Oracle-Statistiken im Data Warehouse effizient nutzen

Reinhard Mense
ARETO Consulting
Köln

Schlüsselworte:

DWH, Data Warehouse, Statistiken, Optimizer, Performance, Laufzeiten

Einleitung

Für die performante Ausführung von Berichten und Ad-hoc-Abfragen eines BI-Systems sind aussagekräftige und aktuelle Statistiken für den Oracle-Optimizer von essentieller Bedeutung. Doch die Erstellung der Statistiken benötigt bei großen Datenmengen oft auch sehr viel Zeit. Somit stellt sich die Frage nach einer geeigneten Strategie für die Aktualisierung der Statistiken eines Data Warehouse (DWH).

Für die Ausführung eines SQL-Statements untersucht der Oracle Optimizer alle möglichen Ausführungspläne und wählt den Ausführungsplan mit den geringsten Kosten. Damit der Optimizer die Kosten für einen Ausführungsplan ermitteln kann, benötigt er möglichst detaillierte Informationen über alle beteiligten Tabellen und Indizes. Diese Informationen werden durch Statistiken zur Verfügung gestellt. Anhand der Statistiken versucht der Optimizer unter anderem vorherzusagen, wie viele Datensätze in einem Schritt des Ausführungsplans, also z. B. bei einem Join, verarbeitet werden müssen. Diese Anzahl Datensätze, Cardinality genannt, bestimmt maßgeblich die Kosten und damit die Auswahl eines Ausführungsplans. D. h. aber auch der ausgewählte Ausführungsplan kann nur gut sein, wenn die Statistiken dem Optimizer die richtigen Informationen geliefert haben. Mit jeder Ausführung des ETL-Prozesses werden umfangreiche Änderungen bzw. neue Daten in das DWH aufgenommen, so dass insbesondere hier auf aktuelle Statistiken großer Wert gelegt werden muss.

Welche Statistiken sollen erstellt im DWH werden?

Die typische DWH-Architektur besteht aus einer Staging Area, einem Core und Data Marts. Während die Staging Area zur temporären Aufnahme der aus den Quellsystemen extrahierten Daten und für die Speicherung von Zwischenergebnissen der Transformationen im ETL-Prozess verwendet wird, werden im Core dauerhaft die konsolidierten und homogenisierten Daten historisch gespeichert. Die Data Marts stellen schließlich die Daten in für die Abfragen optimierte Star- oder Snowflake-Schemata zur Verfügung.

Da die Berichte und Ad-hocAbfragen in der Regel nur auf die dafür optimierten Data Marts ausgeführt werden, sind für die Datenbankobjekte dieser Schicht aktuelle Statistiken besonders wichtig, um eine gute Performance der Abfragen zu erzielen. Dabei sollten die Statistiken sowohl für alle Tabellen und Indizes der Data Marts als auch für die häufig zusätzlich erstellen Materialized Views erzeugt werden.

Auch wenn auf der Staging Area und der Core-Schicht keine Auswertungen erfolgen, ist zu bedenken, dass im Rahmen der ETL-Prozesse intensive Abfragen auf die Tabellen und Indizes dieser Schichten

erfolgen. Insbesondere bei einem ELT-Tool, wie z. B. dem OWB, werden die ETL-Prozesse vollständig in der Datenbank ausgeführt, so dass auch hier aktuelle Statistiken von großer Bedeutung sind, um eine gute Performance der ETL-Prozesse zu ermöglichen.

Lokale und globale Statistiken

Im Core und in den Data Marts werden die Bewegungsdaten und die Fakttabellen in der Regel partitioniert. Statistiken können sowohl für die einzelnen Partitionen (lokale Statistiken) also auch für die gesamte Tabelle (globale Statistiken) erstellt werden.

Greift eine Abfrage nur auf eine Partition zu, werden vom Optimizer lediglich die lokalen Statistiken der einzelnen Partition ausgewertet. Erfolgt jedoch der Zugriff auf mehr als eine Partition werden sowohl die lokalen Statistiken als auch die globalen Statistiken vom Optimizer ausgewertet. Abfragen, die die Daten mehrerer Partitionen auslesen, sind im DWH keine Seltenheit, so dass in jedem Fall neben den lokalen auch die globalen Statistiken aktuell zu halten sind.

