Ongestructureerde data vormen een key issue

DW2.0

Bill Inmon

DW2.0 is de nieuwe generatie in datawarehousing. Datawarehousing begon in het midden van de jaren tachtig en sindsdien is er veel vooruitgang geboekt in architectuur, technologie en informatiesystemen. Deze nieuwe ontwikkelingen zijn allemaal verweven in DW2.0.

Bij de eerste generatie datawarehousing draaide het vooral om transactiedata die werden geïntegreerd en op disk opgeslagen. In dit vroege begin werd ook ETL geboren. Er zaten ook veel aspecten en functies niet in, omdat men toen vond dat ze niet in een datawarehouse thuishoorden.
DW2.0, de volgende generatie, heeft vele geïntegreerde features die niet in de eerste generatie werden aangetroffen:

- gekwalificeerde en bewerkte ongestructureerde data in verschillende vormen;
- geïntegreerde metadata, zowel business metadata als technische metadata;
- online high performance data, waarop updates mogelijk zijn;
- reference masterdata;
- profile data records.

Bovendien bevat DW2.0 continue time span data.

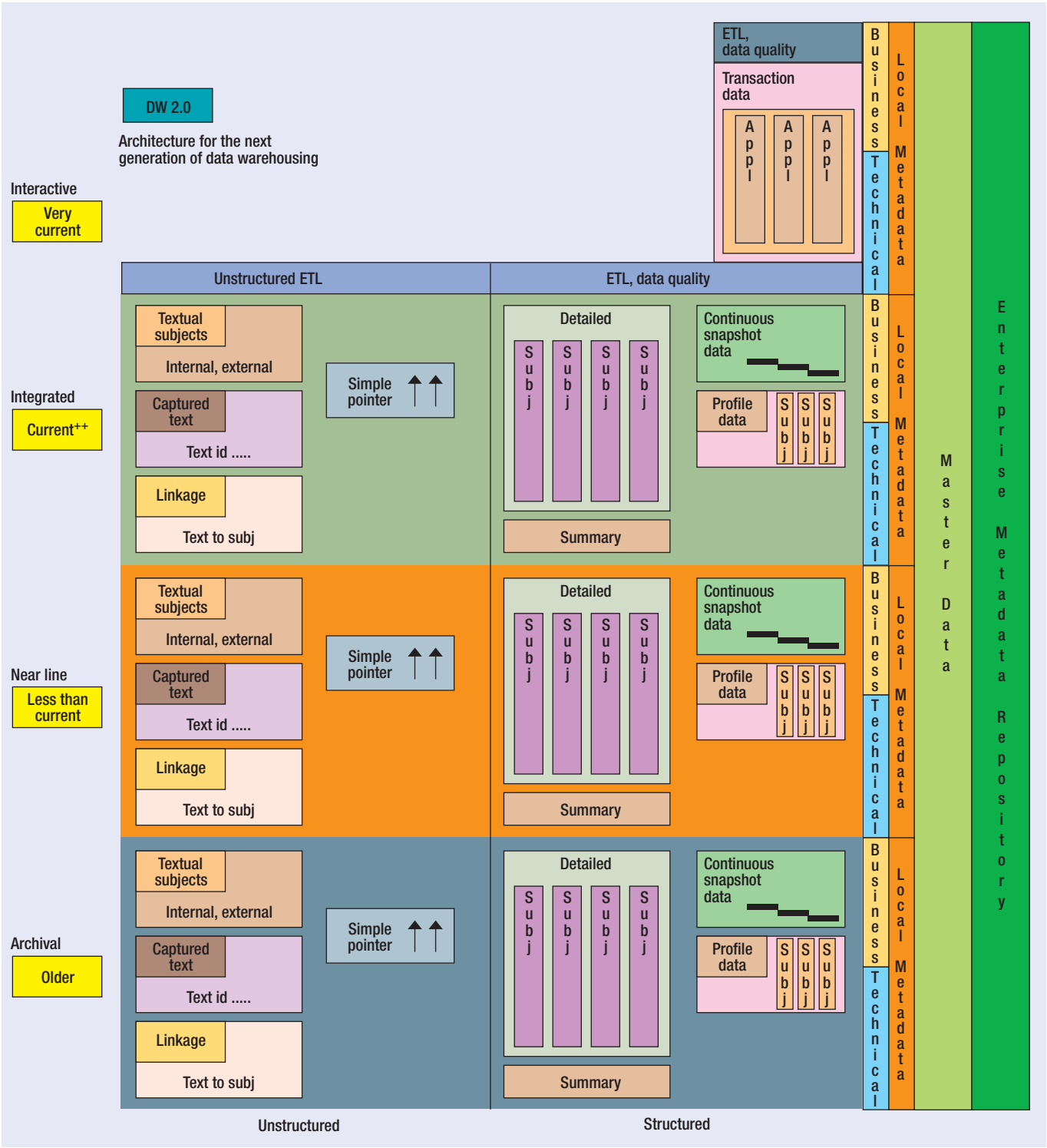
Life cycle van data

DW2.0 herkent de life span, de life cycle, van data doordat data worden verzameld, gebruikt en afgedankt. Als data het systeem binnenkomen, zijn ze vers en nieuw. Dan beginnen ze te verouderen. Eerst gaan ze over in middelbaar, dan in bejaard en uiteindelijk worden ze gearchiveerd. Al deze technologieën en verfijningen worden stevig samengebundeld in het DW2.0 datawarehouse framework. In afbeelding 1 is te zien dat DW2.0 vier belangrijke sectoren heeft: een interactieve sector, een geïntegreerde sector, een near-line sector en een archiveringssector. Data komen in DW2.0 binnen in de vorm van applicaties. In de regel voeren die applicaties transacties uit, die snelle responstijd en hoge beschikbaarheid vereisen. Normaal gesproken is er erg weinig sprake van data-integratie tussen applicaties.
Als de transactiedata naar de geïntegreerde laag doorgaan, worden ze geïntegreerd. Natuurlijk kunnen data naar de geïntegreerde laag doorgaan zonder dat ze de applicatiesector passeren. Er is een brede variatie aan data in de geïntegreerde sector. De applicatiedata die uit de geïntegreerde sector zijn gekomen, zijn geïntegreerd en omgevormd naar gedetailleerde subject area. Andere data die zich in de geïntegreerde sector bevinden zijn: tekstobjecten; captured tekst; tekstlinks; continue snapshots van data; profile data.
Bovendien zitten er in het geïntegreerde niveau lokale business metadata, lokale technische metadata en ondernemings metadata, gewoonlijk neergezet in een enterprise metadata repository.
Dezelfde vormen van data bevinden zich in de near-line sector en in de archiveringssector.

