SQL Server Statistiken

Die Rolle von Statistiken bei der Erstellung von Ausführungsplänen

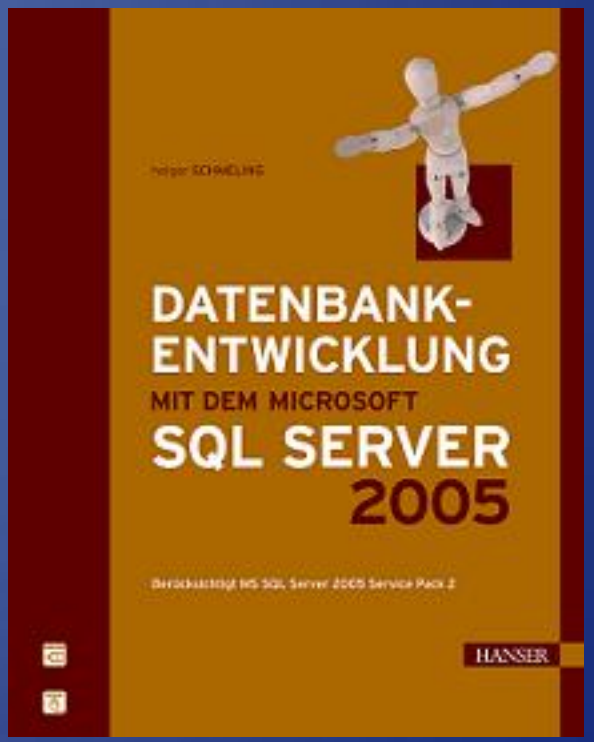
DBCC SHOW_PROFILE('Holger Schmeling')

- SQL Server seit 1995 (Version 6.5)
- Freiberuflicher Consultant seit 1996
  - Datenbank-Architektur, -Design, -Administration, - Entwicklung, -Optimierung
  - Trainer und Autor
  - München
  - http://www.sqlserver-online.com

DBCC SHOW_PROFILE('Holger Schmeling')

SQL Server 2008 Performanceoptimierung
Addison-Wesley, 2009
ISBN10: 3827327784

Datenbank-Entwicklung mit dem Microsoft SQL Server 2005
Hanser Verlag, 2007
ISBN10: 3446225323