Histogramme

Werden die Daten z. B. durch eine WHERE-Bedingung gefiltert, versucht der Optimizer vorherzusagen, wie viele Datensätze gefiltert werden. Ohne Histogramme geht der Optimizer von einer Gleichverteilung der Werte aus. D. h.:

Anzahl Datensätze = Anzahl unterschiedlicher Werte für die Filterspalte / Gesamtzahl der Datensätze

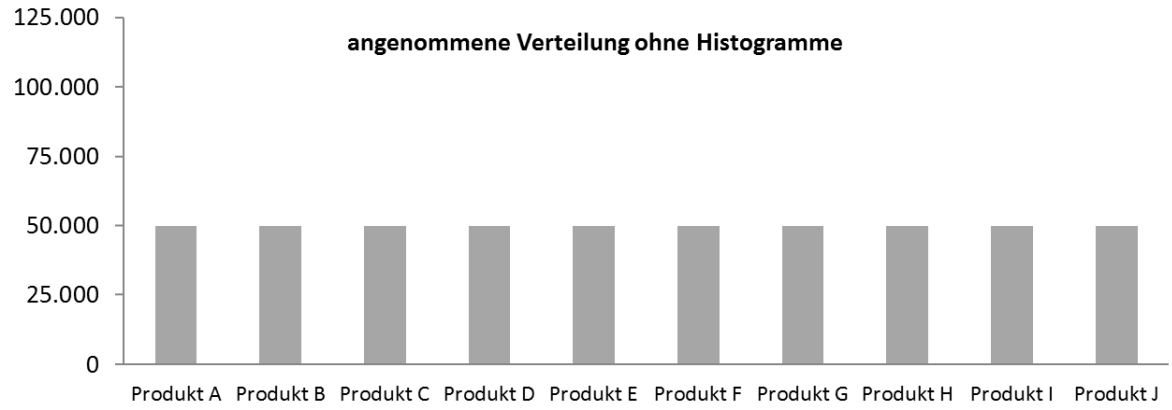

Abb. 1: Vom Optimizer angenommene Verteilung von z. B. Verkaufsdaten ohne Histogramme

In der Realität liegt jedoch nicht immer eine Gleichverteilung vor, so dass der Optimizer die falsche Anzahl Datensätze ermittelt und damit auch die Gefahr besteht, dass ein ungünstiger Ausführungsplan gewählt wird. Für Spalten, die von der Gleichverteilung der Daten deutlich abweichen, kann es dehalb sinnvoll sein, Histogramme zu erzeugen. Histogramme ermöglichen dem Optimizer auch für die Spalten mit nicht gleichverteilten Werten eine genauere Vorhersage der zu erwartenden Datensätze.

