwd.

Er is gekozen om de weather tabel te koppelen aan de consignment tabel. Dit zou in business-taal betekenen dat we bij een bezorging naar het weer kunnen kijken. Er is een losse parcel-tabel gemaakt, in plaats van een koppeltabel (parcel\_of\_consignment). Dit is gedaan omdat parcel te veel informatie bevat om als los object te zien binnen het domein en database.

*Zie bijlagen datavault.sql voor de implementatie en de sql code.*

## Beschrijving van conceptueel model naar sterschema

Vanuit het Datavault PDM is de datamart/sterschema gemaakt en binnen het ETL proces aangevuld.

Consignment staat centraal in ons model. De KPI’s zijn allemaal gevraagd op basis van consignments. De KPI’s stellen vragen over het verschil in aantal/prijs over verloop van tijd en op basis van het weer. Daarnaast worden er vragen gesteld over de express-type van een pakketjes, hier wordt de express-ervaring van de personeelsmedewerker gebruikt.

Kunnen de KPI’s hiermee beantwoord worden?

* Hoelang duurt het bezorgen van een levering door een nieuwe koerier T.o.v. een ervaren koerier.

Zal gerealiseerd worden tot een: Staafgrafiek.

* Hoeveel procent van de express leveringen worden niet op tijd bezorgd.

Zal gerealiseerd worden tot een: Donut grafiek.

* Hoeveel consignments worden er op slechte weersomstandigheden dagen bezorgd T.o.v. normale dagen.

Zal gerealiseerd worden tot een: Histogram.

* Hoeveel consignments worden er gemiddeld in een week overgeplaatst.

Zal gerealiseerd worden tot een: Histogram.

* Zijn er genoeg koeriers voor de hoeveelheid parcels op een dag.

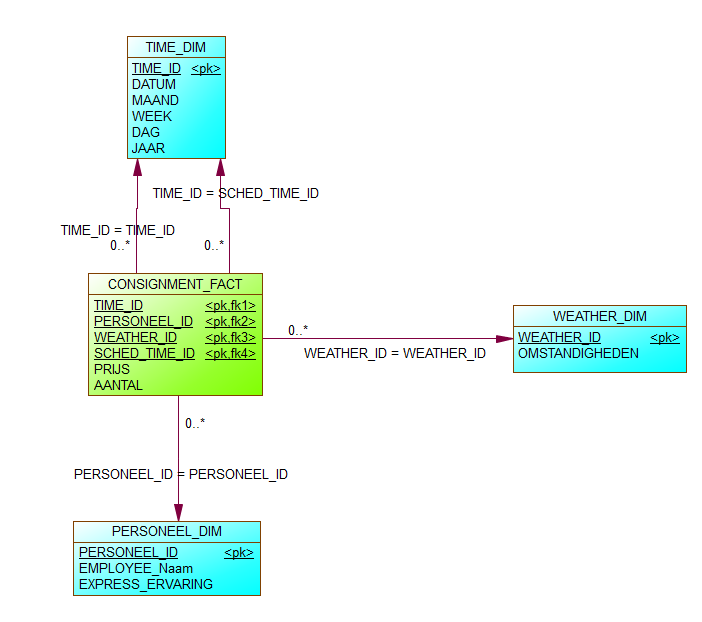
Zal gerealiseerd worden tot een: Staafgrafiek

* Wat is het verschil in omzet in omstandigheden van goed en slecht weer per activiteit.

Zal gerealiseerd worden tot een: Wafelgrafiek en divergerende grafiek.

We zien dat de hoeveelheid overgeplaatste parcels per dag hiermee niet weergegeven kunnen worden op basis van dit sterschema. Dit kregen wij ook niet voor elkaar met de pythoncode. Dit zou verbeterd kunnen worden in een volgende versie, waarbij wellicht de origin/destination van de consignment toegevoegd kan worden.

## Sterschema



*Zie bijlagen sterschema.sql voor de sql code).*

## ETL proces

Er zijn twee processen uitgevoerd:

1. Van de source-database naar de datavault
2. Van de datavault naar de datamart (sterschema)

Deze processen zijn, zoals eerder verteld, uitgevoerd in Python (zie bijlage voor volledige script). Dit is gekozen omdat dit veel gevraagd wordt in bedrijven, en dit beter past binnen ons idee van kennis over product.

