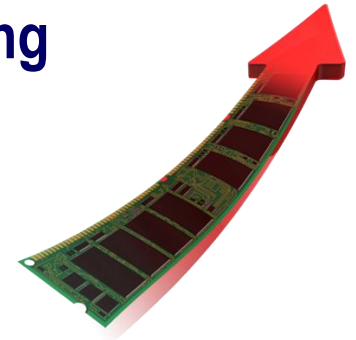


Exadata und In-Memory

Datenbewirtschaftung und Analyse „Extrem“
mit Exadata und InMemory
(Erfahrungsbericht)



Christian Haag, DATA MART Consulting
Consulting Manager
Oracle DWH Team



Inhalt

- **Thematik (Einordnung des Projektbeispiels)**
- **Anforderungen und Ziele**
- **Herausforderungen**
- **Architektur / Komponenten**
- **Facts and Figures**
- **Lösungsbeschreibung / Erfahrungen**
- **Dos and Don'ts**
- **Fazit**

Inhalt

- **Thematik (Einordnung des Projektbeispiels)**
- Anforderungen und Ziele
- Herausforderungen
- Architektur / Komponenten
- Facts and Figures
- Lösungsbeschreibung / Erfahrungen
- Dos and Don'ts
- Fazit

Thematik – Worum geht es?

- **Aufbau eines Data Warehouses**
- **Verarbeitung großer Datenmengen aus dem Online Marketing Bereich**
- **Massive Parallelverarbeitung**
- **Ad-hoc Anfragen auf Tabellen >500 Mio. Datensätze**
- **Auswertung und Analyse von komplexem Kundenverhalten**
 - **Kundensegmentierung**
 - **Trendanalysen**
 - **Produktanalysen**

Inhalt

- Thematik (Einordnung des Projektbeispiels)
- **Anforderungen und Ziele**
- Herausforderungen
- Architektur / Komponenten
- Facts and Figures
- Lösungsbeschreibung / Erfahrungen
- Dos and Don'ts
- Fazit

Anforderungen und Ziele – Wozu?

- **Schnelle Abfragen auf großen Datenbestand**
- **Auswertung vieler Kennzahlen und Attribute**
- **„on-the-fly“ Aggregationen**
- **Unvorhersehbare Abfragen**

Anforderungen und Ziele – Wozu?

- Hohe Flexibilität
- Einfache Wartbarkeit
- Reduzierung von Maintenance Tasks (Index)
- Kurze Aktualisierungszyklen (mehrmals täglich)

Inhalt

- Thematik (Einordnung des Projektbeispiels)
- Anforderungen und Ziele
- **Herausforderungen**
- Architektur / Komponenten
- Facts and Figures
- Lösungsbeschreibung / Erfahrungen
- Dos and Don'ts
- Fazit