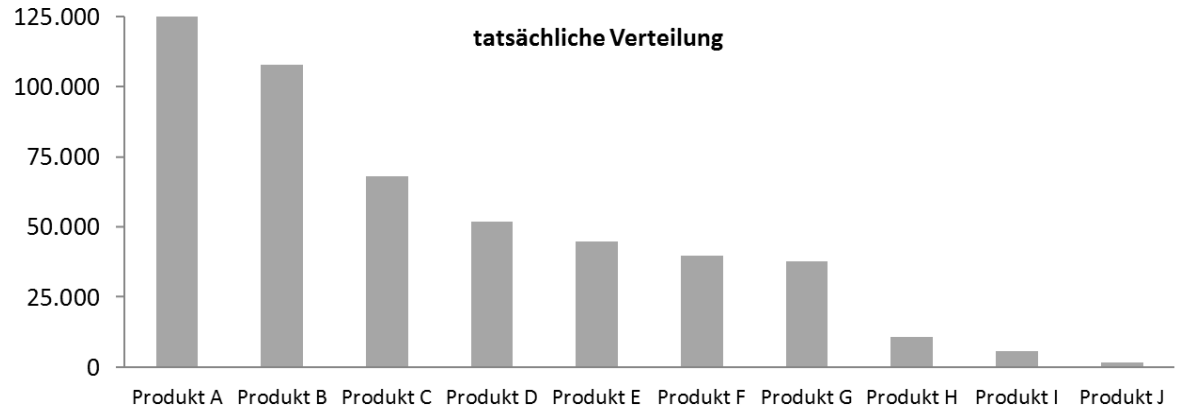

Abb. 2: Tatsächliche Verteilung der Verkaufsdaten

Im DWH wird in der Regel über die Attribute der Dimensionstabellen gefiltert, so dass Histogramme für diese Tabellen in Betracht zu ziehen sind. Dabei sollten Histogramme nur für Spalten mit einer ungleichen Verteilung erzeugt werden. Durch Joins der Dimensionstabellen mit den Fakttabellen findet implizit auch eine Filterung der Fakttabellen statt. Deshalb sind Histogramme auch für die Foreign Key-Spalten der Fakttabellen von Bedeutung.

Extended Statistics

Erfolgt eine Filterung der Daten über mehr als eine Spalte, nimmt der Optimizer an, dass die Werte dieser beiden Spalten unabhängig voneinander sind. Enthält z. B. eine Produkt-Dimension im DWH 10.000 Produkte, die hierarchisch in 100 Produktgruppen und 10 Produktkategorien gleichmäßig verteilt sind und wird das folgende SQL-Statement abgesetzt

```
SELECT *
  FROM dim_produkt
 WHERE produktgruppe_id = 1
   AND produktkategorie_id = 1
```

so erwartet der Optimizer 10.000 / 100 / 10 = 10 Datensätze als Ergebnis. Da die Werte für Produktgruppen und Produktkategorien aber aufgrund ihrer hierarchischen Abhängigkeit korrelieren, werden tatsächlich 10.000 / 100 = 100 Datensätze als Ergebnis geliefert.

Abfragen dieser Art treten im DWH häufig auf und stellen insbesondere bei Dimensionstabellen für den Optimizer ein Problem dar, da die Korrelation der Spalten nicht erkannt wird. Extended Statistics können hier Abhilfe schaffen. Extended Statistics erlauben es für die kombinierten Werte mehrerer Spalten Statistiken zu erstellen. Diese Statistiken erlauben es dann dem Optimizer die korrekte Anzahl der Datensätze auch bei korrelierenden Spalten zu ermitteln.

Für Standards Berichte kann die gezielte Erzeugung von Extended Statistics Performance-Probleme aufgrund von Filtern über korrelierende Attribute beseitigen. Bei Ad hoc-Abfragen stößt der Einsatz der Extended Statistics jedoch an seine Grenzen, da nicht immer vorhersehbar ist welche korrelierenden Attribute die Benutzer in ihren Abfragen in welcher Kombination verwenden. Extended Statistics für sämtliche denkbare Kombinationen von korrelierenden Attributen zu erzeugen ist viel zu aufwendig, so dass eine andere Lösung her muss.

Dynamic Sampling

Wenn keine Extended Statistics vorliegen kann der Einsatz von Dynamic Sampling dem Optimizer helfen, die richtige Cardinality vorherzusagen. Beim Dynamic Sampling werden vom Optimizer für das auszuführende SQL-Statement zusätzliche Statistiken für die beteiligten Objekte ermittelt. Der Datenbankparameter OPTIMIZER_DYNAMIC_SAMPLING gibt dabei an, wann Dynamic Sampling zum Einsatz kommt und welche zusätzlichen Informationen durch Dynamic Sampling ermittelt werden sollen. Wird dieser Parameter auf das Level 6 oder höher gesetzt, werden bei Filtern über mehrere Spalten Informationen über Korrelationen dieser Spalten gesammelt. Je höher das Level ist, desto mehr zusätzliche Informationen werden gesammelt. Werden viele Informationen gesammelt benötigt der Optimizer zwar mehr Zeit für das Ermitteln des Ausführungsplans, aber er kann ggf. den besseren Ausführungsplan wählen, was insbesondere im DWH bei Abfragen auf große Datenmengen ein Vorteil sein kann.