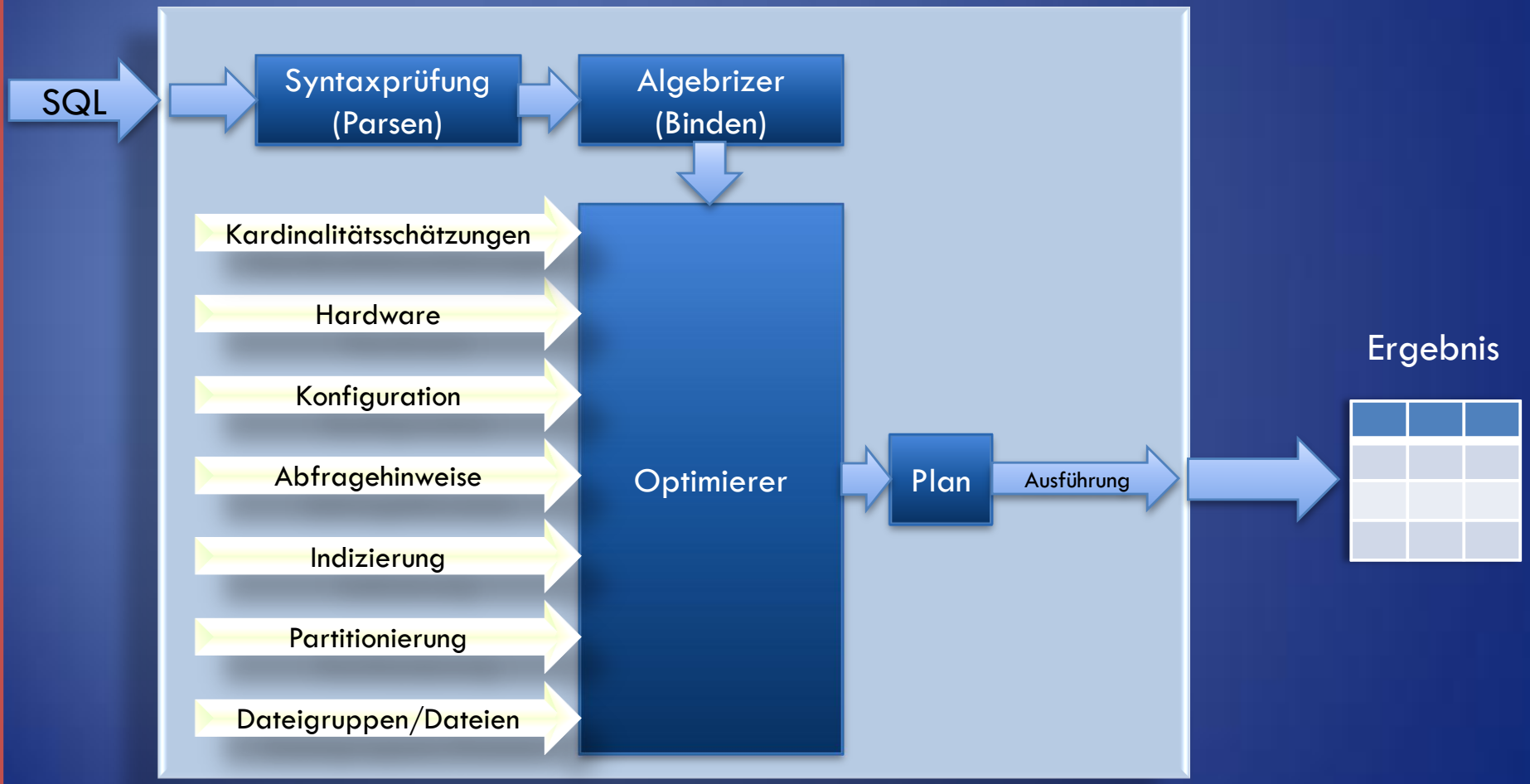
Agenda

- Wozu Statistiken?
- Organisation von Statistiken
- Verwalten von Statistiken
- Probleme mit Statistiken
- Zusammenfassung/Best Practices

Agenda

- Wozu Statistiken?
- Organisation von Statistiken
- Statistiken verwalten
- Probleme mit Statistiken
- Zusammenfassung/Best Practices

Wozu Statistiken?



Wozu Statistiken?

- Kardinalitätsschätzung
  - Auswahl geeigneter physischer Operatoren zur Ausführung einer logischen Operation
  - Zum Beispiel: Suche (Seek ⇔Scan), JOIN (Merge, Hash, Nested Loop)
  - Für Abfragen allozierter Speicher
- Stichprobe der Originaldaten
- Reduzierung der Datenmenge für die Kardinalitätsschätzung
- Beschleunigung der Optimierungsphase

Wozu Statistiken?

- Demo01-Cardinality

Agenda

- Wozu Statistiken?
- Organisation von Statistiken
- Verwalten von Statistiken
- Probleme mit Statistiken
- Zusammenfassung/Best Practices

Organisation von Statistiken

- Vor- und Nachteile
- Arten von Statistiken
- Informationen abfragen
- Beispiel

Organisation von Statistiken

- Allgemeine Probleme mit redundanten Daten
  - Anomalien
  - Asynchronität
- Allgemeines Problem mit allen Statistiken
  - Informationsreduktion
  - Relevanz / Charakteristik der Stichprobe

Statistiken sind wie Bikinis. Sie zeigen Verlockendes und verbergen doch das Wesentliche.

Aaron Levenstein

Organisation von Statistiken

- Generell zwei Arten von Statistiken
  - Index-Statistiken
    - Für jeden Index automatisch erstellt
  - Spalten-Statistiken
    - Bei Bedarf automatisch erstellen lassen oder manuell erstellen
- Einspaltige und mehrspaltige Statistiken

Organisation von Statistiken

- Informationen abfragen
  - SQL Server Management Studio
    - Problem mit der Darstellung von „reinen“ Spaltenstatistiken (ohne Bezug zu einem Index)
  - DBCC SHOW_STATISTICS
  - TSQL
    - sys.stats
    - sys.stats_columns
    - stats_date()
    - sp_helpstats (veraltet!)
    - sys.sysindexes (veraltet!)
      - rowcnt, rowmodctr