Herausforderungen

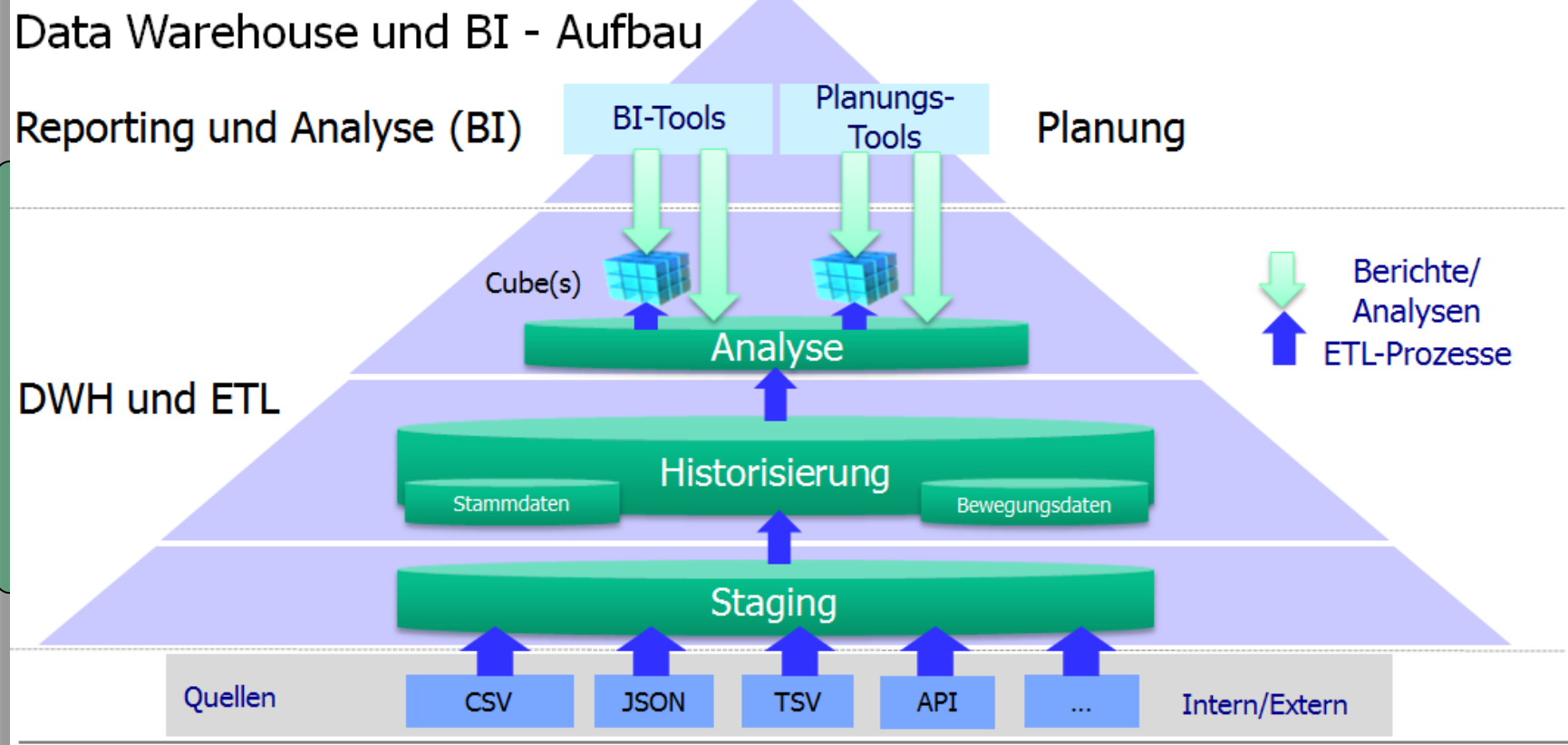
- **Daten mehrmals täglich performant zu verarbeiten**
- **Unabhängige Quellverarbeitung**
- **Hohe Verfügbarkeit der Daten ohne Ausfallzeiten für Beladungszyklen**
- **Ladeprozesse zu standardisieren**
- **Einfaches Datenmodell für hohe Flexibilität**
- **Datenqualität sicherstellen durch standardisierte Prüfverfahren**

Inhalt

- Thematik (Einordnung des Projektbeispiels)
- Anforderungen und Ziele
- Herausforderungen
- **Architektur / Komponenten**
- Facts and Figures
- Lösungsbeschreibung / Erfahrungen
- Dos and Don'ts
- Fazit

Architektur / Komponenten

Erfahrungsbericht - Exadata und In-Memory



Komponenten

- Oracle PL/SQL Packages
- Eigene PL/SQL Packages als Werkzeugkasten
- Linux Skripte (aus DB erzeugt)
- Java



Inhalt

- Thematik (Einordnung des Projektbeispiels)
- Anforderungen und Ziele
- Herausforderungen
- Architektur / Komponenten
- **Facts and Figures**
- Lösungsbeschreibung / Erfahrungen
- Dos and Don'ts
- Fazit

Facts and Figures – Hardware / Software

■ Exadata X3-2 Quarter¹ Rack:



- 2 Datenbank Server, jeder mit:
 - 2 x Eight-Core Intel® Xeon® E5-2690 Processors (2.9 GHz)
 - 256 GB Memory erweitert auf 512 GB
- 48 TB Datenvolumen (nutzbar) + Expansion Rack
- Oracle 11g (11.2.0.3) Database Cluster mit mehreren Instanzen
- Datawarehouse auf 1 Instanz !
 - Ressourcenteilung
 - 32 Cores
 - 60GB SGA

Facts and Figures – Data Warehouse

- Gesamtgröße DWH (3 Jahre):
 - ~2 Mrd Datensätze
 - ~10TB (EHCC for Query High)
 - ~ 150GB In-Memory Tabellen (In-Memory Compression for Query high)
- ~10Mio. Datensätze / Tag
- ~1000 Datendateien / Tag
- 3 Quellsysteme die unabhängig verarbeitet werden
- Online Reporting Systeme die bedient werden (AdHoc Reporting, Report Verteilung)
- Tabellen In-Memory mit bis zu 100 Spalten
- 28 Abfragen gleichzeitig in unter 9s

Inhalt

- Thematik (Einordnung des Projektbeispiels)
- Anforderungen und Ziele
- Herausforderungen
- Architektur / Komponenten
- Facts and Figures
- **Lösungsbeschreibung / Erfahrungen**
- Dos and Don'ts
- Fazit

Lösungsbeschreibung – Wie wurde es gemacht?