Wie sollen die Statistiken erstellt werden?

Aufgrund der großen Datenmengen im DWH ist eine möglichst effiziente Erzeugung der Statistiken von großer Bedeutung. Die Dauer für die Erzeugung der Statistiken hängt wesentlich vom Umfang der für die Analyse verwendeten Stichprobe ab. Für partitionierte Tabellen, wie z. B. den besonders großen Fakttabellen, können die globalen Statistiken seit Oracle 11g außerdem inkrementell erzeugt werden.

AUTO_SAMPLE_SIZE

Der Umfang der für die Erzeugung der Statistiken zu analysierenden Daten kann mit Hilfe des ESTIMATE_PERCENT-Parameters beim Aufruf der DBMS_STAT.GATHER_TABLE_STATS-Prozedur als Prozentsatz (Sample Rate) angegeben werden. Ein niedriger Prozentwert führt zu einem geringen Umfang der Stichprobe für die Analyse. Je geringer der Umfang der Stichprobe desto schneller erfolgt die Erzeugung der Statistiken, gleichzeitig nimmt man aber auch eine geringere Genauigkeit der Statistiken in Kauf.

Als Default-Wert für ESTIMATE_PERCENT ist jedoch nicht ein fester Prozentwert angegeben, sondern der Wert DBMS_STATS.AUTO_SAMPLE_SIZE. Dieser Wert bewirkt, dass die Oracle-Datenbank selbst den geeigneten Prozentwert für das Erzeugen der Statistiken ermittelt. Wird AUTO_SAMPLE_SIZE in Oracle 11g verwendet, kommt dabei außerdem ein neuer sehr effizienter Algorithmus für das Erzeugen der Statistiken zum Einsatz. Die von diesem Algorithmus erzeugten Statistiken haben fast die gleiche Genauigkeit wie die Statistiken auf Basis einer 100% Sample Rate.

Die nachfolgende Abbildung zeigt die Laufzeiten einer Oracle 11g-Datenbank für das Erzeugen der Statistiken für eine 40 Millionen Datensätze umfassende Fakttabelle mit unterschiedliche Sample Rates im Vergleich zur Nutzung von AUTO_SAMPLE_SIZE.

