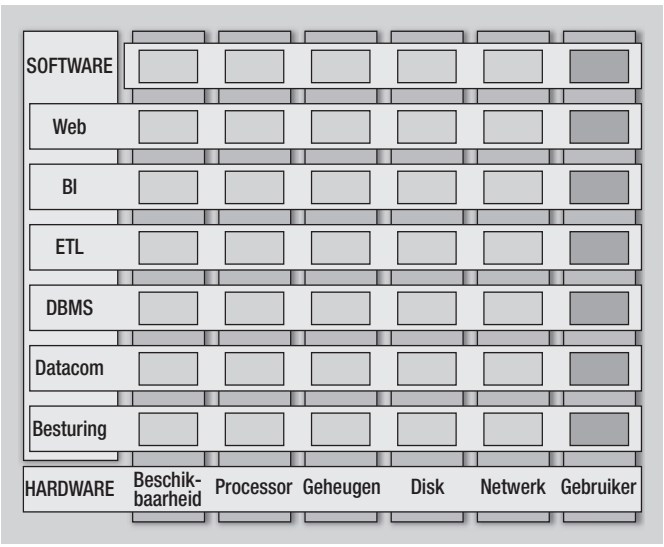
Business Intelligence voor de beheerder

Monitor het Datawarehouse

Martin Misseyer

Met een beetje geluk zal in een datawarehouse-project uitgebreid aandacht worden besteed aan het selecteren en inrichten van de infrastructuur. Met nog meer geluk worden hier (aspirant) beheerders bij betrokken die hun kennis en kunde inbrengen. Met een beetje pech is de ingebrachte kennis niet voldoende om de groei in omvang van het datawarehouse te kunnen voorspellen, laat staan de omgeving – juist vanwege het grillige gebruik – beheerbaar te houden. Met wat meer tegenslag escaleren periodiek allerlei perikelen rondom infrastructuur, vergt de bestelling van nieuwe resources te veel (doorloop)tijd, heeft beheer het gedaan, wordt de infrastructuur(keuze) vervloekt en kan er een discussie ontstaan over verwachtingsmanagement. Welkom bij de problematiek rond datawarehouse systeem-resources.

De infrastructuur in een datawarehouse staat geregeld onder druk. In sommige gevallen is er op basis van een beschikbaar budget een – te krappe – behuizing aangeschaft, groeit het datawarehouse (veel) sneller dan verwacht, en/of is de inrichting van de omgeving inefficiënt (onnodig complexe processen ontwikkeld, onhandig data(base)-ontwerp et cetera), waardoor de toch al dure resources worden opgeslokt. Tevens is beschikbaarheid van beheer- en infrastructuurexpertise een cruciale factor en ook speelt de aanwezigheid van een relevant meet- en monitor-instrumentarium een belangrijke rol. Zeker in het geval dat er onvoldoende kennis is omtrent datawarehousing in het algemeen en datawarehouse operations, kunnen instrumenten hulp bieden bij het bewaken van de kostbare infrastructuur-resources van het datawarehouse.



Afbeelding 1: Hardware-, software- en gebruikers-matrix.

