Cursus “Database Concepts”

Katrien Deleu, Joachim Nielandt

20 september 2021

Denkvraag		Studeeraanwijzingen	
Leerdoelen		Beeld-/geluidsfragment	
Extra informatie		Voorbeeld	
Niet vergeten		Tools / Apps	
Opdracht / Oefening		Website	
Presentatie / Power-Point		Zelfstudie	
Samenwerking		(Zelf)toets	

Legende van de gebruikte iconen.

Inhoudsopgave

Hoofdstuk 1

Inleiding

In dit eerste hoofdstuk van de cursus Database Concepts wordt een heel brede inleiding gegeven: gaande van de historiek van gegevensdragers over de evolutie van data tot de evolutie van systemen om data te beheren en te analyseren. Daarenboven worden ook een aantal gerelateerde functieprofielen uit het werkveld vermeld waarvoor de vraag veelal het aanbod van mensen met gepaste competenties ruimschoots overtreft.

In de afgelopen jaren is steeds meer duidelijk geworden dat het belang van data spectaculair stijgt. Gegevens zijn immers de primaire grondstof geworden voor het functioneren van bedrijven en van onze samenleving. Terecht wordt dan ook dikwijls de stelling geponeerd: **"data is the new oil"**.

- Een aantal basisbegrippen in verband met gegevens, informatie en databanken, databasemanagementsystemen uitleggen en in een context correct gebruiken.
- De rol van een databasemanagementsysteem toelichten.
- Het onderscheid tussen een operationele database en een datawarehouse snappen en de rol van een datawarehouse en verwante begrippen uitleggen.
- Het belang van kennis van databanken aantonen en de invloed dat *data science* op ons leven situeren.

1.1 Data versus informatie

Computersystemen worden tegenwoordig gebruikt voor de automatisering van allerlei aspecten van menselijke activiteit. Een belangrijke activiteit, die niet meer weg te denken is uit de huidige informatiemaatschappij, is het automatisch beheren van informatie met het oog op latere verwerking en hergebruik. In eenvoudige gevallen komt dit louter neer op het bijhouden van gegevens voor later onderzoek. In meer geavanceerde toepassingen mondt dit uit in complexe taken voor de ondersteuning van diverse activiteiten binnen een onderneming. Naast het doorzoeken, zijn het toevoegen, verwijderen, aanpassen en consistent houden van gegevens typische basistaken die voorkomen bij gegevensbeheer.

Doorheen deze cursus maken we een duidelijk onderscheid tussen de termen **data** (of **gegevens**) en **informatie**. Met **data** bedoelen we 'gegeven feiten'. Zo vormen getallen, symbolen, woorden, berichten, foto's, video's,.. op zich data. Voorbeelden van gegevens of data zijn dan: 45; "Piet"; 12.5; ...

Pas wanneer een gebruiker of een groep gebruikers betekenis aan deze data kan toekennen spreken we van **informatie**. 45 wordt dan de leeftijd van een sollicitant. 'Piet' wordt de voornaam van een werknemer. 12,5 wordt de gemiddelde temperatuur op 11 januari 2019. Een foto is een verzameling pixels. Dat is pure data. Wanneer een persoon op een foto in Facebook getagd wordt ontstaat er extra 'informatie'.

In de context van informatiebeheer is het verschil tussen de concepten **data** en **informatie** belangrijk. De meeste computersystemen zijn vrijwel enkel in staat om op een efficiënte manier data te beheren, betekenis wordt toegevoegd door de gebruikers. Dit heeft rechtstreeks geleid tot de term **database** die we als volgt kunnen omschrijven: **een database is een collectie van persistente data**. Het woord 'persistent' geeft hierbij aan dat de data gedurende een zekere tijd worden opgeslagen in het permanent geheugen van een computersysteem.

De huidige databasetechnologie, die ingezet wordt voor bedrijfsdatabases, is overwegend gericht op het beheer van data (gegevens), eerder dan op het beheer van de betekenis van de data. Op het wereldwijd web zien we dat er steeds meer betekenis in de data herkend wordt, denk maar aan intelligente zoekmachines zoals Google. Men spreekt in dit verband over **semantische databases**. De data worden vergezeld van betekenis en context en die informatie kan zowel door mensen als machines geïnterpreteerd worden.

Een computersysteem voor het beheer van databases wordt een **databasemanagementsysteem (DBMS)** genoemd.