Het doel van het ETL proces is om de gegevens uit een bron-systeem te halen, deze te transformeren in het formaat van het doelsysteem, en het laden van de gegevens in het doelsysteem.

Het proces begint met het opzetten van een verbinding met het bron-systeen met behulp van PyODBC: een ODBC wrapper voor python. Vervolgens worden de gegevens uit de verschillende tabellen in de database gehaald met behulp van eenvoudige SELECT SQL-Queries. De gegevens worden per tabel opgeslagen in een python-lijst voor verwerking.

De transform-stap bestaat in ons systeem uit verschillende sub-stappen, waarbij de gegevens worden getransformeerd naar het gewenste formaat. De identifier van elke record is uniek gemaakt.

De load-stap begint weer bij het verbinden met een database met behulp van PyODBC.

Er zijn een aantal besluiten genomen tijdens het schrijven van de python-code, zoals:

1. De PK's worden gegenereerd in de python-code door te checken op dubbele entries van de originele PK's.
2. De datum-tijd-gegevens worden genormaliseerd met datetime.datetime.
3. Weersomstandigheden is TRUE of FALSE. TRUE betekent dat er slecht weer was op dat tijdstip (mist, sneeuw, regen, ijs, etc.)
4. Src\_town en dest\_town zijn 0 of 1: 0 betekent Arnhem, 1 betekent Nijmegen.

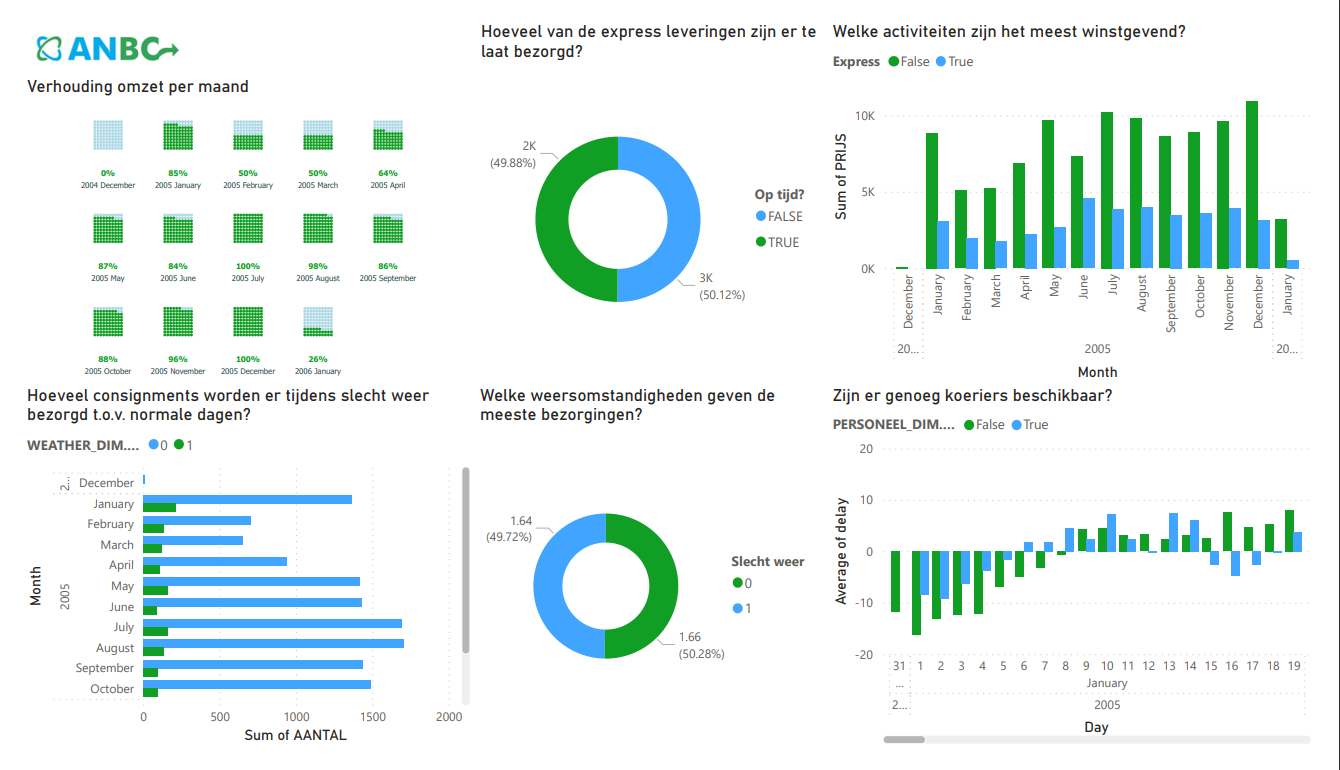


Figure 3: Power Bi dashboard.

