ONDERZOEKSVERSLAG DATA WAREHOUSE EN BUSINESS INTELLIGENCE

The Cloud Consultants

Versiebeheer

Versie	Auteurs	Datum	Opmerking
0.1	A. van Dalen S.A. Twardowski	29-10-2020	Eerste versie
0.2	A. van Dalen	02-11-2020	BI en software toegevoegd

Inhoudsopgave

Introductie

Om een gedegen afweging te maken of het gebruik van een data warehouse en business intelligence voordelen bieden voor de organisatie is het belangrijk om een aantal vragen te beantwoorden.

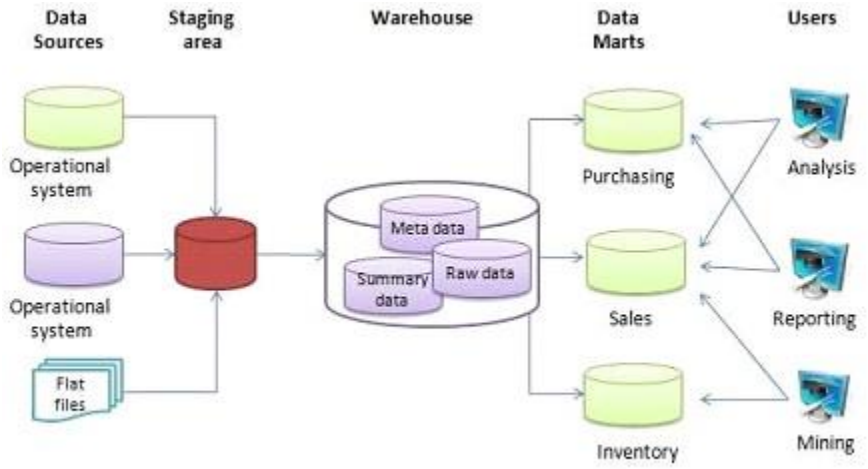
Dit onderzoeksrapport zal antwoord geven op de volgende vragen:

- Wat is een data warehouse en wat is het nut van een data warehouse?
- Welke terminologieën onderkennen we binnen een data warehouse en wat houden deze terminologieën in?
- Welke verschillende data warehouse schema's bestaan er en wanneer worden deze schema's toegepast?
- Welke software kan worden gebruikt voor het opzetten van een data warehouse en welke ondersteunende software wordt gebruikt om de data warehouse te vullen met bedrijfsdata?
- Wat is business intelligence en wat is de toegevoegde waarde voor een organisatie?
- Welke software kan worden gebruikt om data uit de data warehouse om te zetten naar informatie die waarde toevoegt voor de organisatie en deze informatie kan visualiseren?

Door antwoord te geven op deze vragen wordt duidelijk wat een data warehouse en business intelligence kunnen betekenen voor de organisatie en kan er een goede afweging worden gemaakt of een data warehouse en business intelligence passen bij de huidige organisatie.

Data warehouse: definitie

Een data warehouse is een systeem dat wordt gebruikt voor rapporteren en data analyse. Een data warehouse is een onderdeel van business intelligence waarop later in dit onderzoek wordt ingegaan. Een data warehouse is een centrale verzameling van data uit een of meerdere bronnen. Een data warehouse bevat huidige en historische data op een centrale plek om te gebruiken voor het maken van data rapportages en analyses.



Figuur 1 Schematische weergave van een data warehouse en omliggende entiteiten

De data voor een data warehouse wordt verkregen uit verschillende bronnen, denk bijvoorbeeld aan een operationele database welke in een bedrijf wordt gebruikt met gegevens over klanten, orders, financiën, voorraad enzovoorts. De data in een data warehouse wordt na het invoegen in principe niet veranderd en wordt niet gebruikt voor de dagelijkse bedrijfsvoering maar uitsluitend voor analyse en rapportage doeleinden.

Om data gereed te maken voor een data warehouse vinden processen plaats. Deze processen zijn: extractie, transformatie en laden, hierover meer in het hoofdstuk terminologieën.

Karakteristieken

Een data warehouse heeft een aantal basis karakteristieken, namelijk:

De data in een data warehouse draait om onderwerpen binnen een organisatie. Denk hierbij aan inzicht krijgen in patronen van verkoop data van klanten of voorraad informatie van een magazijn.