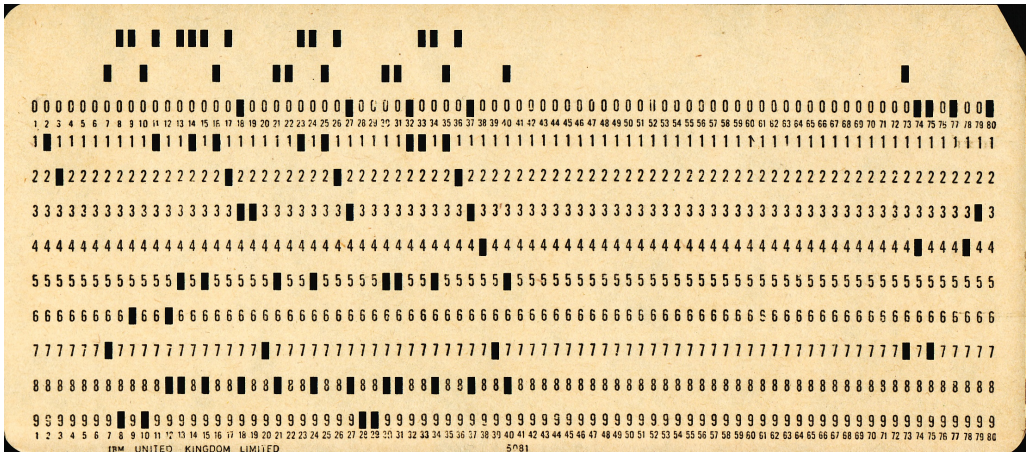
1.2 Gegevensdragers door de eeuwen heen

Gegevensbeheer is een zeer oude activiteit. Vrijwel alle moderne beschavingen danken hun bestaan grotendeels aan hun vermogen om gegevens en kennis fysiek te registreren met het oog op bewaring en kennisoverdracht. Sinds de codex van Hammoerabi (een van de oudste schriftelijke wetten, ongeveer 4000 jaar geleden) waren steen, hout, perkament en papier achtereenvolgens de dominante informatiedragers. Aanvankelijk deed men weinig moeite om de geregistreerde informatie te **structureren**, later begon de mens het nut te erkennen van het groeperen van data volgens een vaste structuur: men kan gestructureerde data veel efficiënter doorzoeken en bewaren.

Eén van de eerste voorbeelden van een gestructureerde informatiedrager: in het begin van de negentiende eeuw werden de besturingsinstructies van het automatisch **Jacquard**-weefgetouw voorgesteld door gestructureerde perforaties in kartonnen kaarten. Herman Hollerith gebruikte ook soortgelijke kaarten bij de machine die hij ontwierp voor de statische verwerking van de Amerikaanse volkstellingsgegevens van 1890. Een nieuwe gegevensdrager, de **ponskaart**, was geboren. We kennen de ponskaart ook in muziekdoosjes en draaiorgels. De ponskaart werd decennia gebruikt bij geautomatiseerde gegevensverwerking door computers, zowel voor de opslag van gegevens als programma's. In de warenhuisketen Colruyt moesten klanten tot in de jaren tachtig per artikel een ponskaart nemen. Deze ponskaarten werden aan de kassa door een kaartlezer ingelezen en zo werd automatisch een kassaticket gemaakt.

Het duurde tot de jaren 1970-1980 voordat ponskaarten geleidelijk aan werden verdrongen door sneller toegankelijke en gemakkelijker te hanteren **magnetische gegevensdragers** met veel grotere opslagcapaciteit. Aanvankelijk waren dat vooral *magneetbanden*. Magneetbanden konden heel veel gegevens bevatten, maar ze hadden het nadeel dat de gegevens sequentieel (de een na de ander) benaderd moesten worden. Dat was niet erg efficiënt.

Daarna kwamen de *magneetschijven*. Zij hadden het voordeel dat de leeskop sneller naar een bepaald blok op de schrijf kon bewegen. Dit kennen we nog van de meeste harde schrijven op onze PC's en servers.



Figuur 1.1: Ponskaart - Pete Birkinshaw. User Punchcard. https://www.flickr.com/photos/binaryape/5151286161/. CC-BY-2.0

Figuur 1.2: Magneetbank bij een IBM-computer uit de jaren zestig. U.S. National Archives and Records Administration - PDM.

Later volgden de **optische gegevensdragers** (CD, DVD, Blu-ray). Deze werden en worden gebruikt voor de archivering van gegevens en voor het afspelen van multimediaal materiaal.

Deze laatste categorie is stilaan aan het verdwijnen ten voordele van **flashgeheugen** (SSD, USB, SD-kaart). Deze zijn veel sneller dan de klassieke magneetschijven omdat er geen mechanische bewegingen (draaiende schijven, leeskoppen) nodig zijn, maar alles elektronisch gebeurt. De prijs per MB van flashgeheugen is veel duurder dan dat van een klassieke magneetschijf, maar die prijs is snel aan het dalen. De invoering van deze snelle gegevensdragers maakt onze PC's en servers vele malen sneller dan voorheen.

We kennen intussen ook al de zogenaamde "in-memory databases". Hier worden de gegevens in eerste instantie in het heel snelle, maar vluchtige werkgeheugen gehouden en gebeurt de stockering op een permanente gegevensdrager slechts in tweede instantie.

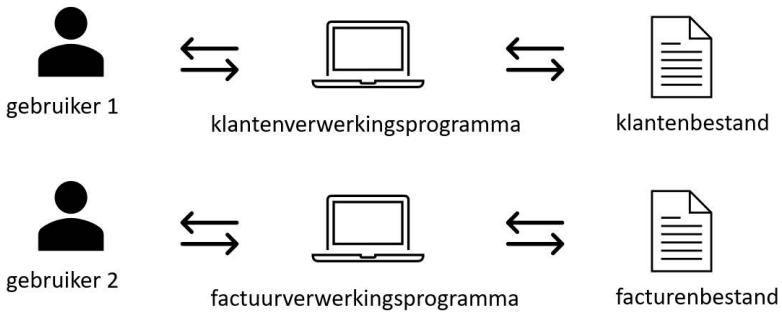
Een niet te stuiten evolutie is dat data steeds meer en meer in de cloud opgeslagen wordt. De cloud staat voor een netwerk dat met alle computers die erop aangesloten zijn een soort 'wolk van computers' vormt, waarbij de eindgebruiker niet weet op hoeveel of welke computers de software draait, de data staat en waar die computers precies staan.