## Ontsluiting

Figure 3: Power Bi dashboard.

**Power BI:**

Hier ziet u een screenshot van het gerealiseerde dashboard. Wij hebben zo veel mogelijk geprobeerd om aan de schermontwerpen binnen het Functioneel ontwerp te houden, maar niet alles gaat altijd volgens plan. Wij hebben een paar wijzigingen moeten doen om het product te realiseren. De grafiek rechts boven ‘*Welke activiteiten zijn het meest winstgevend’* hadden wij een divergerende grafiek voor in gedachte maar hebben gekozen voor een normale staafgrafiek omdat de optie voor een divergerende grafiek niet aanwezig was.

Zo is van de dashboard van links naar rechts af te lezen, 1e grafiek het jaar omzet, hoeveel omzet heeft elke maand opgeleverd. 2e grafiek de te late en optijd bezorgde express leveringen in donut formaat. 3e grafiek de winstgevende activiteiten over het jaar. 4e grafiek het aantal consignments bij goed en slecht weer in een staafgrafiek formaat. 5e grafiek een donut die de aantal consignments per slecht en goed weer weergeeft. 6e grafiek die de aantal vertragingen van de consignments aangeeft per ervaren en onervaren koerier en of er genoeg koeriers in dienst zijn.

Wij hebben een aanpassing moeten verrichten in de benamingen van de grafieken en een andere naamgeving voor het beantwoorden van de kpi van “Hoeveel consignments worden er op slechte weersomstandigheden dagen bezorgd t.o.v. normale dagen.” Naar welke weersomstandigheden geven de meeste bezorgingen? Hiervan hebben wij ervoor gekozen een donut grafiek te maken omdat de aflezing hierdoor in 1 opzicht beter te zien was. Hierin is nu in 1 opzicht te zien of er tijdens goed of slecht weer de meeste bezorgingen zijn.

Bij de grafiek van rechts onder hebben wij ook keuzes gemaakt om 2 KPI’s in 1 te beantwoorden in een andere naamgeving en vorm van grafiek. Om deze 2 KPI’s te beantwoorden zijn wij gaan kijken naar de delay factor. Om een inzicht te krijgen op welke periodes er het meeste delay is, kun je eruit constateren of er genoeg koeriers aanwezig zijn binnen het bedrijf en of de leveringen verschil in delay hebben tussen ervaren en onervaren koeriers.

Afbeelding met grafiek

Automatisch gegenereerde beschrijving**Tableau:**

Figure 4: Tableau dashboard.

Hier zien we het dashboard gerealiseerd binnen tableau. De twee toolingen komen erg met elkaar overeen in resultaat. Wat erg verschilt zijn de grafische vormgevingen, bijvoorbeeld wat ons opviel is dat de wafel grafiek binnen tableau erg onhandig te lezen is, binnen BI is het in 1 opzicht te zien welke maanden de meeste omzet hebben, binnen tableau zijn hij grote en kleine vierkantjes die moeilijk af te lezen zijn.