De data in een data warehouse is geïntegreerd vanuit een of meerdere verschillende bronnen. Inconsistenties uit data van verschillende bronnen moeten worden weggenomen.

Een data warehouse bevat zowel huidige als historische gegevens. De historie van gegevens over een lange periode helpen bij het vinden van patronen en helpt bij voorspellingen tijdens de analyse. Een data warehouse kan data bevatten over een periode van tot wel tien jaar of langer.

De data in een data warehouse wordt alleen gelezen en dus niet gewijzigd of verwijderd. Er zijn natuurlijk uitzonderingen op deze regel, denk bijvoorbeeld aan het verwijderen van persoonsgegevens wanneer een klant hierom vraagt conform de wet persoonsgegevens. (Wikipedia, 2020)

Terminologieën

Bij het omschrijven en het gebruik van een data warehouse worden veel technische terminologieën gebruikt. Om verduidelijking te scheppen worden in dit hoofdstuk veel voorkomende terminologieën uitgelegd.

OLTP en OLAP

OLTP staat voor online transactional processing. Een OLTP systeem bevat en verwerkt data in real-time doormiddel van transacties. Een voorbeeld van een OLTP systeem is een SQL database met ERP gegevens. OLTP richt zich op snelle verwerkingen binnen het systeem.

OLAP staat voor online analytical processing. Een OLAP systeem, bevat historische data en gebruikt ingewikkelde query's om historische data te analyseren. Een data warehouse is een OLAP systeem. OLTP richt zich op flexibiliteit boven snelle verwerkingen.(Stichdata, 2020)

ETL

ETL staat voor extraction, transform en load. ETL omslaat het ophalen van data uit bronnen(extraction), het uniformiseren en transformeren van de data(transform) en het wegschrijven van de data(load). De data wordt vaak opgeslagen in een data warehouse. (Wikipedia, 2020)

Metadata

Metadata is data die andere data beschrijft. Er zijn verschillende soorten metadata te onderkennen, een aantal van deze soorten metadata:

Beschrijvende metadata wordt gebruikt voor identificatie. Beschrijvende metadata bevat informatie over bijvoorbeeld titel, auteur en uitgever.

Structurele metadata bevat informatie over de opbouw van data. Structurele metadata bevat informatie over bijvoorbeeld types, relaties en versies.

Administratieve metadata bevat informatie over het beheer van middelen. Administratieve metadata bevat informatie over bijvoorbeeld middel soort, rechten en creatie datum.

Metadata wordt gebruikt door software om de juiste informatie te selecteren om deze vervolgens weer te geven of te transformeren. (Wikipedia, 2020)

Metadata repository

Een metadata repository bevat verschillende typen metadata:

Bedrijf metadata bevat informatie over eigenaarschap, bedrijfsdefinities en veranderingsbeleid.

Operationele metadata bevat informatie over de status van data. De data kan actief, gearchiveerd of verwijderd zijn. Ook bevat operationele metadata informatie over de historie van bewerkingen en migraties van de betreffende data.

Metadata die wordt gebruikt voor het vastleggen van linken naar bronnen waar de data warehouse gebruik van maakt. Deze metadata wordt gebruikt voor data extractie, data partitionering, data opschoning, transformatie regels en verwijder regels.

Algoritmes voor opsomming van data. Deze algoritmes worden gebruikt voor het weergeven van data. Tevens bevat het informatie over granularity, aggregation en summarizing. (Fontys University of Applied Sciences, 2020)

Granularity

Granulariteit geeft aan in welke mate er detail van entiteiten aanwezig is in een data warehouse. Hoe hoger de granulariteit des te meer informatie er is beschikbaar is over een entiteit. Er kan bijvoorbeeld worden gekozen om bepaalde historische data samen te voegen om ervoor te zorgen dat de query's die worden gebruikt om informatie weer te geven goed blijven functioneren doordat de hoeveelheid data niet uit de hand loopt. (Wikipedia, 2020)

Aggregate

Aggregaties zijn vooraf gecalculeerde samenvoegingen van samengevatte data die wordt opgeslagen in een nieuwe aggregatie tabel. De data wordt bijvoorbeeld samengevat doormiddel van het verwijderen van dimensies. Aggregaties worden gebruikt om de snelheid van het uitvoeren van query's binnen een data warehouse te vergroten. De snelheidswinst kan een factor honderd of zelfs een factor duizend zijn omdat er een veel kleiner aantal regels hoeven te worden aangesproken tijdens het uitvoeren van een query. (Wikipedia, 2020)

Data cube

Een data cube geeft informatie weer in meerdere dimensies. De dimensies worden weergegeven in de verschillende kolommen.