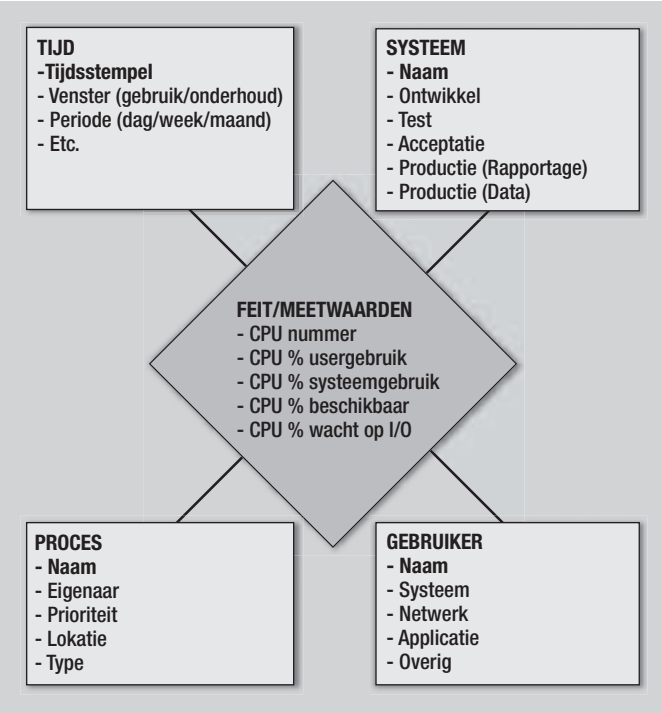
Beheerder als systeem-manager

Een doorsnee BI- en datawarehousing-omgeving omvat web-, BI-, ETL-, DBMS- en datacommunicatie-server software. Voor al deze server software geldt dat het, indien beschikbaar en actief, basis systeem-resources utiliseert als CPU (processorcapaciteit), I/O (virtueel- en/of fysiek geheugen, disk, netwerk). In besturingsterminologie gesproken, kan men stellen dat de beheerder feitelijk de 'systeem-manager' is met een bepaalde management-informatiebehoefte, hoe concreet en/of operationeel georiënteerd de beheerder ook is. De regelkring waarmee de systeem-manager mee te maken heeft is maximaal concreet, dit in tegenstelling tot een manager van een afdeling, een werkmaatschappij of directeur van een onderneming.

> Instrumenten kunnen hulp bieden bij het bewaken van de kostbare resources van het datawarehouse

De 'natural drive' van de systeem-manager is het onder controle hebben én houden van de infrastructuur en de zich hierop afspelende activiteit. Daardoor heeft de systeem-manager een tamelijk concrete informatiebehoefte, namelijk zo actueel mogelijke detailinformatie over alle zich afspelende essentiële activiteiten. Deze activiteiten zijn op hoofdlijnen weergegeven in afbeelding 1.

Een goede systeem-manager zou moeten kijken naar:
- systeemactiviteit in het verleden → rapporteren (MIS);
- actuele systeemactiviteit → monitoren (BAM);
- verwachte, toekomstige, systeemactiviteit → voorspellen (DSS).



Afbeelding 2: Basis van het dimensioneel model DWH Monitor.

Het eerste type werkzaamheden correspondeert met een traditioneel MIS (Management Information System) of BI-reporting. Het tweede type werkzaamheden correspondeert met operationele BI-reporting. Wanneer deze wordt gecombineerd met inhoudelijke reporting, komt men terecht bij BAM (Business Activity Monitoring). Op 'strategischer' niveau zou men ook kunnen denken aan een systeem-georiënteerde Balanced Scorecard (IT-BSC?). Het derde type werkzaamheden correspondeert met decision support-achtige werkzaamheden, waarvoor doorgaans een DSS-instrumentarium wordt gehanteerd (Decision Support System).

Informatiebehoefte

Wanneer men vanuit het theoretisch uitstapje terugkeert naar de operationele werkelijkheid van de systeem-manager, is te stellen dat de basis-informatiebehoefte van waaruit men kan rapporteren, op een simpele wijze vast te leggen is in enkele dimensionele modellen. Vanzelfsprekend zijn deze gebaseerd op dimensies en feiten. Voor de DWH Monitor is een basismodel te definiëren, bestaande uit de dimensies:

- Tijd, het maximale detailniveau is bijvoorbeeld elke 1, 5, 10 of 15 minuten;
- Systeem, indien sprake is van meerdere systemen (OTAP en/of multi-tier);
- Gebruiker, kan een eindgebruiker zijn, maar ook een fictieve (functionele key);
- Proces, elk te monitoren proces, een dummy voor overig.

Met het bovenstaande model kan men de volgende meetwaarden in ieder geval vastleggen:

- Beschikbaarheid, meet algemeen (van het systeem) en per softwarecomponent (BI, enzovoort);
- Activiteit, meet of proces/gebruiker niet alleen bestaan, maar ook wat ze doen;
- Percentage CPU-gebruik, meet voor hoeveel procent de CPU wordt geclaimd;
- Geheugenbeslag, meet hoeveel intern geheugen is gealloceerd aan het proces.

De basis van het dimensioneel datamodel is in afbeelding 2 schematisch weergegeven. Op basis van dit model zijn diverse varianten en aggregaten te definiëren, zoals bijvoorbeeld de beschikbaarheid per systeem, per software-component of geheugenbeslag. Naast het getoonde model bestaat er de mogelijkheid om specifieke aanvullende modellen te definiëren voor processor, virtueel geheugen, disk en netwerk. Indien men over een SMP-configuratie en/of flinke storage-omgeving beschikt is het wellicht wenselijk om te weten of en zo ja hoe, op detailniveau het besturingssysteem met de dure resources omgaat.

