Oracle BI EE mit großen Datenmengen

Christian Casek
Riverland Solutions GmbH
München

Schlüsselworte:

Oracle BI EE, Oracle BI Applications, Informatica, RPD, große Datenmengen, Performance, Performanceoptimierung, Laufzeitreduzierung, schnelle Antwortzeiten, ETL, Ladezeit, Datenbank, Datenbank-Tuning, Datenbankoptimierung, Hints, Reports, Auswertungen, Analysen, Dashboards, Answers

Einleitung

Eine essentielle Anforderung für ein Business Intelligence System ist die Performance, welche über die Akzeptanz des Systems bzw. der Software entscheidet. Man stelle sich nur vor, dass ein CEO oder CIO mehrere Minuten oder gar eine viertel Stunde auf Berichte warten muss, bevor sie diese für den eigentlichen Zweck nutzten können. Solch ein Business Intelligence System würde in der heutigen Zeit sehr schnell an Akzeptanz verlieren und als nicht betriebsfähig abgestempelt werden. Doch gerade ein Business Intelligence System dient den Entscheidern um „betriebsfähig“ zu sein, um Ihr Unternehmen mit fundiertem, aktuellem Wissen über das Geschäft steuern zu können.

Um dieser Problematik entgegen zu wirken, ist es maßgeblich, dass bereits bei der Auswahl des BI Systems die Möglichkeiten der Performanceoptimierung und der Skalierbarkeit in Betracht gezogen werden müssen. Dieses besondere Augenmerk muss bei den nächsten Phasen des Projekts, der Anforderungserfassung und der darauffolgenden Erstellung des technischen Designs, mit der gleichen Sorgfalt beibehalten werden.

Dieser Vortrag wird gezielt auf die Schwierigkeiten mit dem Umgang von großen Datenmengen eingehen. Es werden hierbei nicht nur theoretische Problemfälle, sondern Beispiele aus der Realität eines Oracle Business Intelligence Projekts bei einem Transportunternehmen aufgezeigt. Das dort eingesetzte BI System umfasst die Module Marketing, Loyalitäts-, Beschwerde-, und Bestellmanagement, welche zu einem sehr hohem Datenvolumen führen. Die einzelnen Module produzieren jeweils zwischen 100 und bis zu 500 Millionen Datensätzen, die für ein ROLAP System eine beachtliche Datengröße und Herausforderung darstellen. Es müssen spezielle Mechanismen von Oracle BI EE und auch weitere Techniken, die nicht direkt im OBI EE vorhanden sind, angewandt werden, um die Performance bei solchen enormen Datenmengen, für die Endanwender akzeptabel zu halten.

Die Problematik wird im Rahmen dieses Vortrags anhand der drei typischen Schichten eines BI Systems dargestellt. Dabei wird die Systematik der Schichten durchleuchtet und mögliche Ursachen und Lösungsansätze zur Behebung von Laufzeitproblemen aufgezeigt.

Es werden sowohl technische Details zu den Schichten, als auch relevante Themen, die im Rahmen eines BI Projekts in Bezug auf Performance eine wichtige Rolle spielen und in Betracht gezogen werden müssen. Des Weiteren wird die Skalierbarkeit des OBI EE Systems in Relation zur Datenmenge und Performance dargestellt.

Bei den drei Schichten handelt es sich um:

- Die Datawarehouse – Schicht (physikalische Datenbankschicht)
- Die logische Schicht des Business Intelligence Systems
- Die Request- und Dashboards – Schicht (Frontends)

Um die Probleme und Lösungsansätze besser zu veranschaulichen, werden reale Beispiele aus Projekten vorgestellt, deren Systemlandschaften mit Oracle Datenbanksystemen, Oracle BI Applications und Oracle BI EE bestehen.

Die Datawarehouse – Schicht (physikalische Datenbankschicht)

Die Basis eines BI Systems ist ein Datawarehouse. Es beinhaltet die relevanten Daten für Analysen, Reporting und Marketing. Die Daten werden aus unterschiedlichen operativen Systemen des Unternehmens extrahiert und in einer konsolidierten Form ins Datawarehouse geladen. Damit eine höchstmögliche Aktualität der Daten gewährleistet werden kann, muss der ETL (Extract – Transform – Load) Prozess täglich durchgeführt werden. Bei großen Datenmengen führt dies schon zum ersten Problem, da die Zeitfenster für den ETL Prozess meinst kurz bemessen sind.

Man stelle sich folgendes Szenario vor: Das CRM System wird von den Vertriebsmitarbeitern eines Unternehmens täglich bis 20:00 Uhr, im Anschluss daran laufen Verbuchungsprozeduren im Hintergrund der Applikation ab, die erst gegen 23:30 Uhr abgeschlossen sind. Somit kann der ETL frühestens um 24:00 Uhr beginnen, wenn man einen Puffer von einer halben Stunde berücksichtigt. Soll das Datawarehouse jeden Tag ab 08:00 Uhr mit aktuellen Werten zur Verfügung stehen, so bedeutet dies ein Zeitfenster von 8 Stunden in denen der komplette ETL abgeschlossen sein soll. Das ist nicht allzu viel Zeit, um beispielsweise mehrere Millionen Datensätze zu extrahieren und im Datawarehouse in einer konsolidierten Form zu laden.