Een voorbeeld van een data cube:

Location="New Delhi"				
Time(quarter)	Item(type)			
	Entertainment	Keyboard	Mobile	Locks
Q1	500	700	10	300
Q2	769	765	30	476
Q3	987	489	18	659
Q4	666	976	40	539

Figuur 2 2-d weergave van verkoop data

In bovenstaand figuur is een data cube afgebeeld met twee dimensies: time en item. Om meer inzichten te krijgen in de informatie kunnen er dimensies worden toegevoegd.

Time	Location="Gurgaon"			Location="New Delhi"			Location="Mumbai"		
	Item			Item			Item		
	Mouse	Mobile	Modem	Mouse	Mobile	Modem	Mouse	Mobile	Modem
Q1	788	987	765	786	85	987	986	567	875
Q2	678	654	987	659	786	436	980	876	908
Q3	899	875	190	983	909	237	987	100	1089
Q4	787	969	908	537	567	836	837	926	987

Figuur 3 3-d weergave van verkoop data

In bovenstaand figuur is een data cube afgebeeld met drie dimensies: Time, Item en Location.

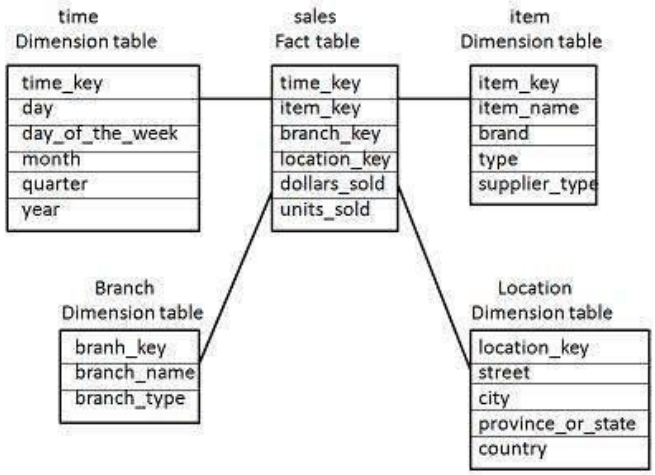
Door het gebruik van data cubes zijn duidelijke inzichten te krijgen in de informatie. (Fontys University of Applied Sciences, 2020)

Data warehouse schema's

Schema's geven een logische omschrijving van een database. Binnen een operationele database zijn we bekend met een entiteit relatie model. Binnen een data warehouse worden andere soorten schema's gebruikt. De schema's die worden gebruikt binnen een data warehouse zijn:

Star schema

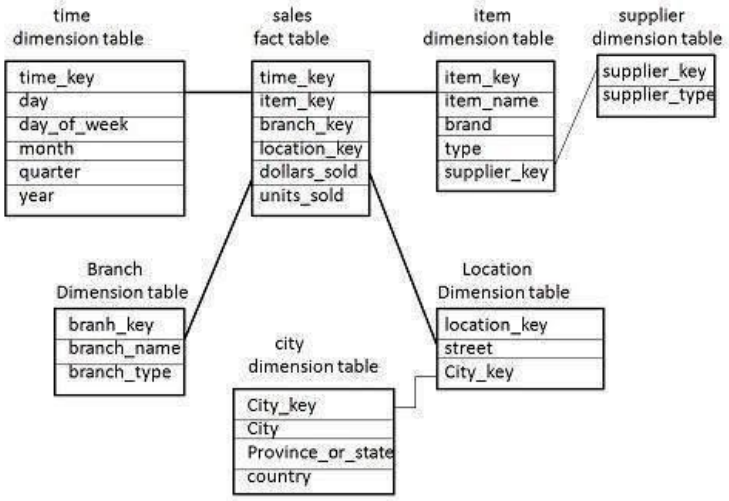
Iedere dimensie in een star schema wordt weergegeven in een aparte tabel. Een centrale fact tabel verwijst naar alle dimensie tabellen doormiddel van verwijzende sleutels. Omdat er een centrale tabel is met linken naar alle omliggende tabellen lijkt dit schema op een ster.



