Migreren datawarehouse naar Exadata

Beschrijving van project binnen Atos Origin

Atos Origin is door een klant gevraagd om de datawarehouse omgeving te migreren. De huidige omgeving bestaat uit 8 Oracle9i standalone databases met een totale omvang van 40 TB. Qua I/O capaciteit, performance en scalability loopt deze datawarehouse omgeving op zijn eind en de gebruikers wachten met smart op een verbetering. Als nieuwe platform is gekozen voor een Oracle11g RAC omgeving op Exadata. In een serie van drie artikelen beschrijven we hoe dit project is aangepakt.

De migratie van het datawarehouse omvat meerdere aspecten waar rekening mee moet worden gehouden. Ten eerste wordt het datawarehouse gemigreerd van Oracle9i naar Oracle 11g release 2. Er zijn diverse nieuwe functionaliteiten geïntroduceerd in versie 11g en bepaalde 9i functionaliteit wordt niet langer ondersteund. Ten tweede wordt er gemigreerd van een Oracle standalone omgeving naar een Oracle RAC omgeving. Dit is een wezenlijk andere architectuur die onder andere consequenties heeft voor de connectivity naar de databases. Ten derde wordt er gemigreerd naar Oracle Exadata. Hier moet met name gekeken worden naar de nieuwe vorm van Exadata Hybrid Columnar Compressie. Maar ook moet nader worden gekeken naar de backup en de monitoring. Ten vierde draait de bronomgeving op Solaris en draait de nieuwe Exadata-omgeving op Oracle Enterprise Linux. Tot slot is hier sprake van een VLDB omgeving met in totaal 40 TB aan data. Dit zal in een window van 48 uur gemigreerd moeten worden, wat strikte eisen stelt aan de migratiemethode en de planning.

Het project wordt daarom opgedeeld in een aantal fases:

- Fase 1: voorbereiding Exadata-machine (setup en configuratie);
- Fase 2: eerste testmigratie waarbij de migratiemethode wordt getest;
- Fase 3: tweede testmigratie waarbij applicatietesten en integratietesten plaatsvinden;
- Fase 4: migratie en inbeheername van het datawarehouse op Exadata.

Parallel aan deze fasering wordt gewerkt aan het inregelen van beheersaspecten als monitoring, backup en disaster recovery.

Inventarisatie van de bronomgeving

De klant beschikt over acht standalone databases (versie 9.2.0.6.0) die draaien op vijf databaseservers. In totaal gaat het om ruim 40 TB aan data, waarbij er twee VLDB databases zijn: één database DWHP van 18 TB en één database DWHKDP van 13 TB. In deze databases zitten een aantal zeer grote gepartitioneerde tabellen. De grootste tabel is bijvoorbeeld 3,4 TB ! Met name de database DWHP heeft veel last van performance problemen die I/O gerelateerd zijn. De twee grootste databases worden via het afsplitsen van mirrors gebackupt en de kleinere databases worden met RMAN naar tape gebackupt. De bronomgeving wordt in kaart gebracht op basis van storage, memory zaken en specifieke init.ora parameters.

Na de inventarisatie van de bronomgeving wordt vastgesteld hoe de databases gaan draaien op de nieuwe Exadata-omgeving. Via een instance mapping document wordt vastgesteld welke databases waar gaan draaien en hoeveel actieve instances elke database heeft en wat de memory instellingen worden voor de instances. De grootste database DWHP krijgt 8 actieve instances op alle nodes om de I/O doorvoersnelheid te maximaliseren. Met behulp van services zal later worden gestuurd dat bepaalde batch processen zich bij deze database op een paar nodes concentreren. Dit om interconnect traffic (dataverkeer tussen de database nodes onderling) te minimaliseren.

Inventarisatie en configuratie doelomgeving

Als doelomgeving geldt een V-2 Exadata Full rack met high performance SAS disken voor productie en een ½ rack met high capacity SATA disken voor de test/uitwijk omgeving. In dit drieluik wordt met name aandacht besteed aan de productie-omgeving. Behoudens de uitwijk is de testomgeving verder niet van belang.