De gebruiker hoeft op deze manier geen eigenaar meer te zijn van de gebruikte hard- en software en is dus ook niet verantwoordelijk voor het onderhoud.

Cloud computing is het via een netwerk – vaak het internet – op aanvraag beschikbaar stellen van hardware, software en gegevens, ongeveer zoals elektriciteit uit het lichtnet.

1.3 Databasemanagementsystemen

Oorspronkelijk werden data bijgehouden in aparte bestanden. We hadden databasetoepassingen die toegang kregen tot die gegevens, daar betekenis aan gaven en die gegevens beheerden. Met een databasetoepassing bedoelen we nu een webapplicatie, een mobiele app, een desktoptoepassing of back-endsoftware die toegang heeft tot de gegevens in de database. Enkele decennia geleden zaten die gegevens in geïsoleerde bestanden. Toepassingen hadden dan vaak exclusieve toegang tot die bestanden (zie Figuur 1.3).



Figuur 1.3: Interactie van gebruikers, via een programma, met data bestanden.

Die softwaretoepassingen (client-toepassingen) moesten exact weten waar de bestanden waren opgeslagen en welke interne structuur ze hadden. Op die manier konden zij die gegevens inlezen, wegschrijven en er betekenis aan geven. Deze betekenis was beschreven in de toepassingen, niet in de bestanden zelf. Dit had duidelijke nadelen:

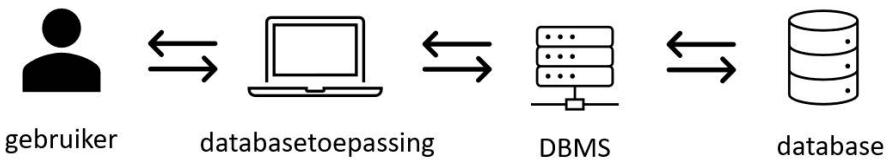
- Gegevens worden gescheiden opgeslagen. Klantgegevens zitten in andere bestanden dan factuurgegevens.
- Er ontstaat duplicatie. Bepaalde gegevens over klanten zullen zowel in het klantenbestand als in het facturenbestand voorkomen.
- De applicaties moeten de interne structuur van de bestanden kennen.
  - Dat maakt de applicaties complexer.
  - Als de interne structuur van een bestand wijzigt of het bestand wordt verplaatst , moet ook de applicatie herschreven worden.
- De gegevens zijn niet goed geïntegreerd. Het vraagt soms toegang tot diverse bestanden om alle gegevens over een verkoop of een klant te verzamelen.
- Buiten de applicaties om, is het nagenoeg onmogelijk om een betekenisvolle toegang tot de gegevens te krijgen.

In de jaren 1970 ontstaan er geïntegreerde databases. Dit betekent dat we de diverse gegevens in één database stoppen.

Een databasemanagementsysteem (DBMS) is de softwarecomponent van een databasesysteem die instaat voor het beheer van de databases. Het DBMS heeft als doel het beheren en bewaren van data in de database en gebruikers toe te laten om, via een databasetoepassing, deze data te manipuleren en te doorzoeken (zie Figuur 1.4).

Het DBMS neemt dus alle beheerstaken omtrent de opgeslagen data op zich:

- De fysieke opslag van de gegevens regelen.



Figuur 1.4: De DBMS schermt rechtstreekse interactie met de database af.

- De juistheid en de volledigheid van de gegevens controleren.
- Het zo snel mogelijk doorzoeken van de data (**performantie**) mogelijk maken.
- De gegevens tegen ongeautoriseerd gebruik beschermen.

Deze taken worden dus bijgevolg uit de applicaties gelicht waardoor deze minder complex worden. Het DBMS houdt in de database ook een beschrijving van de structuur van alle opgeslagen data bij:

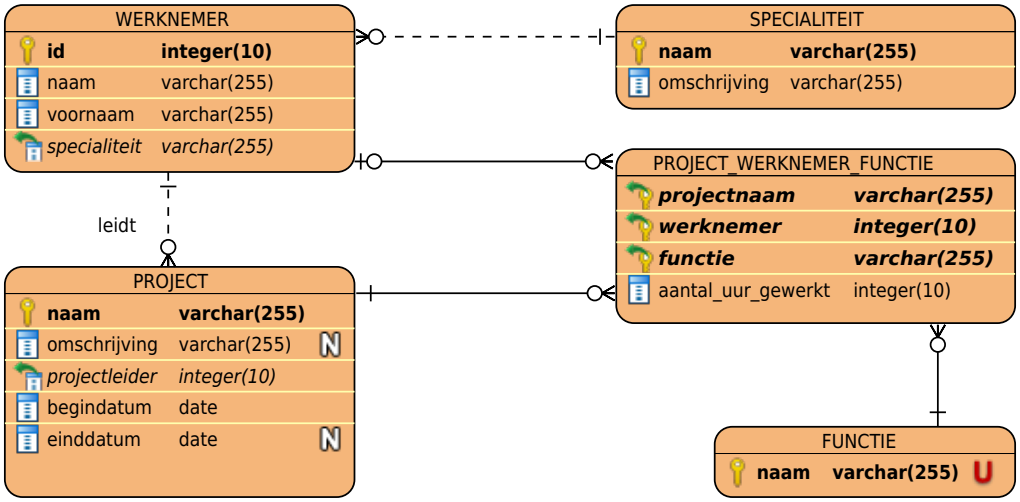
- Welke soort data worden er opgeslagen?
- Welke verbanden zijn er tussen de data?
- Welke waarden kunnen er opgeslagen worden?
- Welke gebruikers zijn er geregistreerd?
- Wie is de eigenaar van bepaalde data?
- Welke rechten hebben ander gebruikers t.o.v. bepaalde data (lezen/wijzigen/schrappen)?