# Testen

## Waarde vergelijkingen:

Table 1 ANC\_B database Employee tabel

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| EMPLOYEENR | BSN | EXPRESS | MAXHOUR | CELLPHONENO | HOMEPHONENO | BIRTHDATE | FIRSTNAME | LASTNAME |
| 1 | 1234567890 | ONWAAR | 40 | 06-12345678 | 026-5671234 | 06-09-1990 | Jan | Janssen |
| 2 | 4567890123 | ONWAAR | 28 | 06-45678123 | 026-1672345 | 01-10-1992 | Mark | Bruinvis |
| 3 | 1237890456 | ONWAAR | 36 | 06-67812345 | 026-3451267 | 06-11-1991 | Sarah | Ollily |
| 4 | 8901234567 | WAAR | 28 | 06-44545678 | 026-4567123 | 12-03-1989 | Maaike | Golf |
| 5 | 8901774567 | WAAR | 40 | 06-12349900 | 026-1299667 | 04-05-1990 | Juan | Oliveira |
| 6 | 7754567890 | ONWAAR | 40 | 06-77545678 | 026-2211234 | 10-09-1990 | Johan | Huizinga |
| 7 | 4567855323 | ONWAAR | 28 | 06-11078123 | 026-4482345 | 02-06-1992 | Oene | Fries |
| 8 | 987890456 | WAAR | 36 | 06-08912345 | 026-8761267 | 09-11-1988 | Faya | Moira |
| 9 | 8901227867 | WAAR | 40 | 06-69945678 | 026-9897123 | 23-10-1989 | Anne | Otting |

*Zie figuur 6 en bijlagen VergelijkingsDataOrgineleDatabase.xlsx bestand voor beter inzicht.*

Table 2 DataMart Personeel id 7 & 8

|  |  |  |
| --- | --- | --- |
| PERSONEEL\_ID | EMPLOYEE | EXPRESS\_ERVARING |
| 0 | Jan Janssen | ONWAAR |
| 1 | Mark Bruinvis | ONWAAR |
| 2 | Sarah Ollily | ONWAAR |
| 3 | Maaike Golf | WAAR |
| 4 | Juan Oliveira | WAAR |
| 5 | Johan Huizinga | ONWAAR |
| 6 | Oene Fries | ONWAAR |
| 7 | Faya Moira | WAAR |
| 8 | Anne Otting | WAAR |

*Zie figuur 7 en bijlagen VergelijkingsDataOrgineleDatabase.xlsx bestand voor beter inzicht.*