Figuur 4 Star schema

Snowflake schema

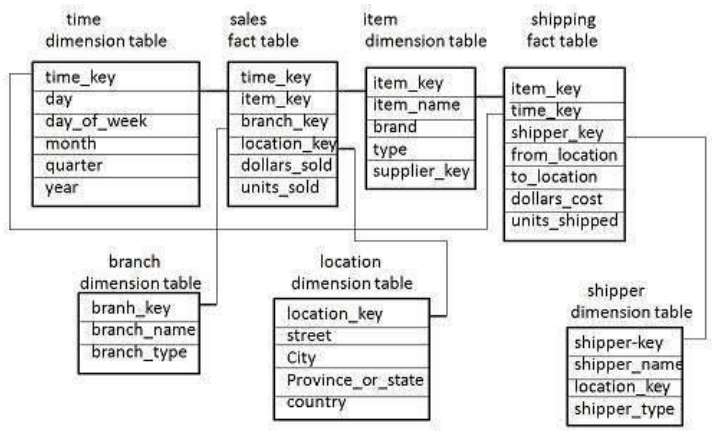
Binnen een snowflake schema worden sommige tabellen genormaliseerd waardoor nieuwe tabellen ontstaan. Er is nog steeds een enkele fact tabel welke wordt gebruikt voor verwijzingen naar de dimensie tabellen.



Figuur 5 Snowflake schema

Fact constellation schema

Een fact constellation schema of galaxy schema bevat meerder fact tabellen met verwijzingen naar dimensie tabellen. Het is mogelijk om dimensies te delen tussen de verschillende fact tabellen. (Fontys University of Applied Sciences, 2020)



Figuur 6 Fact constellation schema

Verschillen tussen een data warehouse en een operationele database

Een data warehouse is een verzameling van data. Een data warehouse is alleen geen traditionele database zoals we gewend zijn. Een data warehouse verschilt ten opzichte van een operationele database op een aantal punten (Fontys University of Applied Sciences, 2020):

Nummer	Data warehouse	Operationele database
1	Omvat verwerking van historische informatie.	Omvat verwerking van huidige data.
2	Wordt gebruikt door managers, leidinggevenden en analisten.	Wordt gebruikt door medewerkers, database administrators en database professionals.
3	Wordt gebruikt om het bedrijf te analyseren.	Wordt gebruikt om het bedrijf te laten functioneren.
4	Richt zich op het lezen van informatie.	Richt zich op het opslaan en verwerken van data.
5	Gebaseerd op het Star schema, Snowflake schema en Fact Constellation Schema.	Gebaseerd op een entiteit relatie model
6	Richt zich op het lezen van informatie.	Is applicatie georiënteerd.
7	Bevat historische data.	Bevat huidige data.
8	Levert samengevoegde en samengevatte data.	Levert primitieve gedetailleerde data.
9	Levert een samengevat multidimensionaal inzicht in de data.	Levert een getailleerd relationeel inzicht in de data.
10	Enkele honderden gebruikers	Duizenden gebruikers.
11	Miljoenen regels kunnen simultaan worden aangesproken.	Tinetallen regels kunnen simultaan worden aangesproken.
12	De opslag grootte kan variëren van 100 GB tot 100 TB.	De opslag grootte kan variëren van 100 MB tot 100 GB.
13	Hoge flexibiliteit.	Hoge verwerkingssnelheid.

Business intelligence (Wikipedia, 2020)

Business intelligence is een term die gebruikt wordt voor het verzamelen van informatie binnen een bedrijf, deze informatie analyseren om uiteindelijk tot conclusies te komen die helpen bij het ondernemen van acties die ten goede komen van het bedrijf. Door het gebruik van business intelligence kan een bedrijf inzicht krijgen in de historie en voorspellingen doen voor de toekomst. De implementatie van business intelligence staat en valt bij het opbouwen van een juiste data set die inzicht kan geven in verschillende business vragen.