Deze beschrijving wordt **meta-data** genoemd (data omtrent de data) en wordt in de database opgeslagen in een onderdeel dat de **data dictionary** genoemd wordt.

De voordelen van een geïntegreerde database beheerd door een DBMS zijn duidelijk:

- De informatie wordt in één database bewaard waardoor het evidenter is om relaties tussen gegevens te leggen (bijv. een verkoop en een factuur).
- Wanneer de database goed ontworpen is (cf. normalisatie uit Hoofdstuk 3), dan worden gegevens niet gedupliceerd, maar slechts één keer bijgehouden. Er kan immers vanuit een factuur verwezen worden naar een klant. Dat verlaagt de kans op fouten.
- De betekenis van de gegevens zit vervat in de database (in het data dictionary). Dat maakt het gemakkelijker om nieuwe applicaties te bouwen die deze data gebruiken.
- De applicaties moeten zich geen (of minder) zorgen maken over de interne structuur van de gegevens, de performantie van de databasebevragingen, de controle op bepaalde facturen die de gegevens correct houden, gebruikerstoegang en beveiliging. Dit wordt door de DBMS afgehandeld op een eenduidige manier.
- De DBMS voorziet ook een gestructureerde, afgeschermde toegang tot de gegevens buiten de applicaties om, bijv. via SQL.

Momenteel zijn de meest gebruikte databasesystemen voor de dagdagelijkse verwerking van operationele taken binnen bedrijven en organisaties de zogenaamde **relationele databasemanagementsystemen (RDBMS)**. Volgens het relationeel databasemodel worden databases logisch gezien als verzamelingen van samenhangende tabellen. Alle relationele databasesystemen werken met een standaard datadefinitie- en datamanipulatietaal, namelijk **SQL (Structured Query Language)**. De meest bekende RDBMS-systemen zijn open-source systemen als **MySQL** en **PostgreSQL**, alsook commerciële systemen als **Oracle RDBMS**, **Microsoft SQL Server** en IBM DB2.



Figuur 1.5: Voorbeeld van een ER diagram.

Data modeler

Een data modeler is in staat om een bedrijfsproces te analyseren en hieruit af te leiden welke data en in welke structuur deze data dient opgeslagen te worden. In Figuur 1.5 zie je een visueel model van de structuur van een database, een zogenaamd entiteit-relatiediagram of ERD.
Vermits data voorkomt in vrijwel elke IT-toepassing vormt het modelleren van data een **essentiële basiscompetentie** voor elke informaticus. Het is dan ook het hoofdthema van deze cursus Database Concepts.

Database administrator (DBA)

Een database administrator is iemand binnen een onderneming die technisch verantwoordelijk is voor de implementatie en het onderhoud van de databases. Enkele belangrijke taken van een DBA zijn:

- Het plannen van initiële benodigde opslagruimte voor nu en in de toekomst, ten bate van zowel het databasemanagementsysteem als van de data zelf.
- Het aanpassen van de databasestructuur naar de eisen en wensen van de applicatieontwikkelaar(s).
- Instaan voor de kwaliteit van de data.
- Het beheren van toegangsrechten van de verschillende gebruikers.
- Het herstellen na falen (backup & restore).
- Het observeren, controleren en optimaliseren van het hele databasesysteem.

1.4 Het toenemend belang van gegevens en informatie

Steeds meer gegevens worden op een digitale manier gecreëerd, beheerd en gebruikt. In het begin van het informatietijdperk hadden we grote mainframes die hun eigen bestanden en databases hadden en die gegevens gebruikten binnen het kader van één bedrijf.

Met de spectaculaire groei van het internet vanaf de jaren ’90 van de vorige eeuw worden gegevens wereldwijd toegankelijk. Alles kan vanop afstand: het regelen van een verkoop, gegevens beschikbaar maken en manipuleren, toepassingen gebruiken en databanken bevragen.

Gordon Moore heeft in de jaren 1960 al voorspeld dat computers elke twee jaar zouden verdubbelen in snelheid. Dit, en de beschikbaarheid van het wereldwijde web, zorgt voor een explosieve toename van het computergebruik en van het gebruik van gegevens.