Een tool, een applicatie of een programma zal metingen moeten wegschrijven in één of meer specifieke (log)bestanden. Deze logbestanden zijn opgemaakt via een standaard formaat. In de meeste gevallen kan worden volstaan met het formaat zoals de (systeem)-tools rapporteren. De tabel in afbeelding 3 geeft een voorbeeld van een zelfgedefinieerd formaat, dat nagenoeg gelijk is aan de uitvoer van een systeem-tool, in dit geval *vmstat* onder Unix.

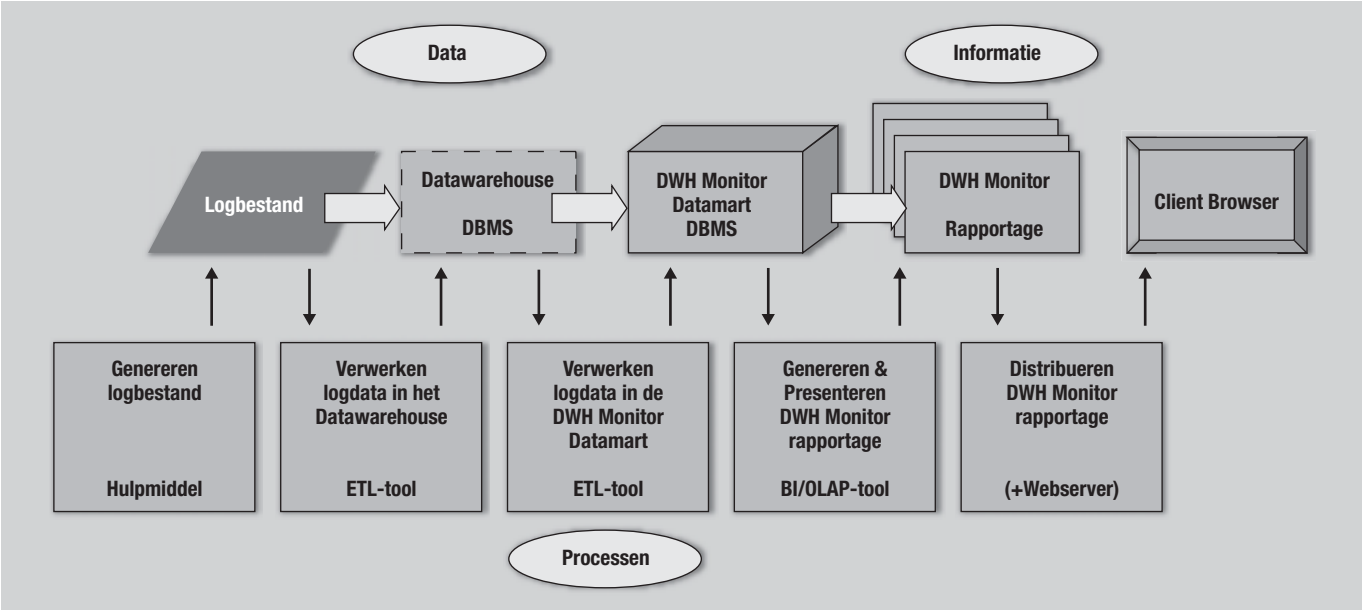
Dataverwerkingsproces

Het dataverwerkingsproces voor de DWH Monitor, zie afbeelding 4, is conform het gebruikelijke dataverwerkingsproces in een datawarehouse. Bij afbeelding 4 zijn enkele korte kanttekeningen te plaatsen:

1. Geen verwerking via een data staging area. Het is de vraag of het nodig is om de data via een staging-laag te verwerken. Gezien het feit dat het een tamelijk eenvoudig (dimensioneel) model betreft, lijkt dit overbodig.
2. Waar vastleggen? Men kan het zich zo uitgebreid (lees: moeilijk) maken als men wil. In principe zou men alle gemodelleerde data-, dimensie-, feit- en aggregaat-tabellen in een datamart kunnen onderbrengen. Enkele puristen zouden willen pleiten voor een opslag van de detaildata in het datawarehouse. Echter,

```
# 1 = Datum/tijd;
# 2 = Virtueel Geheugen max.
# 3 = Gebruik virtueel geheugen;
# 4 = CPU Systeem %;
# 5 = CPU Gebruiker;
# 6 = CPU Beschikbaar %;
# 7 = Wacht I/O;
#1                  2       3    4    5    6    7
2004-01-01-16:05:00 4096MB  26%  0.0  0.1  99.9 0.0
2004-01-01-16:10:00 4096MB  28%  0.0  0.5  99.5 0.0
2004-01-01-16:15:00 4096MB  32%  0.0  3.5  96.5 0.0
```

Afbeelding 3: Uitvoerformaat, vergelijkbaar met vmstat.