Herausforderung	Lösung
Daten mehrmals täglich performant zu Verarbeiten	Parallelisierung, Partitionierung, Komprimierung,
Unabhängige Quellverarbeitung	Logische Parallelisierung durch geplante Jobs
Hohe Verfügbarkeit der Daten ohne Ausfallzeiten	Partition Exchange Loading mit vorheriger Memory Population
Ladeprozesse zu standardisieren	Metadaten Repository mit notwendigen Parametern (Dateiverzeichnisse, FTP Parameter, Prozessketten, etc.)
Datenmodell für hohe Flexibilität	Kein Star-Schema sondern flache, breite Strukturen
Datenqualität sicherstellen durch standardisierte Prüfverfahren	Allgemeine Prüfverfahren (Statistiken, Grenzwerte, Plausibilität)

Lösungsbeschreibung - Erfahrungen

- **Gutes Partitionierungskonzept zwingend notwendig**
- **Möglichst viele Bulk Operationen (CTAS, INSERT APPEND, PARALLEL INSERTS auch wegen Komprimierung)**
- **Viele Joins bei In-Memory vermeiden**
- **Tabellen für In-Memory möglichst flach halten da Column Store Indizes angelegt werden**
- **Partition Exchange Loads (Wenn komplette Partitionen ersetzt werden können) !**

Lösungsbeschreibung - Erfahrungen

- Intelligentes SQL (immer Partitionskriterien verwenden, Hints verwenden da Optimizer nicht immer beste Lösung findet)
- Regelmäßige Datenbankstatistiken fahren
- Gute Partitionskriterien damit Abfragen in kleine Häppchen aufgeteilt werden können
- In-Memory Population muss gut geplant sein, da sonst Daten in der Schlange hinten angestellt werden
- Grundregeln des Data Warehousing gelten auch für Exadata und In-Memory allerdings erlauben diese Mittel neue Architekturen

Inhalt

- Thematik (Einordnung des Projektbeispiels)
- Anforderungen und Ziele
- Herausforderungen
- Architektur / Komponenten
- Facts and Figures
- Lösungsbeschreibung / Erfahrungen
- **Dos and Don'ts**
- Fazit

DOs – „Lessons Learned“

- Sinnvolle Verteilung der Tabellen und Partitionen damit die Cell Server und In-Memory Compression Units gut ausgelastet sind
- Parallele Abfragen erzeugen
- Memory Compression testen (Query High, Low) => CPU overhead kann performance reduzieren
- Aggregationen nutzen Key Vector um Daten zu reduzieren
- Unbedingt Komprimierung verwenden (Performance für Abfragen, Reduktion des Speicherplatzes) => Performance beim Laden nur marginal langsamer
- 500 Millionen Datensätze können in unter 10s ausgewertet werden

Don'ts - „Lessons Learned“

- **Case-When Statements in Where Clauses vermeiden, da zeilenbasiert und damit nicht In-Memory Abfrage**
- **Zu viele Joins Vermeiden da es die Vorteile der In-Memory Option reduziert**
- **Viele Updates und Deletes vermeiden, besser Reload**
- **„Like“ Operationen sind sehr langsam bei In-Memory**

Inhalt

- Thematik (Einordnung des Projektbeispiels)
- Anforderungen und Ziele
- Herausforderungen
- Architektur / Komponenten
- Facts and Figures
- Lösungsbeschreibung / Erfahrungen
- Dos and Don'ts
- **Fazit**

- Exadata und In-Memory ist ein extrem performantes System für Datenbewirtschaftung und Analyse, das:
 - es erlaubt extrem extreme Datenmengen zu verarbeiten und ad-hoc auszuwerten
 - eine Oracle DB darstellt mit ein paar Besonderheiten:
 - IMCU (In Memory Compression Unit)
 - Column Store & In-Memory Storage Indexes
 - SMART Scans
 - Spezielle Architektur ist bei allen Prozessen zu berücksichtigen
 - Beliebig skalierbar ist
 - Intelligentes Storage und Memory Management bietet
 - Kosten sparen kann (im Vergleich zu hohen Kosten in Data Centern für einzelne Komponenten, hier Appliance)
 - Einfach wartbar ist, da alles auf einer Maschine und per Schalter umgelegt wird:

```
„ALTER TABLE <table_name> INMEMORY“
```




Fragen





Christian Haag

Consulting Manager

Tel. +49 6102 7316 0 / Fax +49 6102 7316 22
Martin-Behaim-Str. 12, D-63263 Neu-Isenburg
c.haag@datamart.de

www.datamart.de