De connectiviteit van gebruikers neemt nog toe met de komst van de smartphone. In de 21ste eeuw heeft iedereen een computer op zak. We zijn permanent verbonden met het internet, niet alleen op kantoor, maar ook thuis, op de trein, op vakantie, tijdens een muziekfestival, enzovoort. Overal gebruiken we gegevens en sturen we via sociale media zelf gegevens de wereld in. We laten een permanent spoor van gegevens achter.

Dit heeft gezorgd voor een disruptie. Hiermee bedoelen we dat bestaande economische en maatschappelijke modellen op hun kop gezet worden. De voortdurende connectiviteit, de steeds grotere beschikbaarheid van gegevens van allerlei aard en het steeds intelligentere gebruik van deze gegevens maakt dat hele bedrijfstakken verdwijnen of op hun kop gezet worden.

We kunnen hier tal van voorbeelden geven:

- De klassieke brievenpost is marginaal geworden in vergelijking met het gebruik van e-mail, WhatsApp berichten, sms, enzovoort.
- Zeldzame producten zijn nu voor iedereen van overal beschikbaar (cf. The long Tail).
- Marketing en reclame is niet enkel meer een zaak van televisie en folders, maar is meer en meer digitaal waarbij elke individu zijn eigen persoonlijke reclame krijgt.

- De klassieke platenindustrie werd bijna van de kaart geveegd door het streamen van muziek, waarbij wij die muziek voorgeschoteld krijgen die het best aansluit bij onze persoonlijke smaak.

- Klassieke reisagentschappen verdwijnen of vervullen een totaal andere rol. Wij plannen en boeken onze reis via diverse webtoepassingen en mobiele apps.

- …

Het aantal voorbeelden is legio. De gemeenschappelijke factor is dat al die nieuwe industrieën en diensten steunen op de beschikbaarheid van grote hoeveelheden gegevens en het steeds intelligenter gebruik van deze gegevens. Naast de klassieke operationele database zien we nieuwe fenomenen zoals data warehouses, business intelligence, data analytics, big data, the internet of things, enzovoort.

1.5 Datawarehouses

Een klassieke RDBMS is een ideaal platform voor zogenaamde operationele of transactionele databases. Dit zijn de databases die de dagelijkse transacties bevatten. We denken hier aan het beheren en factureren van verkopen, het registreren van nieuwe personeelsleden, enzovoort. In een dergelijke database komen er voortdurend toevoegingen, wijzigingen en schrappingen voor. Deze databases zijn ideaal voor de dagelijkse transacties, maar maken rapportering over het verleden niet altijd eenvoudig. Daarom hebben we een datawarehouse nodig.

Het **datawarehouse (DWH)** is een database waar de data zo gestructureerd is dat deze rapportering en analyse vlot toelaat. Het vormt de centrale opslagplaats voor verder te analyseren data. Het datamodel in een datawarehouse is zo opgebouwd dat de database ideaal is om gegevens te raadplegen en te analyseren, maar niet geschikt is voor dagelijkse wijzigingen. Dat vraagt andere gegevensstructuren. Een datawarehouse is gebouwd in functie van raadpleging en niet van wijziging.

Een datawarehouse zal ook gegevens groeperen uit verschillende bronnen: de operationele databases uit verschillende systemen, allerhande bestanden, externe gegevens, enzovoort. Deze gegevens moeten worden gefilterd en omgezet naar de structuur van het DWH. Dit is een proces dat geregeld herhaald moet worden aangezien de operationele databases voortdurend met nieuwe gegevens gevuld worden. Dat proces van het overzetten van de gegevens uit diverse gegevensbronnen naar een datawarehouse noemen we **ETL (Extract, Transform, Load)**. Er bestaan diverse ETL-tools die dat proces van het extraheren, het transformeren en het laden van data vanuit de operationele systemen zo automatisch mogelijk laten gebeuren.

Het belang van een goed doordacht DWH mag niet onderschat worden, via het DWH moet immers één versie van de waarheid **(single version of the truth)** gegarandeerd worden. In een operationeel systeem zitten gegevens vaak verspreid over verschillende systemen: diverse RDBMS'en, Excel bestanden, enzovoort, waarbij dubbele of tegenstrijdige waarden niet uitgesloten zijn. Met een DWH moet een eind gemaakt worden aan het bestaan van verschillende informatiesilo's in de organisatie en de daarbij horende tijdrovende gegevensconsolidaties en controles op datakwaliteit en integriteit. Gegroepeerde gegevens over een klant bevinden zich op één plaats en bevatten de volledige en gezuiverde gegevens over die klant.

Voordelen van het gebruik van datawarehouses

- Een **"single version of the truth"** wordt gegarandeerd

- Operationele systemen worden niet extra belast tijdens het uitvoeren van data-analyse-taken.
- Historie (vb. wijzigingen woonplaats klant) worden bijgehouden.
- De opslagstructuur is gemodelleerd in functie van uiterst performante analyses.

Datawarehouse developer

Aangezien het doel van het opzetten van een datawarehouse erin bestaat om rapportering en analyse te ondersteunen, worden de inhoud en het ontwerp ervan bepaald door de eisen van de organisatie. In de praktijk betekent dit dat businessmedewerkers definiëren welke informatie hen helpt betere beslissingen te nemen, waarna experten een datawarehouse en ETL-proces kunnen opzetten zodat de businessmedewerkers de juiste rapporten kunnen krijgen.