Organisation von Statistiken

- Beispiel

- Demo02-CreateTestTable

Organisation von Statistiken

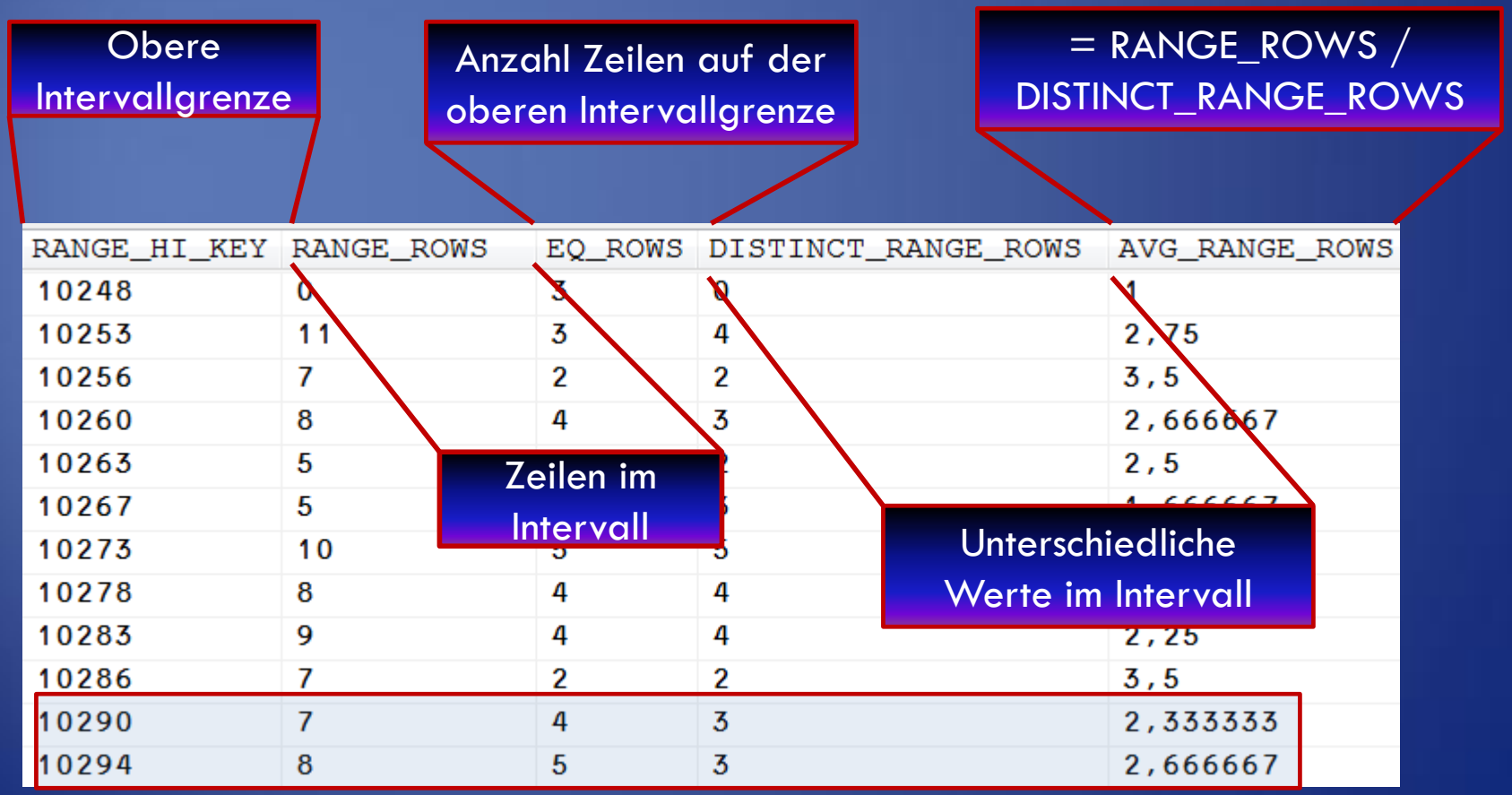
- Informationen - Allgemeine Information

Name	Updated	Rows	Rows Sampled	Steps
IxT1_x	Feb 6 2010 12:24PM	100000	100000	200

  - Zeitpunkt der letzten Aktualisierung
  - Anzahl der Tabellenzeilen
  - Anzahl der für die Stichprobe entnommenen Zeilen
  - Anzahl der Einträge im Histogramm (maximal 200)