Afbeelding 4: Het 'klassieke' dataverwerkingsproces, zoals ook toegepast voor de DWH Monitor.

wanneer deze data feitelijk operationeel gaan worden gebruikt, het 'echte' monitoringproces, dan zou een fysiek gescheiden operational data store ook niet misstaan. Dit klinkt spannender dan het is en houdt feitelijk alleen in dat de data om een bepaalde hoeveelheid minuten worden ververst.

3. Wie mag waar bij? Vanzelfsprekend dient duidelijk te worden bepaald wat generieke informatievoorziening is (rapportages) en specifieke informatievoorziening (query, OLAP, predictie), om de belasting van de infrastructuur niet al te veel te beïnvloeden. Het 'en masse' exploratief onderzoek doen met de

DDL voor dimensies

```
CREATE TABLE dwhmon_dim_systeem (
  systeem_id              INTEGER NOT NULL     ,
  systeem_naam            VARCHAR2 (100)       ,
  systeem_type            VARCHAR2 (50)        ,
  systeem_omschrijving    VARCHAR2 (255)       ,
  systeem_lokatie         VARCHAR2 (255)       ,
  ip_adres1               VARCHAR2 (50)        ,
  ip_adres2               VARCHAR2 (50)        ,
  otap                    VARCHAR2 (50)        ,
  datum_actief            TIMESTAMP            ,
  datum_inactief          TIMESTAMP            ,
  cpu_aantal              INTEGER              ,
  cpu_type                VARCHAR2 (50)        ,
  cpu_model               VARCHAR2 (50)        ,
  geheugen_fysiek         INTEGER              ,
  geheugen_virtueel       INTEGER              ,
      PRIMARY KEY (systeem_id));
```

Toelichting

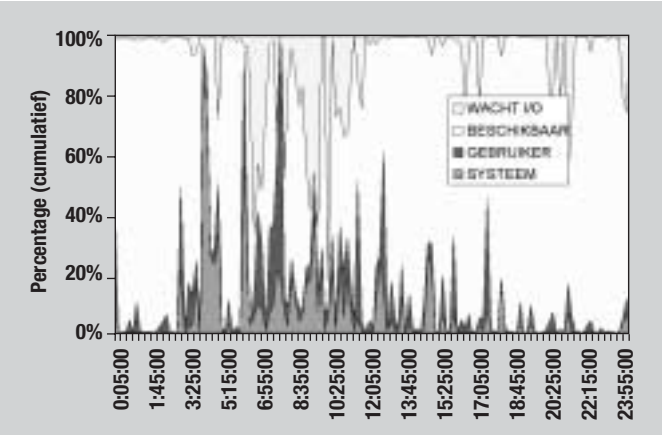
Dimensie SYSTEEM (tabel)
Is bedoeld om:
- een complete 'straat' systemen in te administreren (OTAP = Ontwikkel, Test, Acceptatie en Productie);
- te registreren dat een omgeving 2- of meer-tier is ingericht (de verschillende applicatie-servers op aparte systemen ingericht);
- relevante KPI's, QoS, normen, drempelwaarden en dergelijke te administreren (uit SLA, ITIL);
- alle relevante systeem-configuratie items mee te nemen, zoals soort en aantal CPU's, hoeveelheid intern en extern geheugen et cetera.

FEITEN

```
CREATE TABLE dwhmon_feit_cpu_detail  (
  systeem_id              INTEGER NOT NULL     ,
  tijd_id                 INTEGER NOT NULL     ,
  cpu_nummer              SMALLINT             ,
  cpu_systeem             DECIMAL (3, 1)       ,
  cpu_gebruiker           DECIMAL (3, 1)       ,
  cpu_beschikbaar         DECIMAL (3, 1)       ,
  cpu_wacht_op_io         DECIMAL (3, 1)       ,
      PRIMARY KEY(systeem_id, tijd_id, cpu_nummer));
```

Feit CPU-DETAIL (tabel)
- Resource gebruik per CPU;
- Naast directe CPU-informatie (belasting), zijn er ook andere data toe te voegen, zoals het aantal processen actief op CPU, hoeveelheid intern geheugen door CPU/processen geclaimd et cetera.