De data in de interactieve sector zijn actueel, hooguit een maand oud. Data in de geïntegreerde sector variëren in leeftijd van één dag tot twee of drie jaar oud, data in de near-line sector zijn zes maanden tot tien jaar oud, en data in de archiveringssector zijn er vanaf vijf jaar tot oneindig.

> In het algemeen vloeien de data compleet van de ene naar de andere sector

De belangrijkste factor die de locatie van de data bepaalt, is de opvraagwaarschijnlijkheid en de noodzaak van snelle toegang. De interactieve sector bevat data die zeer snel moeten worden opgevraagd en een hoge opvraagwaarschijnlijkheid hebben.
De geïntegreerde sector bevat data die met een redelijke snelheid moeten worden opgevraagd en een redelijk hoge opvraagwaarschijnlijkheid hebben. De data die zich in de near-line sector bevinden hebben een bescheiden opvraagwaarschijnlijkheid



Afbeelding 1: DW2.0 architectuur. (Copyright Bill Inmon.)

en vereisen slechts een redelijke performance. De data in de archiveringssector hebben een zeer lage opvraagwaarschijnlijkheid en hebben genoeg aan een lage toegangssnelheid. Feitelijk is het gebruikelijk om enkele data in de archieven te houden waarvan wordt aangenomen dat de opvraagwaarschijnlijkheid nul is. Bedrijven bewaren die data om statutaire redenen, terwijl er geen enkele hoop is dat ze ooit nog worden opgevraagd. In andere gevallen bewaren bedrijven data om zeker te weten dat áls de data ooit opgevraagd moeten worden, ze weten waar de data zich bevinden. Hoewel de opvraagwaarschijnlijkheid nul of bijna nul is, als ze ooit nodig zijn kunnen ze worden teruggehaald met een minimum aan investering.

Datavolumes

Als we naar de bulk van data kijken dan bevat de interactieve omgeving bijna een snufje van de data. De geïntegreerde sector

bevat meer data, de near-line sector bevat veel data en de archiveringssector de meeste data. Vanuit datavolume-perspectief zijn er dus grote verschillen tussen de sectoren. Elk van de sectoren van DW2.0 vereist technologie die optimaal is voor de betreffende sector: er is geen 'one size fits all' technologie.
In het algemeen vloeien de data compleet van de ene naar de andere sector. Er zijn twee uitzonderingen. De eerste als applicatiedata naar de integratiesector gaan om te worden geïntegreerd, de tweede als de data naar de archiefomgeving gaan. Om verschillende redenen kunnen data aanzienlijk worden getransformeerd als ze naar de archiveringssector gaan:

- om de data uit de structuur en technologie te verwijderen, die over 20 jaar misschien niet meer wordt ondersteund;
- om de data te herstructureren met het oog op snellere en flexibeler toegang tot de archiveringsomgeving enzovoort.