Organisation von Statistiken

- Informationen-Histogramm
  - Nur für die führende Spalte



RANGE_HI_KEY	RANGE_ROWS	EQ_ROWS	DISTINCT_RANGE_ROWS	AVG_RANGE_ROWS
10248	0	3	0	1
10253	11	3	4	2,75
10256	7	2	2	3,5
10260	8	4	3	2,666667
10263	5	2	2	2,5
10267	5	3	3	1,666667
10273	10	5	5	2
10278	8	4	4	2
10283	9	4	4	2,25
10286	7	2	2	3,5
10290	7	4	3	2,333333
10294	8	5	3	2,666667

Organisation von Statistiken

- Informationen – Gesamt-Dichte

All density	Average Length	Columns
0.001	4	x

  - Berechnet durch *1/(unterschiedliche Anzahl entnommener Werte)*
  - Vom Optimierer verwendet, wenn genauere Abschätzung über das Histogramm nicht möglich ist.

Organisation von Statistiken

- Informationen abfragen
  - TSQL & DBCC

- Demo03-ObtainingInfo

Organisation von Statistiken

- Demo04-AutoCreate

Agenda

- Wozu Statistiken?
- Organisation von Statistiken
- **Verwalten von Statistiken**
- Probleme mit Statistiken
- Zusammenfassung/Best Practices

Verwalten von Statistiken

- Erzeugen von Statistiken
  - Automatisch
  - Manuell
  - Neu in SQL Server 2008: Gefilterte Statistiken
- Aktualisieren von Statistiken
  - Automatisch
  - Manuell

Verwalten von Statistiken

- Automatische Erzeugung von Statistiken
  - Option auf Datenbankebene
  - Eingeschaltet, wenn *model* Datenbank nicht modifiziert
  - Fehlende Spalten-Statistiken werden bei Bedarf ergänzt
    - Einspaltig und ungefiltert
  - **Immer** bei CREATE INDEX oder ALTER INDEX … REBUILD
  - Nicht bei ALTER INDEX … REORGANIZE
  - Kann im Profiler beobachtet werden: *Performance/Auto Stats*

Verwalten von Statistiken

- Automatische Erzeugung von Statistiken
  - DEMO
    - SSMS und ALTER DATABASE

Verwalten von Statistiken

- Manuelle Erzeugung von Statistiken
  - Wie?
    - CREATE STATISTICS
    - sp_createstats
  - Warum sollte man?
    - Erzeugung von Statistiken benötigt Zeit.
    - Proaktives Vorgehen, also Erzeugung im Wartungszeitfenster
    - Automatische Erstellung kann nicht alles
      - Gefilterte Statistiken
      - Anzahl der für die Stichprobe entnommenen Zeilen
      - Mehrspaltige Statistiken
  - Best Practice
    - AUTO_CREATE_STATISTICS ON
    - Manuelle Erzeugung ausgewählter (bekannter) Statistiken
- Demo05-CreateStatistics

Verwalten von Statistiken

- Gefilterte Statistiken
  - Neu in SQL Server 2008 (wie auch gefilterte Indizes)
  - Werden nicht automatisch erzeugt. Ausnahme: Index-Statistik für gefilterten Index
  - Wozu gut?
    - „Virtuell“ mehr als 200 Einträge im Histogramm
    - Maßgeschneiderte Statistiken
    - Partitionierte Tabellen bzw. Indizes
      - Beispiel: 90% historische Daten, 10% aktuelle (aktive) Daten
    - Eventuell niedrigerer Wartungsaufwand
      - Beispiel: Nur 10% der Tabellenstatistiken müssen aktualisiert werden
      - Problem mit Aktualisierung (siehe unten)

Verwalten von Statistiken

- Demo06-FilteredStats

Verwalten von Statistiken

- Aktualisierung von Statistiken
  - Automatisch
    - Synchron
    - Asynchron
  - Manuell
    - UPDATE STATISTICS
    - sp_updatestats
- Hat Re-Compilierungen zur Folge!