1.6 Data-analyse

In deze sectie worden er kort een aantal termen uitgelegd die in de wereld van de data-analisten vaak aan bod komen. Sommige termen zijn zeer breed, andere meer specifiek. Het is voornamelijk belangrijk dat je ze kort kan schetsen en een plaats kan geven. Steeds vaker zie je ook het concept *artificiële intelligentie* passeren, wat een zeer ruim begrip is. De technieken die hieronder worden geschetst kunnen vaak ook onder deze noemer worden geparkeerd, hoewel de term ook vaak misbruikt wordt.

1.6.1 Business Intelligence

Business intelligence (BI) start met het verzamelen van gegevens binnen de eigen handelsactiviteit. Het kan omschreven worden als het proces van gegevens omzetten in informatie, dat vervolgens zou moeten leiden tot kennis en aanzetten tot adequate actie.

Business intelligence heeft als doel competitief voordeel te creëren en organisaties slimmer te kunnen laten werken. Het wordt als een waardevolle kerncompetentie beschouwd.

Business intelligence is gericht op het verzamelen en analyseren van informatie over klanten, beslissingsprocessen, concurrentie, markttoestand en algemene economische, technologische en culturele trends, teneinde beslissingsondersteunende informatie (intelligence) te verkrijgen.

In een organisatie hebben managers op diverse niveaus behoefte aan verschillende soorten informatie, dit veelal op een zeer geaggregeerde wijze. Met geaggregeerde gegevens bedoelen we gegevens die uit verschillende bronnen werden samengevoegd en vaak ook gefilterd zijn. Een voorbeeld hiervan is het gemiddelde inkomen van mannelijke Belgen tussen 35 en 45.

Typisch voorbeeld van een methode die hiervoor ingezet kan worden is de **“Balanced Scorecard”** waarin **KPI’s (Key Performance Indicators)** gevisualiseerd worden. Dit houdt het topmanagement op de hoogte van de mate waarin de vooropgestelde targets al dan niet bereikt worden. Een KPI is een precieze maatstaf die aangeeft hoe goed een onderneming presteert op een bepaald terrein. Wanneer een manager bijvoorbeeld het percentage teruggestuurde producten uit de categorie kleding voor de regio Vlaanderen kent, kan hij nagaan of dit overeenkomt met de vooraf gestelde doelen.

Op het tactisch en operationele niveau worden veelal **“Dashboards”** gebruikt. Een dashboard is één pagina waarop data op een overzichtelijke manier visueel voorgesteld worden en waarop

je vaak zelf kunt filteren op diverse parameters. Via deze interface worden eveneens KPI's voorgesteld. De focus ligt hier eerder op de ondersteuning van dagdagelijkse operationele acties (zie Figuur 1.6.

Figuur 1.6: Management dashboard voor visualisatie KPI's.

Bij het analyseren van gegevens, met oog op het verkrijgen van belangrijke bedrijfsinzichten die het nemen van efficiënte beslissingen kunnen ondersteunen, wordt vaak gebruik gemaakt van OLAP-tools (Online Analytical Processing) waarmee gegevens op een interactieve wijze gevisualiseerd worden.

BI-analist

Een BI-analist analyseert data en creëert rapporten en analyses met nieuwe inzichten die door de business gebruikt kunnen worden. Dit profiel vormt bij uitstek een brugfunctie tussen Business en IT.

Data steward

Een data steward is een ondersteunende ICT'er, verantwoordelijk voor de kwaliteit en de consistentie van de data.

1.6.2 Datamining

Het proces van datamining kan grofweg beschreven worden als een reeks methodes en technieken die dienen om patronen te herkennen in grote hoeveelheden data. Het is belangrijk hier te onthouden dat het echt om de *patronen* gaat, en niet om de analyse of toepassing van de geleerde resultaten.

Eens men met datamining een interessant patroon gevonden heeft kan men dan aan de slag om verdere analyse mogelijk te maken.

Er zijn verscheidene voorbeelden van datamining toepassingen en ze worden toegepast in een ruime hoeveelheid aan domeinen. Zo worden er bijvoorbeeld patronen ontdekt in het rijgedrag van automobilisten. Als er op een bepaalde plaats plotseling sterk geremd wordt zal dit potentieel

als een *anomalie* kunnen herkend worden, wat dan kan leiden tot een waarschuwing of aangepaste wegsignalisatie. Datamining wordt ook gebruikt om je gedrag in winkels te analyseren. Het patroon van je aankoopgedrag wordt gebruikt om je aangepaste reclame of suggesties te kunnen toesturen.

1.6.3 Machinelearning

Machinelearning (ML) algoritmen dienen om op basis van reeds voorhanden data (trainingsdata) een model op te bouwen. Dit model stelt dan de patronen en typische kenmerken voor die te vinden zijn in de trainingsdata. Op basis van het model kan men dan *nieuwe* (tot op heden onbekende) data analyseren.

Een valkuil die kan optreden in machinelearning is dat de trainingsdata te *nauw* is, met andere woorden, dat de trainingsdata de realiteit niet voldoende beschrijft. Daardoor kan er in het model *bias* optreden (een zekere voorkeur tonen, bepaalde zaken voorrang geven, ...), wat mogelijk niet gewenst is. Het is daarom belangrijk steeds in het achterhoofd te houden welke data je aan het systeem voedt.