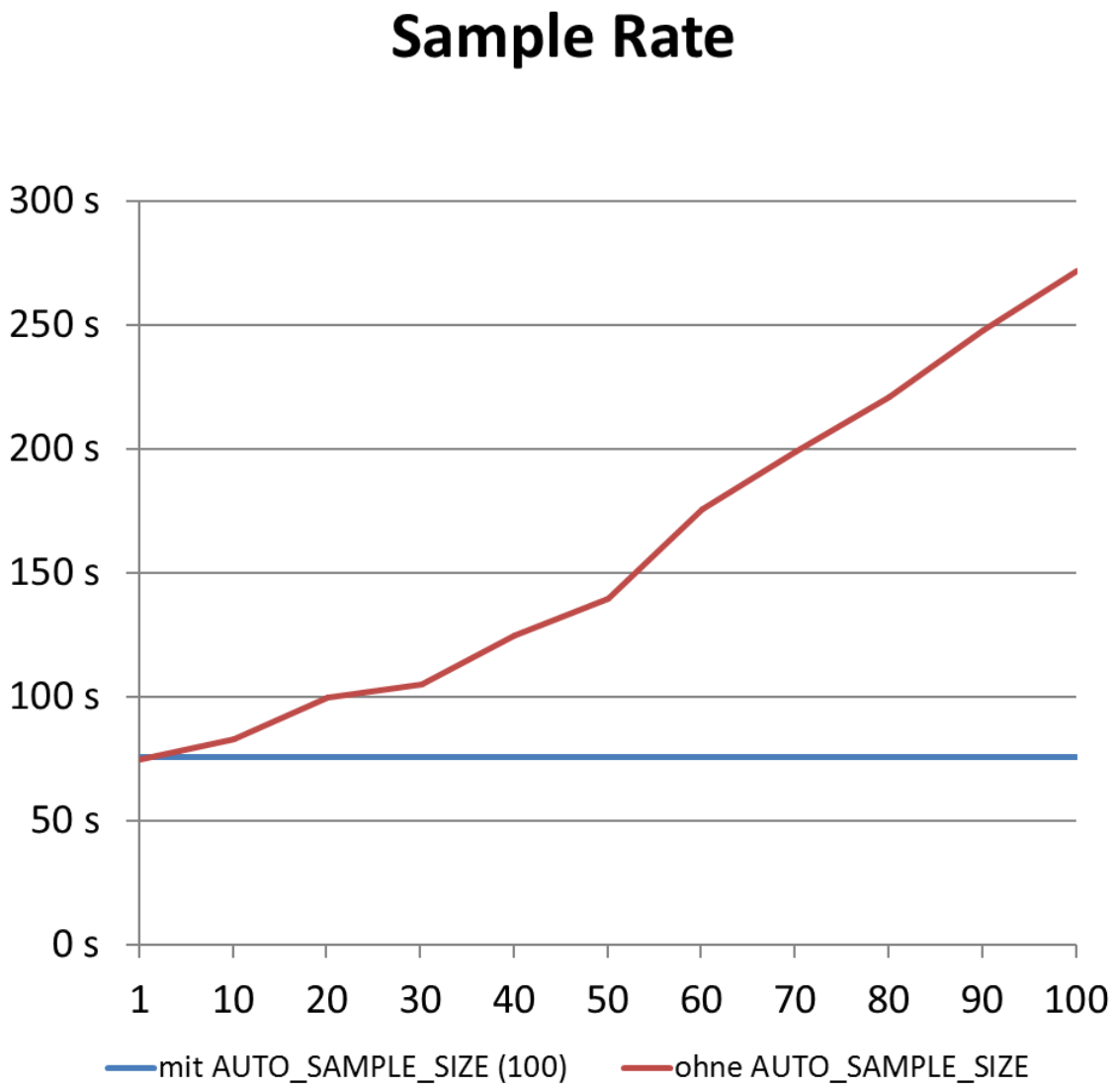

Abb. 3: Laufzeiten für Statistiken mit unterschiedlichen Samples Rates im Vergleich zu AUTO_SAMPLE_SIZE

Inkrementell Statistiken

Das Erstellen der globalen Statistiken ist insbesondere für große Fakttabellen sehr aufwendig und benötigt deshalb oft viel Zeit. Da Fakttabellen jedoch in der Regel partitioniert sind, und meist nur die Daten einer Partition oder weniger Partitionen sich ändern, sollte man den Einsatz inkrementeller Statistiken in Betracht ziehen. Inkrementelle Statistiken können seit Oracle 11g verwendet werden. Dazu müssen die lokalen Statistiken der einzelnen Partitionen aktuell sein und die inkrementellen Statistiken für die betroffenen Tabellen mit DBMS_STATS.SET_TABLE_PREFS aktiviert werden (Preference INCREMENTAL auf TRUE setzen).

Durch das Aktivieren der inkrementellen Statistiken speichert die Oracle-Datenbank für jede Partition der Tabelle ein sogenanntes Synopsis-Objekt im SYSAUX-Tablespace. Die Synopsis-Objekte beinhalten statistische Metdaten für die einzelnen Partitionen und Spalten in den Partitionen. Im Vergleich zu den vollständigen Partitionsdaten sind die Synopsis-Objekte sehr klein. Wird eine neue

Partition hinzugefügt oder eine bestehende Partition geändert, müssen zunächst für diese Partition die lokalen Statistiken aktualisiert werden. Anschließend kann die Oracle-Datenbank anhand der Synopsis-Daten die globalen Statistiken erzeugen ohne die gesamte Tabelle lesen zu müssen. Dadurch wird die Laufzeit für die Erzeugung der globalen Statistiken erheblich reduziert.