Afbeelding met tekst, beeldscherm, scherm, schermopname

Automatisch gegenereerde beschrijving

Figure 5: Resultaat ANBC\_B EMPLOYEE tabel

Afbeelding met tafel

Automatisch gegenereerde beschrijving

Figure 6: Resultaat ANBC\_STER data mart PERSONEEL\_DIM

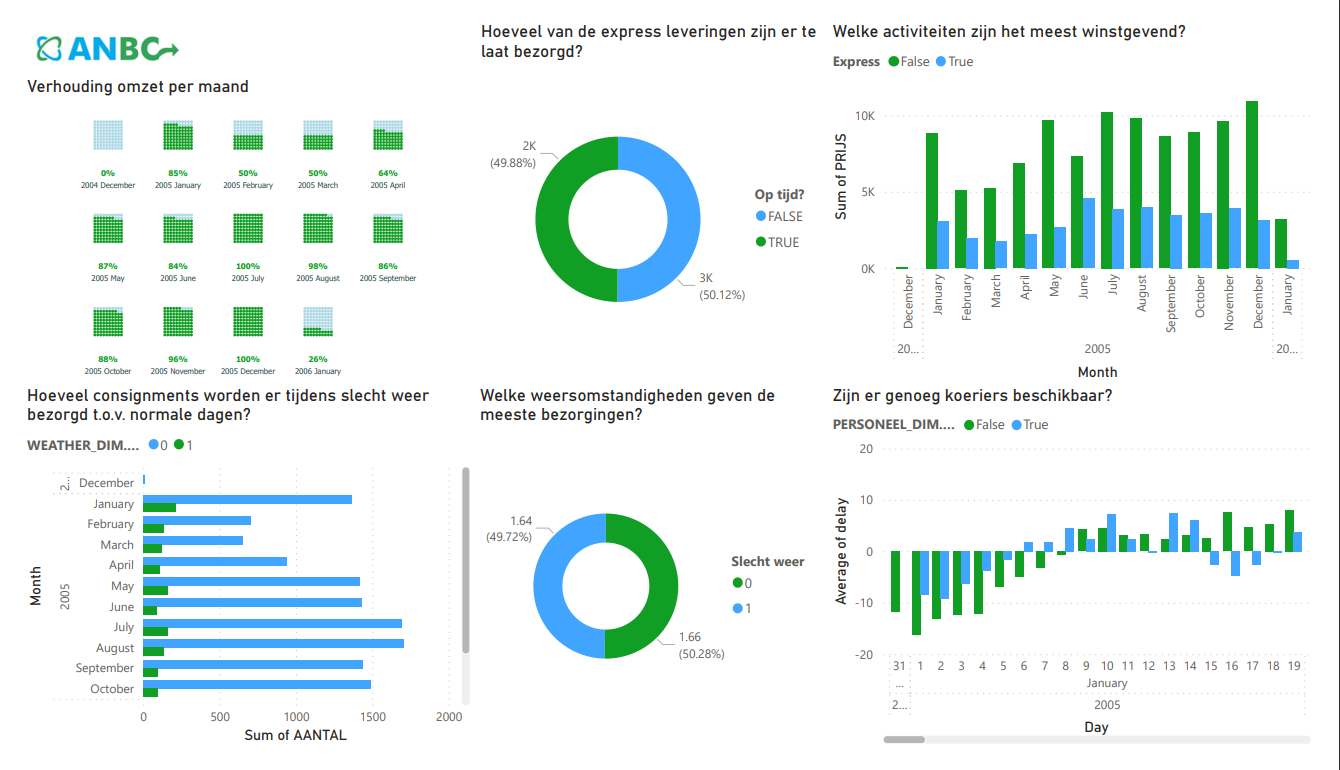


Figure 7: Screenshot dashboard

## SQL resultaten op de brondata:

Voor de volgende 2 KPI’s gaan wij controleren of de uitkomsten in het dashboard overeenkomen met de SQL berekeningen op de brondatabase.

De KPI luid: “*Hoeveel procent van de express leveringen worden niet op tijd bezorgd*.”

**Afbeelding met grafiek

Automatisch gegenereerde beschrijvingDashboard resultaat:**

Figure 8

**SQL resultaat:**

Dit is de resultaat voor berekenen op de brondatabase van de te laat bezorgde express leveringen.

USE anbc\_ster

SELECT SUM(AANTAL) as telaat

FROM CONSIGNMENT\_FACT

INNER JOIN PERSONEEL\_DIM ON CONSIGNMENT\_FACT.PERSONEEL\_ID = PERSONEEL\_DIM.PERSONEEL\_ID

INNER JOIN TIME\_DIM ON CONSIGNMENT\_FACT.PICKUP\_TIME\_ID = TIME\_DIM.TIME\_ID

INNER JOIN TIME\_DIM SCH ON CONSIGNMENT\_FACT.SCHED\_TIME\_ID = SCH.TIME\_ID

INNER JOIN WEATHER\_DIM ON CONSIGNMENT\_FACT.WEATHER\_ID = WEATHER\_DIM.WEATHER\_ID

where SCH.DATUM < TIME\_DIM.DATUM

AND PERSONEEL\_DIM.EXPRESS\_ERVARING = 1

Dit is de resultaat voor berekenen op de brondatabase van de op tijd bezorgde express leveringen.

USE anbc\_ster

SELECT SUM(AANTAL) as OPTIJD

FROM CONSIGNMENT\_FACT

INNER JOIN PERSONEEL\_DIM ON CONSIGNMENT\_FACT.PERSONEEL\_ID = PERSONEEL\_DIM.PERSONEEL\_ID

INNER JOIN TIME\_DIM ON CONSIGNMENT\_FACT.PICKUP\_TIME\_ID = TIME\_DIM.TIME\_ID

INNER JOIN TIME\_DIM SCH ON CONSIGNMENT\_FACT.SCHED\_TIME\_ID = SCH.TIME\_ID

INNER JOIN WEATHER\_DIM ON CONSIGNMENT\_FACT.WEATHER\_ID = WEATHER\_DIM.WEATHER\_ID

where SCH.DATUM >= TIME\_DIM.DATUM

AND PERSONEEL\_DIM.EXPRESS\_ERVARING = 1

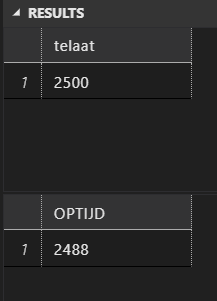


Figure 9: Resultaat te laat en optijd bezorgd.

De KPI luid: “*Hoeveel consignments worden er op slechte weersomstandigheden dagen bezorgd T.o.v. normale dagen*.”.