- Demo SSMS und TSQL

Verwalten von Statistiken

- Automatische Aktualisierung von Statistiken
  - Statistiken enthalten redundante Daten.
  - Originaldaten und Statistiken sind in der Regel nicht zu 100% synchron.
  - Wann ist eine Statistik nicht mehr aktuell?
    - Anzahl Zeilen > 500
      - (20% + 500) Änderungen
    - Anzahl Zeilen <=500
      - 500 Änderungen
    - Änderung der Zeilenanzahl von 0 auf einen Wert >0
    - Überwachung der Änderungen für Statistik-Spalten: *colmodctr*
    - Temporäre Tabellen (#): Alle 6 Änderungen

Verwalten von Statistiken

- Automatische Aktualisierung von Statistiken
  - Erst bei Bedarf, nicht etwa bereits nach Datenänderungen!
  - Ausschließen einzelner Tabellen von automatischer Aktualisierung
    - Wie?
      - UPDATE/CREATE STATISTICS … WITH NO_RECOMPUTE
      - CREATE INDEX … WITH (STATISTICS_NORECOMPUTE)
      - sp_autostats
    - Warum?
      - Manuelle Aktualisierung für ausgewählte Tabellen, z.B. weil gefilterte Statistiken manuell erstellt wurden.
      - Automatische Aktualisierung zu häufig, z.B. weil sich die Zeilenanzahl oft von 0 auf >0 ändert

Verwalten von Statistiken

- Manuelle Aktualisierung von Statistiken
  - Warum?
    - Automatische Aktualisierung ist nicht perfekt.
    - Statistiken, die von automatischer Aktualisierung ausgeschlossen wurden.
    - 20% Schwelle für Invalidierung oftmals zu hoch (Beispiel folgt)
    - Automatische Aktualisierung online, während Normalbetrieb evtl. nicht optimal.
    - Größe der Stichprobe kann angegeben werden.
    - Erforderlich für gefilterte Statistiken
      - Datenänderungen in Bezug auf die Filterbedingung
      - Gefilterte Statistiken altern schneller (Beispiel folgt).

Verwalten von Statistiken

- Manuelle Aktualisierung von Statistiken
  - Wie?
    - UPDATE STATISTICS
    - sp_updatestats
      - Testet auf Erreichen der Invalidierungs-Schwelle

- Demo07-UpdateStatistics

Agenda

- Wozu Statistiken?
- Organisation von Statistiken
- Verwalten von Statistiken
- Probleme mit Statistiken
- Zusammenfassung/Best Practices

Probleme mit Statistiken

- Es gibt keine Statistik
- Aktuelle Statistik existiert, wird aber nicht verwendet
- Statistik ist zu ungenau
- Veraltete Statistik
- Mehrspaltige Statistiken
- Keine Unterstützung von Statistiken für abhängige Spaltenwerte
- Erzeugung/Aktualisierung benötigt Ressourcen
- Unangemessene Speicheranforderung für Abfrageausführung

Probleme mit Statistiken

- Es gibt keine Statistik
  - AUTO_CREATE_STATISTICS ist OFF und keine Statistik manuell erzeugt
  - Tabellen Variable
  - XML und Geo-Daten
  - Remote Abfragen
    - OPENROWSET/OPENQUERY
    - Eine Reihe von DMVs
      - sys.dm_tran_current_transaction
  - Datenbank ist Read/Only
    - Datenbank Snapshots!
      - Demo08-NoStatisticsForSnapshot

Probleme mit Statistiken

- Aktuelle Statistik existiert, wird aber nicht korrekt verwendet
  - Ursache ist meist ineffizientes TSQL
    - Lokale Variablen in TSQL Skripten
    - Ausdrücke in Prädikaten (Non-foldable expressions)
      - Besser: Berechnete Spalten
  - Probleme mit Parametrisierung
    - Erster Parametersatz entscheidend für Planerstellung
    - Parameterwerte in gespeicherten Prozeduren nicht verändern
    - OPTION RECOMPILE nach Bedarf

- Demo09-UnusedStats

Probleme mit Statistiken

- Statistik ist zu ungenau
  - Entnommene Stichprobe ist zu gering
    - Abhilfe durch manuelle Erstellung/Aktualisierung mit Angabe der Stichprobengröße (z.B. FULLSCAN)
  - Granularität ist zu groß
    - 200 Einträge im Histogramm können für große Tabellen zu wenig sein.
      - Beispiel: 5.000.000 Zeilen bei 200 Histogramm-Schritten ergibt 25.000 Zeilen pro Histogramm-Wert.
      - Kein Problem, wenn Spaltenwerte gleich verteilt sind, sonst evtl. schon
    - Abhilfe durch gefilterte Statistiken/Indizes.
      - Vorsicht mit automatisch erzeugten Spaltenstatistiken.

