

SYBASE®

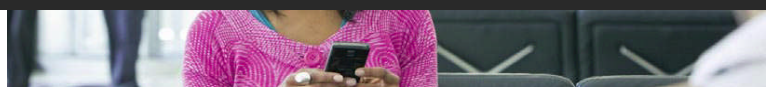
Sybase IQ: Verblüffend einfach. (Vol. 03)

Die gesamte Klaviatur: Anwendung von Indizes in Sybase IQ

SYBASE IQ: VERBLÜFFEND EINFACH. TECHNISCHE TIPS & TRICKS



www.sybase.com



Inhalt

Der Nutzen der gesamten Klaviatur - Anwendung von Indizes in Sybase IQ	1
Stimmen der Oktaven... Allgemeines zu Sybase IQ Indexstrukturen	2
Akkorde: Indextypen in Sybase IQ	2
Basis: Der Fast Projection (FP) Index	3
Der Low Fast Index (LF)	6
Der High Group Index (HG)	6
Der High Non Group Index (HNG)	6
Der Compare Index (CMP)	7
Der Containment Index (WD)	7
Der Index auf Datum und Zeit (DTTM DT TM)	7
Join Index	7
Einsatz von Indizes in Sybase IQ	8
Index Advisor	8
Anlegen von Indizes nach Eindeutigkeit und Kardinalitätsanalyse	8

Principal Author

Klaus Riemer klaus.riemer@sybase.com

Revision History

Version 1.0 - März 2010

Der Nutzen der gesamten Klaviatur - Anwendung von Indizes in Sybase IQ

In dieser Ausgabe der „Sybase IQ – verblüffend einfach“- Reihe möchten wir Ihnen die Anwendung von Indexsystemen in Sybase IQ - welche sich wesentlich von den üblichen Verfahren zeilenorientierter Datenbanken unterscheidet - vorstellen. Wir werden Ihnen aufzeigen, welchen Nutzen Sie aus diesen Indexsystemen ziehen können, welche konkreten Einsatzmöglichkeiten es gibt und welche Regeln Sie beachten sollten.

Was Sie wissen sollten und wovon wir ausgehen:

Worauf wir in diesem Dokument nicht eingehen werden, d.h. was wir voraussetzen ist, dass Ihnen die gängigen Indexierungssysteme B-Tree oder Bitmap ein Begriff und die Aufgaben von Indizes bekannt sind - Indizes werden bspw. gezielt eingesetzt, um Schlüsselbeziehungen zwischen Eltern – Kind Tabellen zu unterstützen und um Abfragen zu tunen. Ferner weisen wir darauf hin, dass bei konventionellen Datenbanksystemen oftmals lediglich ein Index per Abfrage zum Einsatz kommt.

Hinweis:

Da Indizes mit zum Teil erheblichem Platzverbrauch einhergehen und stets gepflegt werden müssen, sind - damit sich die Ladeperformance nicht verschlechtert – bei konventionellen Datenbanksystemen vor dem Laden in der Regel Indizes vorab zu löschen. Bei Sybase IQ ist dies nicht der Fall. Hier werden die Indizes vor dem Einfügen neuer Daten weder gelöscht noch neu angelegt, sondern beim Laden automatisch gepflegt. Reorganisationsmaßnahmen sind nicht erforderlich.

Stimmen der Oktaven .. Allgemeines zu Sybase IQ's Indexstrukturen

Bei spaltenorientierten Datenbanken sind Zeilen nur „virtuell“ und der Verbund eines logischen Datensatzes wird erst zur Ausführungszeit ermittelt. Ferner werden nur die Spalten verwendet, die in der jeweiligen Abfrage im Ergebnis oder in den Filterbedingungen genannt wurden - nicht verwendeten Datensatzelemente werden von Sybase IQ nicht angetastet. Gleichzeitig sind die Datenseiten „dicht“ gepackt und redundanzfrei – d.h. statistische Verteilungen etc. finden keine Anwendung und die sonst erforderlichen Pflegeaufwände entfallen.

Da „Tabellen“ in Sybase IQ eine Sammlung ausgewählter Spalten sind und sich nicht, wie normalerweise üblich, alle Felder eines Datensatzes in einer Seite befinden, empfehlen wir Ihnen die Datenspalten sofort und nur in einer Indexstruktur darzustellen. Das hat zum Vorteil, dass der Aufwand eine Basistabelle zu erstellen entfällt.

Ein weiterer Vorteil, der sich daraus ergibt, ist der, dass für jede Spalte, die in einer Abfrage genutzt wird, sofort eine Index-Anwendung gefunden wird. D.h. für jede Spalte einer Abfrage wird ein Index verwendet, nicht ein einzelner Index für die gesamte Abfrage.