Afbeelding 5: Voorbeeld-definitie van dimensies- en feitentabellen voor de DWH Monitor.



Afbeelding 6: DWH Monitor-rapport over het gebruik CPU resources gedurende een etmaal.

DWH Monitor, zal de productiviteit niet ten goede komen. Bovendien is het belangrijker om over een instrument als de DWH Monitor te beschikken dan om te discussiëren over de wijze waarop de onderliggende data gemodelleerd en beschikbaar zijn.

Voorbeeld implementatie

Afbeelding 5 toont als voorbeeld een tabel met de definitie (DDL) van de belangrijkste dimensie- en feitentabellen. Vanzelfsprekend is men vrij om allerlei eigenschappen, meetwaarden toe te voegen of te schrappen.

Afbeelding 6 toont een DWH Monitor-rapport over het gebruik van CPU resources gedurende een etmaal. Dit is een typisch voorbeeld van een rapport dat zonder probleem met een systeemhulpmiddel is te produceren. In de DWH Monitor is het zowel op detailniveau, dat wil zeggen per CPU, te genereren, als op generaal niveau, dat wil zeggen het CPU-totaal.

> Het zal enkele keren per jaar voorkomen dat vraagtekens worden gezet bij de inrichting van de infrastructuur

Afbeelding 7 geeft een voorbeeld van DWH Monitor-periode-rapport voor disk-ruimtegebruik. Het geeft de ontwikkeling weer van het ruimtebeslag van door een externe bron aangeleverde extractiebestanden. De bovenste horizontale lijn geeft het totaal weer van de vrije ruimte en de bezette ruimte. De 90 procent-norm wordt weergegeven door de tweede horizontale lijn. Als de lijn van de bezette ruimte de lijn van de 90 procent-norm passeert dient er direct te worden ingegrepen.

Als laatste wordt kort een 'complexere' analyse beschreven, welke met behulp van een standaard (BI) tool kan worden uitgevoerd. Een handig rapport is een beknopt en globaal dagoverzicht met betrekking tot het gemiddeld gebruik van de de kostbare CPU resources. Om de hoeveelheid data niet al te groot te laten worden, zijn de volgende beperkingen meegenomen. Het betreft (a) een rapportage per gekozen etmaal, (b) er worden gemiddelde waarden per uur getoond van de meetwaarden systeem, gebruiker, beschikbaar en wacht op I/O.

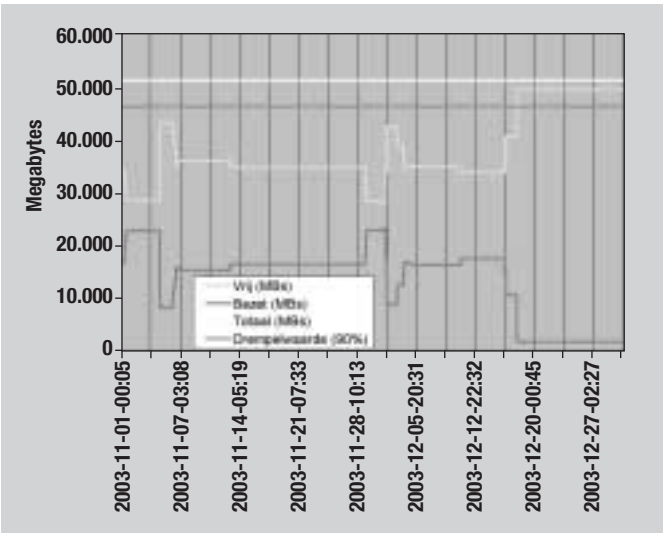
Afbeelding 8 geeft indicatief de knelpunten met betrekking tot het CPU resource-gebruik over 24 uur weer. Dit rapport is prima te combineren met (business) informatie over bijvoorbeeld het aantal ingelogde gebruikers, het aantal processen, de hoeveelheid data (in Kilobytes of records gelezen en geschreven), netwerkverkeer enzovoort. Het zal de geïntereseerde lezer zijn opgevallen dat afbeelding 8 een 'aggregaat' is van afbeelding 6.