Een voorbeeld van machinelearning technieken is bijvoorbeeld het filteren van e-mail. Hierbij kan op basis van een trainingsdata een automatische filtering of tagging plaatsvinden van je e-mailberichten. Een ander voorbeeld is een chatbot, die op basis van een grote hoeveelheid gekende tekst *leert* hoe je moet chatten. De kwaliteit van je input is wederom belangrijk[1].

De supercomputer Watson van IBM heeft de onderzoekslaboratoria verlaten en wordt reeds ingezet in verscheidene Amerikaanse ziekenhuizen. Door symptomen en medische dossiers van patiënten te toetsen aan de enorme hoeveelheden literatuur die hij in zijn geheugen heeft opgeslagen, kan hij een diagnose stellen. Met een hoge efficiëntie: analyse van de resultaten toonde aan dat diagnoses van Watson voor 90 procent betrouwbaar waren, tegenover 50 procent voor menselijke artsen.

Je zal machinelearning soms ook als *predictive analytics* omschreven zien.

1.6.4 Predictive analytics

Het specifiek proberen voorspellen van toekomstige of onbekende evenementen wordt bestempeld als *predictive analytics*. Hierbij wordt statistiek, datamining en machine learning toegepast om op basis van gekende feiten een voorspelling te maken.

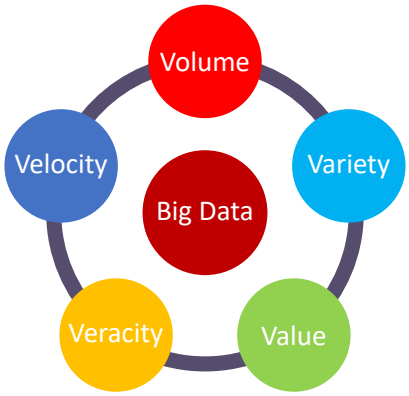
Dit wordt bijvoorbeeld toegepast in *predictive maintenance*, waarbij er wordt voorspeld wanneer een machine (of een onderdeel) door slijtage kapot zal gaan. In plaats van een duur jaarlijks onderhoud kan er dan bijvoorbeeld besloten worden een jaartje te wachten. Of, wanneer er een vroeger onderhoud nodig is wegens een potentieel dringend probleem, kan er een duur probleem vermeden worden door tijdig in te grijpen. Op basis van aankoopgedrag van een klant kan bijvoorbeeld tijdig een afhaker (iemand die zijn klandizie stop) gedetecteerd worden. Dit probleem noemt men *churn prediction*, wat zeer relevant is in de commerciële wereld.

1.7 Big Data

Men spreekt van Big Data wanneer men werkt met meerdere datasets die te groot zijn om met reguliere databasemanagementsystemen verwerkt te worden. Hierbij zijn vijf factoren belangrijk, die in het Engels mooi starten met een V, waardoor men spreekt van de *five V's of Big Data* (zie Figuur 1.7).

[1] Zo werd er een chatbot opgebouwd op basis van Twitter berichten. De chatbot bleek uiterst racistisch en brutaal uit te vallen, doordat de input data nu eenmaal niet *proper* was.

- **Volume** staat voor de hoeveelheid data. Dit gaat dus over het aantal bytes dat de dataset inneemt.
- **Velocity** is de snelheid waarmee data binnenkomt en opgevraagd wordt.
- **Variety** staat voor de diversiteit van de data. Zo gaan bijvoorbeeld verschillende databronnen dezelfde data op verschillende manieren voorstellen. Data kan gestructureerd opgeslagen zijn (e.g., in databanken), maar ook met weinig of ontbrekende structuur (e.g., in tweets, audio, video).
- **Veracity** staat voor data kwaliteit. Zijn de data proper, accuraat, compleet, …?
- **Value** gaat over de waarde die aan de data kan gehecht worden: kan er winst mee gemaakt worden?

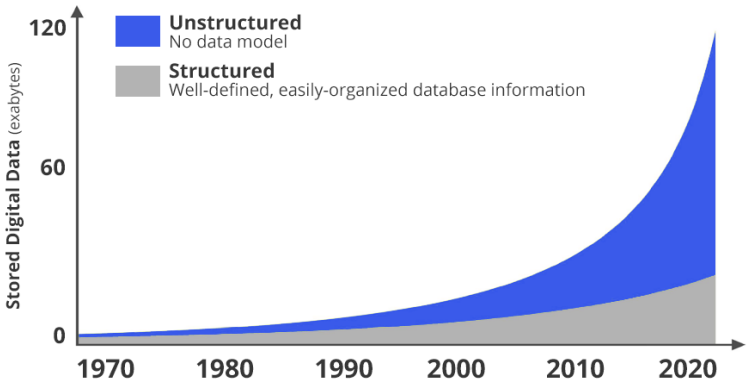


Figuur 1.7: De 5 V's van Big Data.