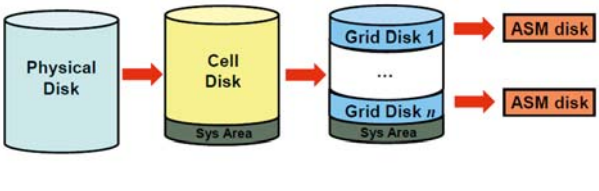
Oracle Exadata

Oracle levert al sinds enige tijd de Oracle Exadata Storage & Database Machine. Een Oracle Exadata omgeving is een door Oracle voorgeconfigureerd 19 inch rack met (bij aanschaf van een 'full' rack) 8 database servers en 14 storage servers. De database en de storage modules vormen dus een geïntegreerd en voorgeconfigureerd geheel. Tussen de componenten is via infiniband een zeer snelle netwerk verbinding geconfigureerd. De storage servers kunnen worden uitgerust met high performance SAS disks (totaal 100 TB bij een full rack) of high capacity SATA disks (totaal 336 TB). Daarnaast beschikken de storage servers over flash disken die kunnen worden ingezet als flash cache. Oracle heeft op diverse niveau's optimalisatie slagen doorgevoerd, waardoor sprake is van een configuratie die vele factoren sneller is dan de traditionele Oracle omgevingen. Zo is er een zeer snel (infiniband netwerk) tussen de storage en de database servers, is een nieuwe vorm van compressie geïntroduceerd (Exadata Hybrid Columnar compressie waarover meer in het tweede artikel) en wordt gebruik gemaakt van een flash cache (totaal 5,3 TB!) en storage indexen. De storage servers bevatten bovendien intelligentie waarbij veel van de queries via een query offload op de storage server zelf worden uitgevoerd en alleen de resultaten worden teruggegeven aan de database server. Deze exadata smart scan functionaliteit is volledig transparant voor de applicatie. De Oracle Exadata machine is schaalbaar (geen downtime nodig) van een kwart machine tot en met 8 Full Exadata machines. Oracle is inmiddels aanbeland bij de 3e generatie Exadata Machines. Na de V-1 en de V-2 generatie worden nu de X2-8 en de X2-2 geleverd aan klanten.

Oracle levert een Exadata-machine op met draaiende (test) database. Alle storage cells zijn geconfigureerd en de grid disken zijn al aangemaakt. Tevens levert Oracle per Exadata-machine een configuratierapport op. Dit vormde het vertrekpunt van het project. In het kader rechtsboven is nader aangegeven hoe de grid disken kunnen worden geconfigureerd, maar dit was dus al door Oracle gedaan bij het opleveren van de Exadata machine.

Disk configuratie met CELLCLI

De full Exadata rack bestaat uit 14 storageservers (ook wel cell servers genoemd). Elke storage server beschikt over 12 fysieke disken van 600 Gb elk alsmede enkele flash disken. M.b.v. de cell commandline interface (cellCLI) kunnen deze disken worden geconfigureerd tot cell disken en van cell disken tot grid disken. (grid disks). De grid disken zijn zichtbaar in +ASM als candidates en kunnen worden toegevoegd aan een +ASM diskgroep:



Onderstaand een voorbeeld hoe dit te werk gaat:

```
 [celladmin@prdcel01 ~]$ cellcli
CellCLI: Release 11.2.1.3.1 - Production on Tue Dec 21 14:05:14 CET
2010

Copyright (c) 2007, 2009, Oracle.  All rights reserved.
Cell Efficiency Ratio: 41,640

CellCLI> create celldisk all harddisk;

Celldisk CD_00_prdcel01 succesfullly created
Celldisk CD_01_prdcel01 succesfully created
. . . . . .

CellCLI> list celldisk;

CD_00_prdcel01          normal
CD_01_prdcel01          normal
. . . . . .

CellCLI> create griddisk all harddisk prefix=data, size=250M;

CellCLI> list griddisk;

DATA1_CD_00_prdcel01            active
DATA1_CD_01_prdcel01            active
. . . .
```

Uiteindelijk was de volgende configuratie qua storage beschikbaar:

```
NAME                              TOTAL_GB   USABLE_GB
------------------------------ ---------- ----------
SYSTEMDG                         4077.50             1851.98
DATA1                           75600.00            33240.56
RECO                            13219.50   5878.21
```

De reden dat de usable storage veel kleiner is dan de total storage is dat ASM is geconfigureerd met redundancy NORMAL. Alle data wordt op +ASM extent niveau dubbel weggeschreven (naar verschillende failure groups) om redundancy in te bouwen. Door deze redundancy is het mogelijk dat een