Conclusies

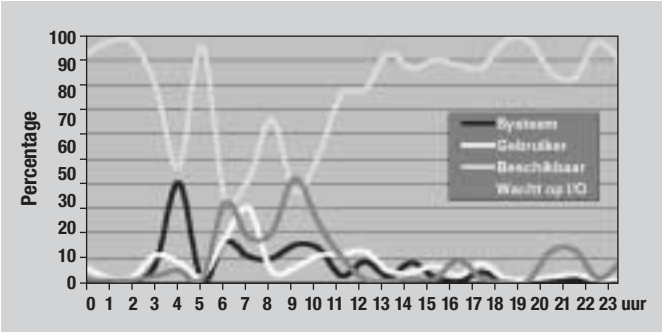
Over het nut en de noodzaak van een DWH Monitor kan men kort of lang discussiëren. Gemiddeld zal het enkele keren per jaar voorkomen dat er de nodige vraagtekens zullen worden gezet bij de beschikbaarheid van resources en de inrichting van de infrastructuur. Dit gegeven alleen al is voldoende om in een paar dagen een eerste versie van een instrument als de DWH Monitor als prototype te implementeren. Hieronder volgen enkele tips voor als men besluit een DWH Monitor op te zetten.

1. 'Think big, act small'.

Zoals bij elk goed DWH-project is hier het 'think big, act small' motto van toepassing. Uitgangspunt is niet dat de hoeveelheid of diversiteit van hetgeen wordt gelogd bepalend is, maar het gebruik. Wanneer een DWH Monitor direct al grootschalig wordt aangepakt, wordt er in feite een compleet DWH-traject gestart, hetgeen voorbij gaat aan het simpele doel van het hebben van een monitoring-instrument voor de beschikbaarheid en gebruik van resources. De moraal van het verhaal is dat er *agile*, iteratief en in belangrijke mate door het DWH beheer dient te worden ontwikkeld en/of begeleid.



Afbeelding 7: Perioderapport voor disk-ruimtegebruik.



Afbeelding 8: Overzicht knelpunten CPU resource-gebruik.

2. Tijddimensie.
Wanneer er op systemen wordt gemeten met behulp van systeemhulpmiddelen, kan het voorkomen dat er niet exact gelijke intervallen worden gehanteerd. Voor verwerkings- en representatiedoeleinden is het wel aan te bevelen om de tijddimensie op basis van vaste meetpunten of -eenheden op te bouwen, bijvoorbeeld elke paar minuten tot heel kwartier. Dit houdt in dat bij verwerking van de metingen, voor zover nodig, metingen naar het dichtst bijzijnde punt op de tijddimensie dienen te worden afgerond.

3. Hergebruik de huidige hulpmiddelen en tools.
De essentie van de DWH Monitor is Business Intelligence voor de beheerder. Dit betekent dat zowel inhoudelijk als technisch, de beheerder zelf ook ' iets' heeft aan een BI- en datawarehousing-omgeving. Het is daarom belangrijk om zoveel mogelijk gebruik te maken van alle voorzieningen, hulpmiddelen en instrumenten die er al zijn.

4. Naar een volwaardig dashboard en BAM.
Degenen die de markt een beetje hebben gevolgd, hebben begrepen dat Gartner's Business Activity Monitoring (BAM) steeds serieuzere vormen begint aan te nemen. Zeker omdat er meer leveranciers BAM-achtige mechanismen ondersteunen. BAM houdt in dat verticaal door de organisatie heen, informatie van verschillende niveaus (business, functioneel IT, technisch IT en infrastructuur) dient te worden gecombineerd om de impact van business-activiteit op IT en vice versa te kunnen monitoren en, indien noodzakelijk, maatregelen te kunnen nemen na geconstateerde incidenten en verstoringen.
Als BAM op een juiste wijze in een organisatie wordt ingevoerd (dat wil zeggen meten in de gehele kolom, monitoren met hoge verversingsfrequentie, informatievoorziening met behulp van dashboards en korte reactietijd tot maatregelen), dan is te verwachten dat geconstateerde incidenten geen 'problem' worden maar tijdig in de kiem worden gesmoord en hun impact minimaal zal zijn.

Dr. Martin P. Misseyer (martin.misseyer@ordina.nl) is Profession Leader bij Ordina VisionWorks.