Wenn bereits alles „automatisch“ indiziert ist, wozu dann noch zusätzliche Indizes anlegen? Diese und andere Fragen werden in diesem Kapitel beantwortet.

Akkorde: Indextypen in Sybase IQ

Befassen wir uns zunächst mit der allgemeinen Indexstrukturen in Sybase IQ. Indizes in Sybase IQ werden als Bitmap Indizes verschiedener Breite, B-Tree Index, Kombination, Lookup-Systemreferenzen und spezielle, dem Datenformat entsprechende, Datenstrukturen implementiert. Entscheidend für die Nutzung eines Indextyps sind die Kardinalität der voneinander verschiedenen (d.h. disjunkten) Werte der jeweiligen Spalte, die Nutzung (Referenzierung) des Werts in einem Datensatz, des Datentyps (int, date, Word, ..) oder der Nutzung (Schlüsselspalte, Anzeige, Filter, Aggregation, Gruppierung..).

Aus all diesen Kriterien kann dann die Auswahl der von Sybase IQ angebotenen Indextypen getroffen werden, die üblicherweise nur unter den Abkürzungen FP, FP(1), FP(2), FP(3), LF, HG, NHG, CMP, WD, DTTM, DT, TM und JOIN bekannt sind.

Bevor nun die Indextypen vorgestellt werden, sei noch daran erinnert, optimale Performance kann nicht nur mit den Indizes erreicht werden, sondern ist immer ein Zusammenspiel aus mehreren Faktoren: Datenmodell, Datentypen, technische Ressourcen und letztendlich die Abfrage selber.

Die Art der gestellten Abfragen ist vorhersehbar, jedoch sind Datenmodell, Datentypen und die Aufteilung der Ressourcen neben den Indizes als beeinflussende Faktoren „konstant“ und werden vor den Indizes definiert. Daher kann die optimale Leistung nur dann erzielt werden, wenn alle konstanten Faktoren optimiert wurden.

Basis: Der Fast Projection (FP) Index

Dieser Index wird automatisch für jede Spalte erstellt, und ist für Sybase IQ eine Art Basisspalte und somit im weitesten Sinne mit der Basistabelle einer zeilenorientierten Datenbank zu vergleichen. Dieser Index wird bereits bei Definition der Tabelle vorbereitet und beim Laden / Einfügen von Daten automatisch angelegt und angepasst und kann somit vom Benutzer nicht gelöscht werden.

Der FP-Index dient zur „Projektion“ d.h. zur Anzeige von Daten, wird jedoch auch bei Summation oder anderen Funktionen, die die Darstellung des Ergebnisses beeinflussen, verwendet. Der FP – Index wird bei Werkseinstellung als „flat“ Index erzeugt, d.h. die Bytebreite des Index entspricht dem Datentyp der Spalte. In Fällen hoher Kardinalität, (Schlüssel, Katalognummern etc.) d.h. im Falle vieler unterschiedlicher Einträge, ist dies eine gute Wahl. Allerdings finden sich bei der Analyse der Daten sehr oft Spalten mit Informationen, die für viele Datensätze zutreffend sind. Typische Fälle sind z.B. Geschlecht (männlich, weiblich, Firma, Unbekannt), aber auch Warengruppen oder Filialbezeichnungen. In diesen Fällen ist der „flat“ FP Index ineffizient bezüglich der Auswirkungen auf Platzverbrauch und Leistung. Deshalb bietet es sich an die FP(1) 1 Byte für bis zu 256 Ausprägungen, FP (2) - 2 Byte für bis zu 1024 Ausprägungen oder - neu in Sybase IQ 15 - FP(3) 3 Byte für bis zu 65535 verschiedene Werte einer Spalte zu verwenden.

Wie erreicht man nun, dass Sybase IQ den jeweils kleinstmöglichen FP –Index verwendet? Die empfohlene und für Administration und Anwender bequemste Variante ist es:

die Option “Minimize Storage“ mit

```
set public.option MINIMIZE_STORAGE='ON'
```

zu setzen. Dadurch wird automatisch, und ohne weiteres Zutun, für alle danach angelegten Tabellen der kleinstmögliche Index angelegt.

Weitere Möglichkeiten bieten sich als Spaltenoption beim Anlegen der Tabelle oder für den Fall, dass sich bereits Daten in der Tabelle befinden über die Prozedur

```
sp_iqrebuildindex.
```

Überschreiten die von einander verschiedenen Werte die Indexkapazität, wird automatisch auf den nächst breiteren Index erweitert. Dies ist jeweils genau einmal der Fall. Die Erweiterung findet dann während des Ladevorgangs oder beim Insert statt und macht sich ggf. durch eine Verzögerung im Ladevorgang bemerkbar, jedoch ist dies für den abfragenden Benutzer transparent.