Business intelligence proces

Een standaard business intelligence proces ziet er als volgt uit:

Men begint met data verzameling uit verschillende bronnen en voegt deze data samen in een data warehouse. Als bronnen kunnen verschillende systemen worden gebruikt bijvoorbeeld een enterprise resource planning systeem.

Vervolgens moet de data uit verschillende bronnen worden geüniformeerd. Omdat de verschillende bronnen wellicht data onder een andere naam opslaan of onder een andere datatype.

De verzamelde en geüniformeerde gegevens worden geanalyseerd en omgezet naar informatie die aansluit op vragen uit de business. Stel men heeft een hoop orderinformatie van afgelopen jaren. Een vraag uit de business kan zijn: welke artikelen worden het meest verkocht per maand. Dit kan inzicht geven in het bestelgedrag van klanten waardoor het bedrijf bijvoorbeeld voor kerst meer voorraad aan kan houden.

Als laatst wordt de informatie verkregen in de vorige stap gepresenteerd op een dashboard of in een business intelligence tool.

Software

Om een goede business intelligence implementatie te maken in combinatie met een data warehouse is het belangrijk om te oriënteren op ondersteunende software. Belangrijk bij de keuze van software is het doel wat het bedrijf op dit moment met business intelligence wilt bereiken. Tevens is het belangrijk om rekening te houden met schaalbaarheid. Mocht er in de toekomst de wens komen om de business intelligence strategie te verruimen moet de software hier wel mogelijkheid toe bieden.

De verschillende soorten software die we hier onderzoeken zijn opgedeeld in drie hoofdgroepen. Data warehouse software voor het opslaan van de data. ETL software voor het ophalen van data uit bronnen, het transformeren van deze data en het opslaan van de data in de data warehouse. Datavisualisatie software voor het weergeven van de informatie die we verkrijgen uit de data warehouse.

De keuze voor software wordt toegelicht in het prototype wat gemaakt wordt om de implementatie van business intelligence in combinatie met een data warehouse aan te tonen. Tevens wordt in het adviesrapport aangegeven welke software aan sluit bij de organisatie.

Data warehouse software

Er zijn een aantal aanbieders van data warehouse systemen die interessant zijn voor het bedrijf. Afhankelijk van de eisen aan het systeem, denk hierbij aan opslag, snelheid of aantal gebruikers, kan een keuze gemaakt worden tussen de verschillende aanbieders. De meeste data warehouse oplossingen zijn gebaseerd op hosting door de aanbieder. De aanbieder biedt opslag, rekenkracht en toegang tot het systeem. Deze services zijn gebaseerd op een abonnement vorm, denk hierbij aan kosten per gebruiker, per gigabyte of per processor. Een aantal van deze aanbieders zijn:

Oracle – biedt met Oracle Autonomous Data Warehouse een hosting oplossingen inclusief opslag en rekenkracht. Het kostenmodel voor deze service is prijs per processor per uur of prijs per omgeving per uur.

Google – biedt met Big Query een hosting oplossingen inclusief opslag en rekenkracht. Het kostenmodel voor deze service is per gigabyte per maand en per insert data. Operaties op de data zijn gratis.

IBM – biedt met Db2 een hosting oplossingen inclusief opslag en rekenkracht. Het kostenmodel voor deze service is per instantie per uur, per processor per uur en per gigabyte per uur.

Microsoft – biedt met Azure Synapse Analytics een hosting oplossingen inclusief opslag en rekenkracht. Het kostenmodel voor deze service is prijs per uur.

Naast deze online hosted oplossingen zijn er ook zogenaamde on-premise oplossingen. Deze kunnen door een bedrijf zelf worden gehost op een eigen server. Een on-premise oplossing is voor een proof of concept de ideale oplossing omdat zaken als beschikbaarheid en performance minder belangrijk zijn als in een productie omgeving.

Een goede on-premise oplossing voor een proof of concept is Microsoft SQL server. Hoewel MSSQL server van oorsprong een relationele database is biedt de software ondersteuning voor het implementeren van een data warehouse.