Oracle Exadata vanuit applicatieperspectief

De Oracle Exadata-omgeving is volledig transparant voor applicaties. De Oracle applicaties communiceren via SQL*Net met de database server. En of daar nu Exadata storage of andere storage aan is gekoppeld maakt niet uit. Wat wel een verandering kan zijn (en ook bij dit project een rol speelt) is dat de Oracle Exadata omgeving een RAC omgeving is, gebaseerd op Oracle11g release 2. Deze versie van Oracle biedt onder andere via SCAN (Single Client Access Name) één entry in de tnsnames.ora waarmee alle nodes in het RAC cluster kunnen worden benaderd. Het is al mogelijk om vanaf Oracle9i gebruik te maken van dit SCAN principe. Maar pas vanaf Oracle client 11g kan van de mogelijkheden als load-balancing en failover gebruik worden gemaakt die met de SCAN functionaliteit worden geboden. Naast deze verandering moet ook de applicatie code 'RAC compliant' zijn. Zo wordt er bij dit project geconstateerd dat veelvuldig gebruik wordt gemaakt van DBMS_PIPE. Deze verouderde Oracle functionaliteit wordt niet door RAC ondersteund. Bij het draaien van applicaties op een Oracle RAC dienen applicaties tevens zo te worden geprogrammeerd dat een reconnect plaats vindt vanuit de applicatie naar de database op het moment dat de verbinding wordt verbroken. Op die manier profiteren de applicaties ten volle van de hoge beschikbaarheid van het totale RAC cluster.

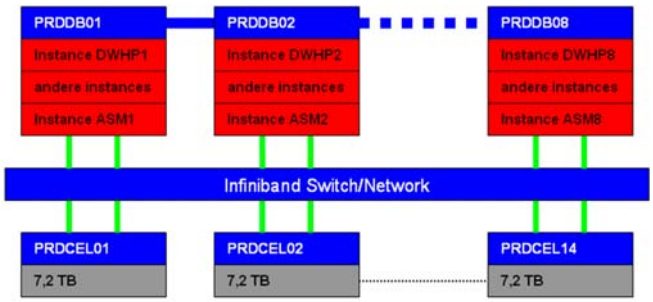
Onderstaand een voorbeeld van een 'SCAN' tns entry

```
DWP =
  (DESCRIPTION=
  (ADDRESS=(PROTOCOL=tcp)(HOST=rac-scan.domein.nl)(PORT=1521))
  (CONNECT_DATA=(SERVICE_NAME=DWP))
  )
```

complete storageserver uitvalt en dat de databases toch gewoon door blijven draaien. Dit is ook handig in het kader van onderhoud. Zo kunnen de storage servers een voor een worden gereboot terwijl de databases dan gewoon blijven draaien. De RECO diskgroup bevat de online redo logfiles en kan daarom klein worden gehouden. Bij deze configuratie is er voor gekozen om de backup en archive locatie uit kostenoogpunt buiten de Exadata machine zelf te plaatsen op Oracle Open Storage. Hierover meer in het derde artikel.

Verder leveren de database servers per node 72 GB aan memory en 8 dual core CPU. Van het memory wordt 10% gereserveerd voor het OS, 20 % wordt gereserveerd voor groei en de resterende 50 GB worden verdeeld over de diverse instances die moeten worden aangemaakt.

Het aanmaken van de databases gebeurt met de Database Configuration Assistant. Deze GUI werkt zeer intuïtief en er wordt voor gekozen om alle 8 instances aan te maken op alle 8 database servers. Na de migratie zullen bepaalde instances van databasesworden gedeactiveerd op basis van het vastgestelde instance mapping plan. Op deze manier wordt de load van het database zo goed mogelijk verdeeld over de gehele Exadata-machine. Alleen database DWHP heeft instances op alle 8 nodes. De andere databases hebben instances op een aantal van de beschikbare nodes. Na afloop is een configuratie als in onderstaande afbeelding aangemaakt.



Nadat de databases zijn aangemaakt worden nog tablespace structuren aangemaakt conform de huidige 9i databases. Dit is ter voorbereiding op de migratie. Omdat van de EHCC compressie veel wordt verwacht worden deze initieel kleiner aangemaakt.

Bij het aanmaken van tablespaces valt direct al de enorme snelheid van de storage op. Daar waar de snelste (niet exadata) omgeving er ruim 15 minuten over doet om een tablespace van 200 GB aan te maken, wordt deze op de nieuwe exadata omgeving aangemaakt in 1:06 minuut. Een tablespace van 1,7 TB is aangemaakt binnentien minuten:

```
create bigfile tablespace ts_dwhp_large
datafile size 1776096M autoextend on next 672M maxsize 8880480M;

Tablespace created.

Elapsed: 00:09:54.40
```

Op deze wijze is de Exadata-omgeving voorbereid op de migratie. In het volgende artikel gaan we nader in op de migratie en de gevolgen op het gebied van compressie en performance.


Rob Lasonder en Jannes Arends zijn als DBA werkzaam bij Atos Origin en maken deel uit van de Global Exadata Competence Group van het bedrijf.