Weitere Maßnahmen zum Index gibt es nicht, sollten die Kardinalitäten jedoch dauerhaft wieder unter die Indexschwelle der aktuellen Variante sinken, empfiehlt es sich den Index mit der Prozedur `sp_iqrebuildindex` wieder zu verkleinern.

Betrachten wir nun die häufig angewendeten und empfohlenen zusätzlichen Indextypen:

Der Low Fast Index (LF)

Der Low Fast Index LF kommt zum Einsatz im Fall von Spalten mit geringer Kardinalität, d.h. weniger als 1000 von einander verschiedene Werte und mindestens 25.000 Datensätze in der Tabelle und ist ideal zur Anwendung in Filtern (Where – Bedingungen oder für Join Prädikate), für Aggregate / Zährefunktionen. Der Index kann nicht angewendet werden, wenn die Anzahl verschiedener Werte 10.000 übersteigt oder im Falle von Spalten mit Bit Datentypen oder Zeichenketten (Char / VarChar) mit mehr als 255 Byte Breite.

Der LF Index sollte nicht bei kleinen Tabellen mit weniger als 25.000 Datensätzen angewendet werden, denn hier steigen die I/O Operationen an. Stattdessen sollte der High Group Index angewendet werden, den wir Ihnen nachfolgend vorstellen.

Der High Group Index (HG)

Dieser Index ist der neben dem LF Index am häufigsten eingesetzte Indextyp. Er eignet sich insbesondere für Filter, Join-Spalten mit Integerdatentypen und für Spalten, die wie der Name schon beschreibt, in GROUP BY Bedingungen eingesetzt werden und in denen mehr als 1000 verschiedene Werte vorhanden sind. Der HG Index sollte ferner dann eingesetzt werden, wenn der LF Index aufgrund zu geringer Zeilenzahlen nicht ausreichend effizient ist und mehrere Spalten in einem Index zusammengefasst werden müssen.

Der HG Index ist eine spezielle Datenstruktur, die aus verschiedenen Indexsystemen besteht und benötigt deshalb den meisten Aufwand an Ressourcen unter allen Indextypen in Sybase IQ. HG Indizes können nicht auf Spalten mit Bit Datentypen oder Zeichenketten (Char / VarChar) mit mehr als 255 Byte Breite eingesetzt werden und generell nicht für Spalten mit Gleitkommazahlen (Float, Real, Double) empfohlen. Dieser Index ist eine Alternative zum LF Index, d.h. es macht wenig Sinn sowohl LF als auch HG Indizes auf derselben Spalte zu verwenden.

HG Indizes werden automatisch angelegt, wenn Primärschlüssel und /oder Fremdschlüssel angelegt oder Constraints zur Eindeutigkeit (UNIQUE) verlangt werden.

Betrachten wir nun die Reihe der spezielleren Indizes, die sich nach der Anwendung orientieren:

Der High Non Group Index (HNG)

Dieser High Non Group Index eignet sich insbesondere für Filterfunktionen in eingegrenzten Bereichen (Range Search / BETWEEN) sowie Aggregaten auf Spalten mit hoher Kardinalität.

Eine mögliche Anwendung bietet sich z.B. auf ganzzahligen Zählspalten (Mengen), eine weitere in Kombination mit HG Indizes, jedoch nicht wenn Eindeutigkeit (UNIQUE) verlangt wird, auf Datentypen mit Gleitkomma, Bit und Zeichenketten > 255 Bytes.

Der Compare Index (CMP)

Der Compare Index wird innerhalb derselben Tabelle angewendet und beschreibt die

Beziehung zweier Spalten identischen Datentyps, Länge und Genauigkeit einer Tabelle mit binärem Vergleich, d.h. (>, <, =).

Die Anwendung dieses Index bietet sich an bei Messreihen, Daten für Histogramme oder in stark denormalisierten Tabellen (R-Cubes) mit Jo.

Der Containment Index (WD)

Dieser Index wird auch Word Index genannt. Dabei wird jedes, durch ein Whitespace (Leerzeichen, Tab..) oder Trennzeichen (Delimiter) wie Komma, separierte Wort indexiert. Hier handelt es sich nicht um Volltextsuche oder Sprachausdrücke - dieser Index ist ausschließlich für Charakter-Spalten, d.h. Char, Varchar oder Long Varchar, gedacht¹ und wird nur mit LIKE oder CONTAINS Befehlen genutzt.

Die Anwendung des Index ist gegeben, wenn Kurzmitteilungen nach Schlüsselbegriffen zu durchsuchen sind, für die Suche in archivierten Emails und Nachrichten, die aus einem Webformular stammen, so wie überall da wo unscharf mit Wildcards gesucht wird.