Probleme mit Statistiken

- Veraltete Statistik
  - Erinnerung: Mindestens 20% Änderungen erforderlich, damit automatische Aktualisierung erfolgt.
  - In vielen Fällen ist diese Schwelle zu groß und manuelle Aktualisierungen sind erforderlich.
  - Achtung bei gefilterten Indizes/Statistiken!
    - Keine Berücksichtigung des Prädikats für Filterung

- Demo10-Product

Probleme mit Statistiken

- Mehrspaltige Statistiken
  - Werden nicht automatisch erzeugt
  - Manuell anlegen
  - Schwer zu ermitteln
    - Erfahrung erforderlich
    - DTA kann bei Analyse helfen

Probleme mit Statistiken

- Keine Unterstützung von Statistiken für abhängige Spaltenwerte
  - Werte in verschiedenen Salten, die voneinander abhängig sind
    - Beispiele
      - Geschlecht ⇔ Größe
      - Alter ⇔ Schuhgröße
  - By Design
  - Abhilfe durch
    - Gefilterte Indizes/Statistiken
    - Abdeckende Indizes

- Demo11-RentalCar

Probleme mit Statistiken

- Erzeugung/Aktualisierung benötigt Ressourcen
  - Entnehmen der Stichprobe
  - Nicht vergessen: Invalidierung gespeicherter Pläne!
  - Automatisch (online) oder manuell (im Wartungsfenster)?

Probleme mit Statistiken

- Unangemessene Speicheranforderung für Abfrageausführung
  - Für Abfrageausführung erforderlicher Speicher wird anhand von Zeilengröße, Zeilenanzahl und Operator bestimmt
  - Verschätzt sich der Optimierer, wird zu viel oder zu wenig Speicher angefordert
    - Zu viel: Verschwendung von Hauptspeicher
    - Zu wenig: Auslagerung auf *tempdb*
      - Ca. Faktor 5-10 mal langsamer!
      - Profiler: *Errors and Warnings/Sort Warnings*

- Demo12-MemAlloc

Agenda

- Wozu Statistiken?
- Organisation von Statistiken
- Verwalten von Statistiken
- Probleme mit Statistiken
- Zusammenfassung/Best Practices

Zusammenfassung/Best Practices

- Bequem sein ist ok! Lass SQL Server die Arbeit erledigen und erlaube automatische Erzeugung sowie Aktualisierung von Statistiken.
- Bei Problemen mit Abfrageleistung => Statistiken für beteiligte Tabellen aktualisieren (evtl. mit FULLSCAN)
- Automatik ja, aber nicht ausschließlich darauf vertrauen!
- Fragmentierte Indizes neu erstellen. Dadurch werden auch Statistiken neu erzeugt (mit FULLSCAN). Achtung! Kein UPDATE STATISTICS nach Index Rebuild!
- Wenn mehrspaltige Statistiken sinnvoll sind => manuell erzeugen. Evtl. DTA zur Analyse heranziehen.
- Gefilterte Statistiken sind nützlich, **müssen** aber manuell aktualisiert werden.
- Nicht mehr als eine Statistik für eine Spalte.
- TSQL Code (keine lokalen Variablen, kein Überschreiben von Parameterwerten in gespeicherten Prozeduren, keine Ausdrücke für Prädikate)

Interessante Links

- Zitate
  - http://www.stubig.com/Wissenschaft/Zitate.html
- Statistics Used by the Query Optimizer in Microsoft SQL Server 2005
  - http://technet.microsoft.com/en-us/library/cc966419.aspx
- Statistics Used by the Query Optimizer in Microsoft SQL Server 2008
  - http://msdn.microsoft.com/en-us/library/dd535534.aspx
- Using Filtered Statistics with Partitioned Tables
  - http://sqlcat.com/msdnmirror/archive/2009/10/20/using-filtered-statistics-with-partitioned-tables.aspx