Lokale metadata leiden een volledig zelfstandig bestaan

Lokale metadata bestaan in de vele technologieën die in de sector worden aangetroffen. Lokale data worden 'gepasteuriseerd' en gezonden naar de enterprise metadata repository. Deze behoudt een snapshot van de metadata die in de lokale metadata-pool zitten. Als er veranderingen aan de metadata moeten worden aangebracht, wordt dat eerst op lokaal niveau gedaan, waarna de metadata naar het enterprise-niveau worden overgebracht. Metadata kunnen na verloop van tijd worden opgeslagen op enterprise-niveau als het gewenst is de wijzigingen op de metadata bij te houden.

Componenten

DW2.0 bevat vele verschillende componenten. Hierna volgen beschrijvingen van de meest interessante en meest voorkomende.

Captured text.
Captured text komt uit de ongestructureerde omgeving en kan bestaan in de vorm van e-mails, documenten, transcripties van telefoongesprekken, en andere tekstuele informatie. In de regel verkeert captured text in dezelfde onbewerkte staat als waarin het zich in de ongestructureerde omgeving bevindt. De ongestructureerde tekst is echter geselecteerd vanwege de relevantie voor de business-omgeving. Het zou niet verstandig zijn om enorme hoeveelheden ongestructureerde tekst in de DW2.0 omgeving te stoppen, tenzij die tekst belangrijk is voor de business. Daarom is de ongestructureerde tekst die zijn weg vindt naar DW2.0 van tevoren bewerkt en gefiatteerd voor doorstroom naar het datawarehouse.

Profile data.
Profile data zijn samengestelde data verzameld uit verschillende bronnen. Profile data zijn een 'thumbnail-plaatje' van heel veel andere data. Kenmerkend voor profile data is het samengestelde klantdossier, dat gemaakt kan worden met data uit zeer verschillende bronnen. Aankopen, betalingen, internetbezoek, demografische gegevens van de klant kunnen allemaal in één profile-bestand worden gezet. Als dat eenmaal is aangemaakt kan dat snel en veilig worden ingezien. Het is niet nodig overal op zoek te gaan naar gegevens en alle brondata te analyseren, als er op een gegeven moment naar een klant moet worden gekeken. Hoewel klantgegegeven zich bij uitstek lenen voor profiling, zijn er vele andere onderwerpen mogelijk.

Gedetailleerde subject area data.
Gedetailleerde subject area data zitten in het hart van het datawarehouse. Gedetailleerde subject area data zijn data afkomstig van applicaties die vervolgens geïntegreerd zijn. Als de gedetailleerde subject area data eenmaal zijn verzameld in het DW2.0 vormen ze de basis voor Business Intelligence. De gedetailleerde subject area data zijn zo fijnkorrelig, dat ze op vele verschillende manieren kunnen worden gevormd en hervormd. De gedetailleerde subject area data ondersteunen finances, accounting, sales, marketing, engineering, human resources enzovoort. De gedetailleerde subject area data hebben normaal gesproken een relationeel formaat. Elk record heeft een timestamp.

Linken van tekst aan een subject.
Als ongestructureerde data worden overgebracht naar de datawarehouse-omgeving – zelfs als ze zijn bewerkt en gescreened – kunnnen de tekstuele data nog nuttiger zijn als ze worden gelinked aan de klassieke transactiedata en gestructureerde data die zich in DW2.0 bevinden. Meestal wordt er gelinkt naar e-mailadressen en telefoonnummers. Maar ook andere links kunnen worden aangebracht naar namen en mutaties van namen. Deze data worden meestal aangemaakt nadat de tekstuele data naar het datawarehouse zijn overgebracht. Ook data die nergens aan gelinkt zijn, kunnen uiterst relevant zijn voor de business.

Continue snapshot data.
Continue snapshot data zijn data die aan elkaar gelinkt worden door een serie 'van' en 'tot' datums. De links zijn logisch, niet fysiek. Er kan geen overlapping zijn van continue snapshot data, maar er kunnen wel gaten door onregelmatigheid in zitten. Continue snapshot data zijn nuttig als er weinig variabelen zijn en als die langzaam wijzigen. Er bestaat een continue definitie van data die kunnen worden gecreëerd. Klantnamen en klantadressen zijn een typisch voorbeeld waar continue snapshot data worden toegepast.