Wann sollen die Statistiken erstellt werden?

Die Oracle-Datenbank bietet die Möglichkeit an, die Anzahl geänderter Datensätze zu protokollieren (MONITORING-Klausel für Tabellen) und mit Hilfe eines ein einmal täglich laufenden Jobs die Statistiken der Tabellen automatisch zu aktualisieren. Dabei werden nur Tabellen aktualisiert, bei denen sich mindestens ein festgelegter Prozentsatz der Datensätze geändert hat. Die Statistiken werden dann als Stale bezeichnet. Per Default ist der Prozentsatz auf 10% eingestellt. Dieses Verfahren eignet sich für OLTP-Systeme häufig gut, da diese Systeme in der Regel nicht umfangreiche historische Daten vorhalten und somit eine relativ geringe Menge geänderter Daten ausreicht, um das Aktualisieren der Statistiken auszulösen.

Beim DWH hingegen ist mit zunehmender historischer Datenmenge eine immer größere Anzahl Datensätze erforderlich, um die erforderlichen 10% Änderungen zu erreichen und somit das Aktualisieren der Statistiken auszulösen. Im DWH kann das insbesondere bei den Fakttabellen dazu führen, dass über einen größeren Zeitraum das Aktualisieren der Statistiken ausbleibt. Die unten stehende Grafik verdeutlicht anhand eines Beispiels das Problem.

Dabei wird davon ausgegangen, dass in einer Tabelle (z. B. einer Fakttabelle) jeden Tag die gleiche Anzahl neuer Datensätze hinzugefügt wird. In den ersten 10 Tagen werden die Statistiken jeden Tag als Stale angesehen und entsprechend erneuert, aber bereits am 11. Tag wird die erforderliche 10%-Grenze nicht mehr erreicht und das Aktualisieren der Statistiken somit nicht ausgeführt. Erst am 12. Tag wird die 10%-Grenze wieder überschritten und die Statistiken werden erneuert. Zu beachten ist, dass die Zeiträume, in denen die Statistiken nicht erneuert werden, mit zunehmenden historischen Datenvolumen immer größer werden. Nach drei Monaten werden bereits 8 Tage lang die Statistiken der Tabelle nicht aktualisiert. Und am Ende des Jahres überschreitet der Zeitraum ohne aktuelle Statistiken mit 32 Tagen bereits einen ganzen Monat.

Insbesondere für Berichte und Abfragen, die auch auf die neuen Daten zugreifen, drohen somit aufgrund von fehlenden Statistiken und unpassende Ausführungsplänen erhöhte Laufzeiten. Für DWH-Systeme ist dieses Verhalten nicht akzeptabel. Deshalb sollte das Erzeugen der Statistiken regelmäßig innerhalb des ETL-Prozesses erfolgen. Damit wird sichergestellt, dass die Statistiken stets aktuell sind, und die Performance der Abfragen sich nicht verschlechtert.