ETL software

Net als bij data warehouse software zijn er ook een hoop aanbieders van ETL software. In de basis hebben de meeste ETL tools dezelfde functies wat wel belangrijk is bij de keuze: de ETL software moet aansluiten op de bronnen die beschikbaar zijn binnen de organisatie en de gekozen data warehouse software. Een aantal van deze ETL software pakketten zijn:

Xplenty – een ETL oplossing in de cloud. Doormiddel van simpele visuele tools kunnen data stromen worden aangemaakt vanuit bronnen naar een doel. Het kostenmodel voor deze service is per connector. Een connector is een data pad van een bron naar een doel.

Oracle Data Integrator – een ETL oplossing gericht op grote bedrijven. Dit product werkt goed samen met andere Oracle producten maar ondersteund ook data warehouse oplossing als IBM Db2, Teradata en Sybase. Het kostenmodel is gebaseerd op prijs per gebruiker of prijs per processor.

Microsoft SSIS – SQL Server Integrated Services van Microsoft is een ETL tool die goed aansluit bij MSSQL server en Azure Synapse Analytics. Deze tool kan worden geïntegreerd met Azure Devops en GitHub wat het geschikt maakt voor bedrijven die al werken met deze development tools van Microsoft. SSIS is onderdeel van MSSQL server en de prijs valt dus in de licentie van de MSSQL server.

Datavisualisatie software

Om de informatie toegankelijk te maken voor de analist of manager is goede analyse software belangrijk. De data wordt omgezet naar informatie waar de gebruiker iets mee kan. Het is belangrijk dat de software goed aansluit bij de andere gekozen software pakketten. Voorbeelden van analyse software zijn:

Tableau – Tableau is er in twee vormen: desktop en server. Desktop is gericht op gebruik vanaf een machine van een gebruiker en past dus goed bij een kleinere organisatie. Server is gericht op het samenwerken en delen van informatie tussen gebruikers en past dus beter bij een grotere organisatie. Het kostenmodel is gebaseerd op prijs per gebruiker per maand.

SAP Business Objects – wanneer een organisatie SAP gebruikt is het gebruik van Business Objects een goede keuze. Business Objects is nauw verweven met het datamodel van de ERP oplossingen van SAP BI waardoor rapporten en dashboards snel kunnen worden gemaakt. Het kostenmodel voor deze software is prijs per gebruiker per jaar.

Microsoft Power BI – Power BI lijkt erg veel op andere Microsoft producten zoals office waardoor gebruikers makkelijk leren werken met deze tool. Power BI werkt goed samen met SQL server en Azure Synapse Analytics. Het kostenmodel is per gebruiker per maand of per cloud omgeving per maand.

Verwijzingen

Fontys University of Applied Sciences. (2020, 10 29). *Data Warehouse (DW) and Business Intelligence (BI) - what is it?* Opgehaald van fhict.infrastructure.com: https://fhict.instructure.com/courses/10205/pages/data-warehouse-dw-and-business-intelligence-bi-what-is-it?module_item_id=531842

Fontys University of Applied Sciences. (2020, 10 29). *Data Warehousing - Schemas*. Opgehaald van fhict.instructure.com: https://fhict.instructure.com/courses/10205/pages/data-warehousing-schemas?module_item_id=531846

Fontys University of Applied Sciences. (2020, 10 29). *Data Warehousing - Terminologies*. Opgehaald van fhict.infrastructure.com: https://fhict.instructure.com/courses/10205/pages/data-warehousing-terminologies?module_item_id=531844

Stichdata. (2020, 10 29). *OLTP and OLAP: a practical comparison*. Opgehaald van stichdata.com: https://www.stitchdata.com/resources/oltp-vs-olap/

Wikipedia. (2020, 10 29). *Aggregate (data warehouse)*. Opgehaald van wikipedia.com: https://en.wikipedia.org/wiki/Aggregate_(data_warehouse)

Wikipedia. (2020, 10 29). *Business intelligence*. Opgehaald van wikipedia.com: https://en.wikipedia.org/wiki/Business_intelligence

Wikipedia. (2020, 10 3). *Data warehouse*. Opgehaald van Wikipedia.com: https://en.wikipedia.org/wiki/Data_warehouse

Wikipedia. (2020, 10 29). *Granularity*. Opgehaald van wikipedia.com: https://en.wikipedia.org/wiki/Granularity

Wikipedia. (2020, 10 29). *Metadata*. Opgehaald van Wikipedia.com: https://en.wikipedia.org/wiki/Metadata