Applicatiedata.
Applicatiedata worden overgebracht naar een omgeving met een responstijd van twee tot drie seconden. Applicatiedata zijn nooit geïntegreerd en maken de update van datawaarden mogelijk, evenals tussenvoegingen en creatie. Applicatiedata komen voor

waar de meeste data worden gegenereerd in het bedrijf, als bijproduct van de uitvoering van transacties.

Tekstuele subjects.
Tekstuele subjects zijn die subjects waarbij de tekst uit de ongestructureerde omgeving wordt georganiseerd. Tekstuele subjects kunnen intern worden gegenereerd, of extern door de creatie van een of meer ontologieën.

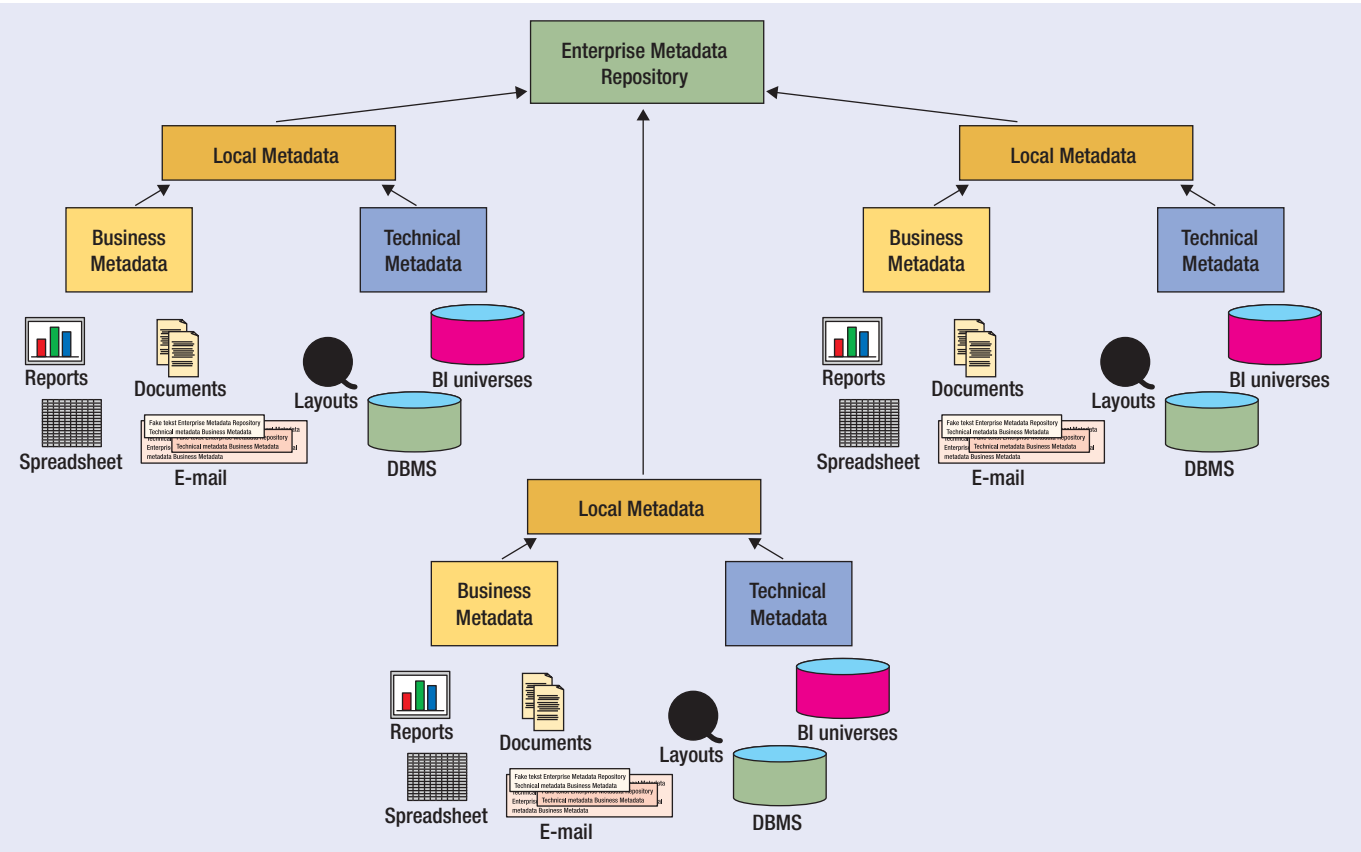
Reference data en masterdata.
Elk bedrijf bezit reference data. Er zijn veel verschillende soorten reference-tabellen. Als reference data door de gehele onderneming heen worden aangewend, is er sprake van masterdata.