Um die begrenzt verfügbare Zeit optimal zu nutzen, sollten die Entitäten, die maßgeblich für die großen Datenmengen verantwortlich sind, sowohl fachlich als auch technisch betrachtet werden. Aus fachlicher Seite sollten vor allem der notwendige Aktualität und der Umfang der tatsächlich benötigten Information hinterfragt werden. Dadurch könnte beispielsweise die Anzahl der zu extrahierenden Spalten deutlich reduziert werden, was einen Laufzeitvorteil mit sich bringen würde. Ist aus fachlicher Seite keine Reduktion mehr möglich, so muss die Optimierung auf der technischen Seite fortgeführt werden. Hier sollten die Funktionen der zugrundeliegenden Datenbanksysteme genutzt werden und in die SQL Abfragen des ETL Prozesses dahingehend optimiert werden.

Die logische Schicht des Business Intelligence Systems

Die logische Schicht des BI Systems verbindet das physikalische mit dem logischen Datenmodell, das die fachlichen Anforderungen abbildet. Im Oracle BI wird das vom Oracle BI Server über das Metadaten Repository (RPD) ermöglicht. Hier können für eine logische Dimension mehrere physikalische Tabellen hinterlegt werden, die unterschiedlicher Granularität sein können. Dadurch können Aggregatstabellen eingebunden werden, die zum Beispiel Umsatzdaten auf Monats- und Jahresebene beinhalten. Das führt dazu, dass bei einer Abfrage durch den Endanwender die Tabelle mit dem zur Abfrage passenden Aggregatslevel ausgelesen wird, was eine Verkürzung der Laufzeit, da nicht jeder einzelne Datensatz zur Laufzeit aggregiert werden muss.

Die Request- und Dashboards – Schicht (Frontends)

Die Answers und Dashboards bilden das Frontend der Oracle BI EE. Hier werden vorgefertigte Reports zur Verfügung gestellt oder Themenbereiche mit denen die Endanwender selbständig Auswertungen erstellen können. Der Benutzer erwartet, dass er zeitnah Antworten erhält und nicht minutenlange Wartezeiten in Kauf nehmen muss. Ein BI System, welches ein solches Verhalten ausweist, ist heutzutage nicht akzeptabel und muss optimiert werden.
Es gibt mehrere technische Ansätze mit denen eine Verbesserung erzielt werden kann, jedoch sollte hier ebenfalls primär im fachlichen Bereich mit einer Analyse der tatsächlichen Anforderungen begonnen werden.
Anhand von konkreten Beispielen wird gezeigt werden, wie auf die Endanwender zugegangen werden kann, um ihnen das Wissen über das BI System, dessen Verhalten und Möglichkeit näher zu bringen und zu verdeutlichen. Dadurch können hilfreiche Erkenntnisse gewonnen werden, wie Dashboards und Themenbereich besser für die Anwender entwickelt werden können.

Theorie ohne Praxis?

Dieser Vortrag wird konkrete Methoden vorstellen, wie in der Realität mit dem Problem von langen Laufzeiten in den Schichten eines BI Systems umgegangen werden kann. Hierbei werden Ansätze aufgezeigt, wie z.B. die Kommunikation mit den Fachabteilung aussehen kann und wie hier bereits im Vorfeld mögliche Probleme auf technischer Seite verhindert werden können.
Des Weiteren werden im Rahmen dieser Veranstaltung unterschiedliche technische Funktionen und Designpattern vorgestellt, mit denen eine Laufzeitoptimierung erzielt werden kann. Diese werden sich über alle drei Schichten eines BI Systems erstrecken und mit realen Beispielen untermauert werden.

Das Ziel dieser Veranstaltung ist es nicht nur theoretische Möglichkeiten vorzustellen, wie etwas umgesetzt oder gelöst werden soll, es soll hier vielmehr gezeigt werden, wie es auf Projekten zu realen Problemen kam, was deren Auswirkungen waren und wie diese auch in der Realität erfolgreich gelöst wurden und für zukünftige Projekte als „Lessons learned“ fungieren kann. Den Zuhörer soll bewusst gemacht werden, dass bei der Konzeption eines BI System die Problematik von großen Datenmengen ein wichtiger Aspekt ist und bei der fachlichen und technischen Architektur nicht außer Acht gelassen werden sollten.

Kontaktadresse:

Christian Casek
Riverland Solutions GmbH
Holbeinstr. 22
D-81679 München

Telefon: +49 89 41 073 860
Fax: +49 89 41 073 862
E-Mail christian.casek@riverland.com
Internet: www.riverland.com