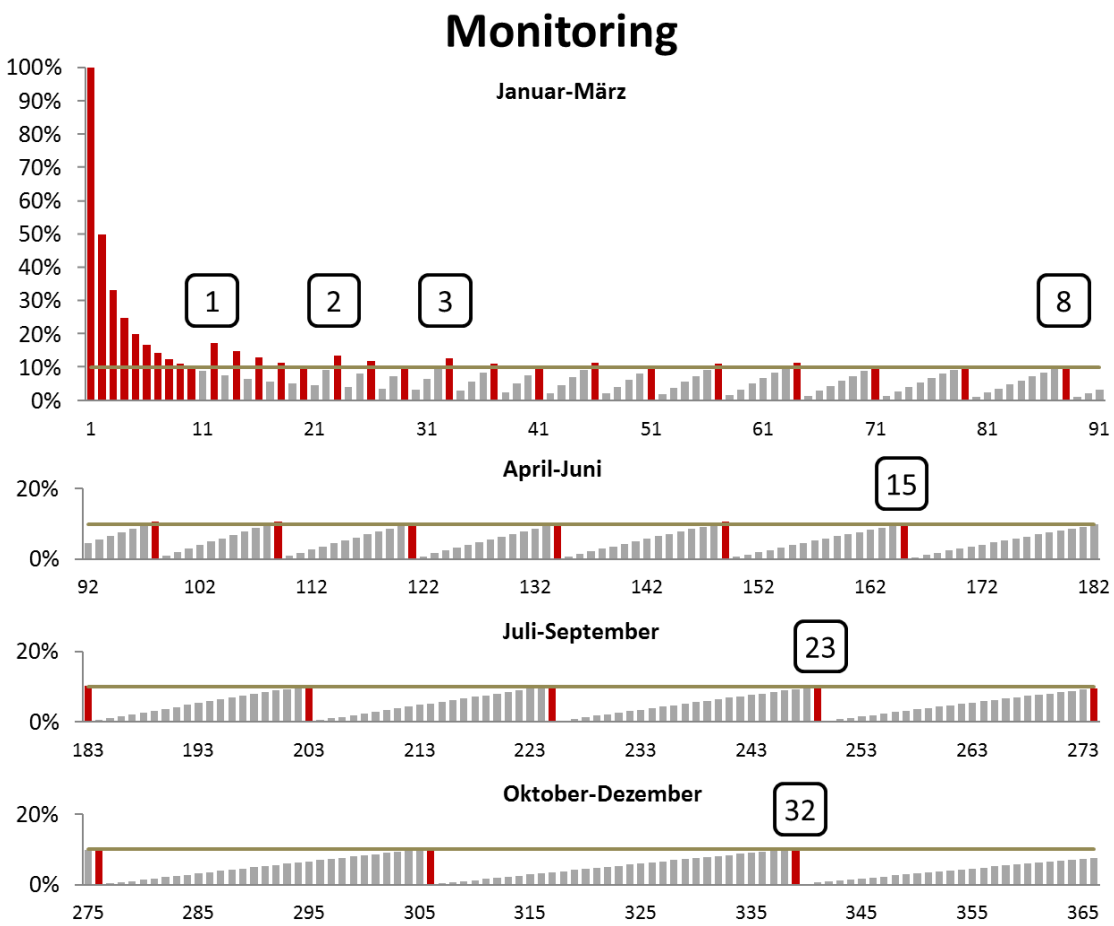

Abb. 4: Automatische Erzeugung der Statistiken mit GATHER_DATABASE_STATS_JOB_PROC

Fazit

Aktuelle Statistiken sind im DWH unverzichtbar, um performante Berichte und Abfragen zu garantieren. Mit den Möglichkeiten von Oracle 11g und der richtigen Strategie ist es möglich, auch für große Datenmengen eines DWH den Aufwand für das Erzeugen aktueller Statistiken gering zu halten. Dabei sollten folgende Punkte beachtet werden:

1. Globale Statistiken für partitionierte Fakttabellen inkrementell erzeugen.
2. Histogramme für ungleichmäßig verteilte Daten verwenden.
3. Extended Statistics für korrelierende Spalten erstellen, die in Standardberichten als Filter genutzt werden
4. Dynamic Sampling mit Level 6 oder höher einsetzen, um auch für Ad hoc-Abfragen gute Laufzeiten bei Filtern über korrelierende Spalten zu erzielen.
5. Statistiken in der Regel mit AUTO_SAMPLE_SIZE erzeugen. Nur bei sehr großen Fakttabellen ggf. gezielt kleine Werte für ESTIMATE_PERCENT angeben.
6. Statistiken im DWH nicht mit dem automatischen Statistik-Job der Datenbank erzeugen.

Kontaktadresse:

Reinhard Mense
ARETO Consulting GmbH
Julius-Bau-Str. 2
D-51063 Köln

Telefon: +49 (0) 221-66 95 75 0
Fax: +49 (0) 221-66 95 75 99
E-Mail reinhard.mense@areto-consulting.de
Internet: www.areto-consulting.de