Samenvattende data.
Terwijl de meeste data in DW2.0 in een gedetailleerde vorm zijn, is er plaats voor samenvattende data. Dat heeft alleen zin als de samenvatting door de hele onderneming heen gebruikt wordt. Als samenvattende data worden aangemaakt, is het ook verstandig om de regels voor samenvatting daarbij vast te leggen – welke data zijn erin meegenomen, welke data niet, hoe is de berekening tot stand gekomen, enzovoort.

Eenvoudige ongestructureerde pointers.
Pointers van de DW2.0 omgeving naar de ongestructureerde omgeving kunnen in sommige gevallen nuttig zijn. Soms zal de bulk van de data die zich in de ongestructureerde omgeving bevinden te groot zijn om naar het datawarehouse te worden overgeheveld. Toch kan er zich waardevolle informatie in de ongestructureerde omgeving bevinden.
Dan is het zinvol om eenvoudige pointers in de ongestructureerde omgeving aan te brengen, zodat toegang kan worden verkregen tot de ongestructureerde data als dat nodig is, echter wel indirect.

Metadata in DW2.0

Er bevinden zich metadata in DW2.0. Metadata vormen een van de belangrijkste componenten, omdat de metadata als het ware als zenuwstelsel van DW2.0 werken. Metadata beschrijven wat er zich in het datawarehouse bevindt en hoe data en componenten aan elkaar gerelateerd zijn. Zonder metadata zouden data in DW2.0 een grote hoop nutteloze gegevens zijn. Er zijn verschillende soorten metadata binnen DW2.0: lokale business metadata, lokale technische metadata en enterprise metadata.
Lokale metadata refereren aan metadata die zich in een bepaalde component bevinden. Enkele voorbeelden van lokale metadata zijn de metadata die zich in de BI-tool bevinden, in een DBMS directory, in een spreadsheet, in een rapport, in een data dictionary enzovoort.
In elk van deze gevallen is er een stukje technologie waar de metadata binnen die technologie zijn opgeborgen. De metadata worden volledig bewaard en beheerd binnen die technologie.
Als er metadata moeten worden toegevoegd of gewijzigd, gebeurt



Afbeelding 2: Metadata in een bedrijfsbrede metadata repository. (Copyright Bill Inmon.)

dat binnen die technologie. Met andere woorden, lokale metadata leiden een volledig zelfstandig bestaan. Het probleem met lokale metadata is dat ze onderdeel uitmaken van een grotere wereld, waarvan ze het bestaan niet kennen.

> Zonder metadata zouden data in DW2.0 een grote hoop nutteloze gegevens zijn

Elke willekeurige eenheid van lokale metadata is onbewust van het feit dat er veel meer andere metadata zijn waaraan ze gerelateerd zijn. Daarom wordt het 'lokaal' genoemd. Binnen de categorie van lokale metadata bestaan business metadata en technische metadata. Business metadata zijn metadata die tekst bevatten die betekenisvol en nuttig zijn voor de business en technische metadata zijn metadata die tekst bevatten die nuttig en belangrijk zijn voor de technicus.

Metadata repository

Binnen elke technologie in DW2.0 bevinden zich dus lokale metadata. Maar er bestaat een noodzaak om de lokale metadata te integreren, hetgeen geschiedt door de inrichting van een bedrijfsbrede metadata repository. Dat is technologie die alle lokale metadata verzamelt en samenbrengt op één plaats. Als de lokale metadata zich daar eenmaal bevinden kunnen ze worden geïntegreerd. Lokale en bedrijfsbrede metadata reflecteren een grotere metadata-structuur.

Lokale metadata, zowel business als technische, worden lokaal verzameld of bestaan lokaal. In dat geval worden de lokale metadata verzameld in de enterprise metadata repository, zie afbeelding 2. Als ze zich daarin bevinden worden de lokale data bewerkt en georganiseerd volgens de behoeften op enterprise-niveau.

Noot

In geval van discussies geeft de originele Engelstalige tekst van dit artikel de doorslag. Deze tekst is te vinden op onze website www.dbm.nl in het hoofdmenu onder Specials/Extra materiaal.

DW2.0 is een geregistreerd handelsmerk van Bill Inmon.

Bill Inmon

William H. Inmon (binmon@inmondatasystems.com) is oprichter en CEO van Inmon Data Systems, gevestigd in Castle Rock, Colorado.