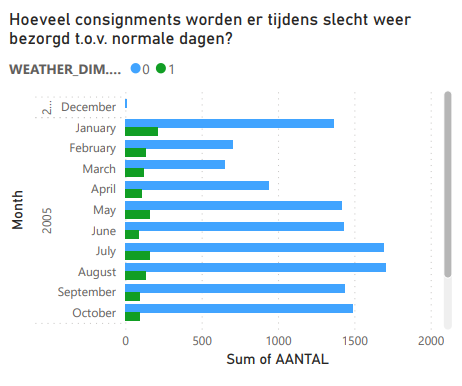
**Dashboard resultaat:**

Figure 10

**SQL resultaat:**

Dit is de resultaat voor berekenen op de brondatabase van het aantal consignments bij slecht weer.

USE anbc\_ster

SELECT SUM(AANTAL) as aantalBezorgingenbijslechtweer

FROM CONSIGNMENT\_FACT

INNER JOIN PERSONEEL\_DIM ON CONSIGNMENT\_FACT.PERSONEEL\_ID = PERSONEEL\_DIM.PERSONEEL\_ID

INNER JOIN TIME\_DIM ON CONSIGNMENT\_FACT.SCHED\_TIME\_ID = TIME\_DIM.TIME\_ID

INNER JOIN WEATHER\_DIM ON CONSIGNMENT\_FACT.WEATHER\_ID = WEATHER\_DIM.WEATHER\_ID

where WEATHER\_DIM.WEATHER\_ID = 1

GROUP BY MAAND, JAAR

Dit is de resultaat voor berekenen op de brondatabase van het aantal consignments bij goed.

USE anbc\_ster

SELECT SUM(AANTAL) as aantalBezorgingenbijgoedweer

FROM CONSIGNMENT\_FACT

INNER JOIN PERSONEEL\_DIM ON CONSIGNMENT\_FACT.PERSONEEL\_ID = PERSONEEL\_DIM.PERSONEEL\_ID

INNER JOIN TIME\_DIM ON CONSIGNMENT\_FACT.SCHED\_TIME\_ID = TIME\_DIM.TIME\_ID

INNER JOIN WEATHER\_DIM ON CONSIGNMENT\_FACT.WEATHER\_ID = WEATHER\_DIM.WEATHER\_ID

where WEATHER\_DIM.WEATHER\_ID = 0

GROUP BY MAAND, JAAR

De data wordt opgehaald vanuit consignment\_fact waar we de andere dimensie tabellen op inner joinen. Vervolgens zoeken we binnen de weather\_dim tabel naar de goed of slecht weer, dit wordt aangegeven met boolean variabele. 1 betekent slecht weer en 0 betekent goed weer. Vervolgens wordt de data gegroepeerd per maand, net zoals in het dashboard.

Van deze waarden worden de gemiddelde van uitgerekend en gerouterneerd als aantal bezorgingen bij slecht weer of als als aantal bezorgingen bij goed weer.

Afbeelding met tafel

Automatisch gegenereerde beschrijvingAfbeelding met tafel

Automatisch gegenereerde beschrijving

Figure 11 Resultaat slecht weer

Figure 12: Resultaat goed weer

## Conclusie testen:

Aan de hand van de test resultaten kunnen wij concluderen dat onze Dashboard onze KPI’s goed beantwoorden. Wij zien dat de ETL processen goed zijn verlopen en de data’s overeenkomen met die van de ANBC\_B database. Ook zien wij dat op SQL gebied op de brondata de KPI’s beantwoord worden en die resultaten overeen komen met wat er op het dashboard staat.

# APA

1. Haider, A. (2018, September 11). *Making Sense of Data Warehouse Architecture*. Data Warehouse Information Center. <https://datawarehouseinfo.com/data-warehouse-architecture/>
2. Crockett, E. (2023, January 25). *10 Best Dashboard Software & Tools for 2023 | Datamation*. Datamation. https://www.datamation.com/big-data/best-dashboard-software-and-tools-2020/

# Bijlagen

## Datavault ETL

*Dubbelklik om volledige script te bekijken*



## Datamart ETL

*Dubbelklik om volledige script te bekijken*



## Datamart.sql

*Dubbelklik om volledige script te bekijken*



## Datavault.sql

*Dubbelklik om volledige script te bekijken*