De beschikbaarheid van grote hoeveelheden data wordt veroorzaakt door een aantal elementen:

- Het intensief gebruik van mediatoepassingen zoals streamingdiensten, sociale media, etc. Dit veroorzaakt een stroom aan berichten, foto's, filmpjes, … wat een grote hoeveelheid ongestructureerde gegevens vormt (zie Figuur 1.8).
- De gigantische hoeveelheid gegevens die via het web verspreid worden, dragen ook bij tot de steeds grotere beschikbaarheid van big data: webshops, wikipedia, blogs, discussiefora, enzovoort.
- IoT, the internet of things. Dat begrip leggen we wat verder in detail uit. Hier krijgen we gestructureerde gegevens die in een hoog tempo door allerlei apparaten gegenereerd worden.
- De steeds grotere beschikbaarheid van grote databanken die de bedrijfsprocessen ondersteunen. We spreken hier dan over gestructureerde data in allerlei systemen. Veel van die data worden ook via de cloud ter beschikking gesteld.

De informatie en potentie van ongestructureerde data is enorm (zie Figuur 1.8). Bioscopen kunnen al op basis van tweets in de week voorafgaand aan een première inschatten hoe druk het in de eerste week gaat worden. Hoe meer positieve tweets over de film, hoe beter hij zal lopen. Goed om te weten als je instaat voor de planning: welke film krijgt volgende week de grootste zaal?



Figuur 1.8: De groei van structured en unstructured data.

Voor dit soort van Big Data is de laatste tijd een heel scala aan nieuwe soorten databases ontstaan. Vaak worden die allemaal samengevat onder de noemer **NoSQL**. Dit is echter een slechte naam die voornamelijk refereert aan het feit dat het niet om relationele databases gaat.

Een populair open-source softwareframework voor gedistribueerde opslag en verwerking voor grote hoeveelheden data is **Hadoop**. Het draait op een cluster van computers en er wordt rekening gehouden met een mogelijke uitval van een computer door **replicatie** van de data over verschillende machines.

De hoeveelheid data neemt exponentieel toe, zo is negentig procent van alle wereldwijd beschikbare data de afgelopen twee jaar geproduceerd. Pas nu wordt de invloed hiervan duidelijk. Nieuwe diensten en analyses rond Big Data zullen onze samenleving veranderen. Alleen organisaties die slim gebruik maken van de bergen aan gegevens over het gebruik van producten en diensten zullen overleven. Helaas ligt de nadruk bij Big Data-initiatieven vaak nog op het genereren en opslaan van enorme hoeveelheden data in plaats van op de analyse ervan.

Chief Data Officer (CDO)

Deze persoon is binnen het managementteam verantwoordelijk voor visie en beleid ten aanzien van data. Voorbeeld van een vraag waar een CDO zich mee bezighoudt: gaan we externe data van de consument die onze producten gebruikt verzamelen en daarmee nieuwe diensten ontwikkelen?

Data Scientist

Is de centrale spil voor een data-gedreven innovatie. Deze nieuwe rol wordt ook wel omschreven als het beroep van de toekomst: “the sexiest job of the 21^{st} century”. Dit sterk profiel combineert heel wat skills: kennis van algoritmen, statistiek, business, programmeren en communicatie. Hij is de specialist in het analyseren van grote hoeveelheden data en het toepassen van machinelearning.

Big Data Practitioner

Heeft de juiste competenties om slimme Big Data oplossingen in praktijk te brengen. De data practitioner past bestaande algoritmes toe op data maar ontwikkelt zelf geen nieuwe algoritmes.

1.8 Internet of Things (IoT)

Internet of Things (IoT) verwijst naar een concept waarbij fysieke objecten (things) verbonden worden met het internet en waarbij deze fysieke objecten zichzelf kunnen identificeren naar andere objecten toe. IoT verwijst dus niet op de eerste plaats naar een nieuwe technologie. Het combineert hoofdzakelijk bestaande technologieën zoals sensortechnologie, elektronica en netwerktechnologie. De vernieuwing zit in de toegenomen verbondenheid via het internet. Internet of Things biedt vooral kansen voor een sterkere integratie tussen de fysieke wereld en verschillende computersystemen.

Voorspellingen[2] gaan er van uit dat er tegen 2025 55.7 miljard IoT-apparaten in gebruik zullen zijn, van smart tv's, koelkasten en auto's tot slimme elektrische meters, grasmaairobots, vochtsensoren en beveiligingscamera's. Allemaal uitgerust met sensoren die gebeurtenissen capteren en alle gegevens netjes doorsturen naar het internet.

Iedereen is het erover eens dat het Internet of Things aan een onstuitbare opmars is begonnen. Het genereert een tsunami aan data waaruit een massa informatie kan worden gepuurd. Op voorwaarde dan wel dat er analytics op losgelaten worden en dit bij voorkeur real-time. IoT is één van de triggers die heel veel big data genereren en data analytics mogelijk maken.

Dit opent meteen de weg naar een heleboel nieuwe scenario's – denk maar aan de detailhandel waar je in real-time boodschappen en promoties naar de klanten kan sturen op basis van, onder andere, de locatie van die klanten in de winkel. Real-time analytics wordt mainstream, de kern van de onderneming.

IoT architect

Een IoT Architect is in staat om innovatieve oplossingen uit te werken die IoT / Mobile-apparaten en data samenvoegen.

De Powerpointpresentatie voor dit hoofdstuk vind je op Toledo.

[2] IDC - https://www.idc.com/getdoc.jsp?containerId=prAP46737220