Der Index auf Datum und Zeit (DTM DT TM)

Die Indizes für Datum, Zeit, Zeitstempel (Timestamp) funktionieren inhaltlich identisch, so dass sie in einem Kapitel vorgestellt werden können. DTM (Date Time) ist die Kombination von Datum (Date DT) und Zeit (Time TM). Sybase IQ stellt neben den oft anzutreffenden kombinierten DateTime Datentypen auch getrennte Datentypen, die nur das Datum oder nur die Zeit speichern. Der Index nutzt die Tatsache, dass das Datum bzw. die Zeit strukturiert abgebildet und gelesen wird, 10000 Jahre, 12 Monate, 31 Tage ... 24 Stunden etc. Damit ist für jedes Element der Struktur die Anzahl der Elemente begrenzt.

Die typischen Probleme einer Integer-Indexierung mit B-Tree's mit der erforderlichen Rebalancierung ist zum einen leistungsfordernd, zum anderen übersteigen die Kardinalität von Zeitstempeln die Fähigkeiten eines üblichen Bitmap-Index. Das Verfahren von Sybase IQ mit der speziellen Speicherung von Datum und Zeitstrukturen ist ideal für Zeitreihenanalysen, extrem platzsparend und performant. Natürlich unterstützt dieser Index nicht nur Zeitreihen sondern darüber hinaus alle Kalkulationen und Berichte mit Bedeutung auf Zeiträumen.

Join-Index

Der Join-Index dient zur Beschleunigung von Joins in Sternmodellen oder Datenmodellen in mehrstufiger Hierarchie. Der Einsatz des Join-Index ist abhängig vom Datenmodell, der Differenz in der Datenmenge der Join-Spalten und der Änderungsfrequenz. Join-Indizes sind „von Natur aus“ multidimensional und es ist möglich

¹ Long Varchar ist als LOB Option für Sybase IQ lizenzierbar.

mehrere Tabellen miteinander zu verknüpfen. Die Effizienz des Index zeigt sich insbesondere auf kleineren Systemen oder bei stark normierten Datenmodellen.

Einsatz von Indizes in Sybase IQ

Nachdem nun alle verfügbaren Indextypen vorgestellt wurden, stellt sich nun die Frage wie beim Einsatz von Indizes vorgegangen werden sollte:

Für neu aufgesetzte Systeme sollte, bevor auch nur das erste Tabellenobjekt angelegt wird, die MINIMIZE_STORAGE Option gesetzt werden.

Index Advisor

Sybase IQ verfügt über einen eingebauten Index-Ratgeber, dies ist der einfachste, jedoch zeitintensivere Weg, um die für die jeweiligen Abfragen empfohlenen Indizes anzulegen. Bei eingeschaltetem Index Advisor werden die Empfehlungen in einer Tabelle abgelegt, danach steht eine Liste der anzulegenden Indizes zur Verfügung. Die Indizes können dann angelegt werden, wenn es im Betriebsablauf sinnvoll ist. Dieses Verfahren ist leider sehr langwierig, da die Abfrageanalyse die erforderlichen Indizes identifiziert und diese anhand gestellter Abfragen anlegt, sodass nicht sofort die volle Leistung zur Verfügung steht.

Anlegen von Indizes nach Eindeutigkeit und Kardinalitätsanalyse

Die Eindeutigkeit von Schlüsseln sollte vor dem ersten Datenladen bestimmt sein, sodass bereits vor dem ersten Laden der Daten die Primärschlüssel bestimmt sind.

Durch die Primärschlüssel und Eindeutigkeit wird implizit eine Indizierung angelegt und die Abfrageoptimierung unterstützt. Nach dem Laden der Daten werden für jede Spalte die Anzahl der voneinander verschiedenen Datenelemente und die Datensatzanzahl bestimmt. Danach werden LF, HG und bei Datum/Zeit-Spalten DTTM Indizes unmittelbar angelegt. Damit ist eine Vollindexierung aller Spalten bei Beginn des Betriebs gegeben. Dies bringt etwas Aufwand bei der Systemeinrichtung mit sich, dafür steht allerdings von Beginn an eine hohe Leistung zur Verfügung. Weiter Indizes werden nach Bedarf vom Index Advisor oder aufgrund der zu den jeweiligen Indizes genannten Anwendungsfälle angelegt.

SYBASE GMBH
PRINZENALLEE 13
40549 DÜSSELDORF
Tel: +49 211 5976 0

www.sybase.de

Copyright © 2010 Sybase, Inc. All rights reserved. Unpublished rights reserved under U.S. copyright laws. Sybase, and the Sybase logo are trademarks of Sybase, Inc. or its subsidiaries. All other trademarks are the property of their respective owners. ® indicates registration in the United States. Specifications are subject to change without notice. 3/09.

SYBASE®