Scalable Data Engineering, Lösungsvorschlag Prüfung WS 2022/23

Henry Haustein

Welche 2 Techniken zum Beschleunigen von SQL-Querys gibt es?

Indizes und Materialized Views

2 SQL-Abfragen wurden gegeben → Welche Optimierungstechniken kann man darauf anwenden?

Indizes lohnen sich insbesondere, wenn es viele Joins gibt, Materialized Views eignen sich, wenn die Query (bzw. eine ähnliche Query) oft ausgeführt wird.

Was ist Database Cracking und wie funktioniert es?

Database Cracking erstellt einen Index während Range-Querys bearbeitet werden. Die Query wird nicht nur bearbeitet, sie sortiert auch (teilweise) den Cracker Index. Kommt z.B. eine Query rein, die nach $5 \leq X \leq 10$ fragt, so erstellt der Cracker Index 3 Bereiche: < 5, zwischen 5 und 10 und > 10. Eine zweite Query, die nach $6 \leq X \leq 7$ fragt, braucht dann nicht mehr die gesamte Liste bearbeiten, sondern kann sich gleich auf den Bereich zwischen 5 und 10 kümmern. Nachdem auch diese zweite Query abgearbeitet wurde, hat man < 5, zwischen 5 und 6, zwischen 6 und 7, zwischen 7 und 10, > 10 als Bereiche des Index.

Welche 2 Typen von materialized views gibt es? Worin unterscheiden sie sich? Wie funktioniert die Wartung?

Die zwei Typen sind: Join Views und Aggregate Views. Sie unterscheiden sich darin, welche Daten sie speichern: Join Views speichern das Ergebnis von Joins und können deswegen sehr schnell sehr groß werden, während Aggregate Views die Ergebnisse von Summen, Counts, Averages, etc. speichern.

Grundsätzlich müssen beide Arten von Views regelmäßig aktualisiert werden, wenn man bei Aggregation Views die *richtigen* Funktionen nimmt (*additive aggregation functions* wie `SUM`, `COUNT`) ist die Aktualisierung auch recht einfach und geht schnell, es muss nicht nochmal die gesamte Tabelle verarbeitet werden, sondern nur der neue Eintrag.

Wie ist der Data Cube aufgebaut? Welche Funktionen gibt es?

Der Data Cube ist eine multidimensionale Datenstruktur, die dazu dient, große Datenmengen in verschiedenen Dimensionen zu analysieren. Die Struktur des Data Cubes basiert auf einer n-dimensionalen Matrix,

die die verschiedenen Datenattribute in den verschiedenen Dimensionen speichert. Die Dimensionen des Data Cubes werden normalerweise als Achsen dargestellt und können beliebige Datenattribute umfassen, wie z.B. Produkte, Regionen, Zeitperioden oder andere relevante Kategorien. Die Werte der Datenattribute in jeder Dimension werden in der Matrix als Zellen gespeichert.

Die Funktionen des Data Cubes umfassen:

- Slice and Dice: Der Data Cube kann verwendet werden, um Daten in verschiedenen Dimensionen zu filtern und zu gruppieren. Die Slice- und Dice-Funktion ermöglicht es, Daten basierend auf bestimmten Kriterien zu gruppieren und die Ergebnisse in verschiedenen Formen und Hierarchien darzustellen.
- Drill-Down: Der Data Cube ermöglicht es, von einer höheren Hierarchieebene zu einer detaillierten Ebene zu navigieren. Mit Drill-Down können Benutzer beispielsweise aggregierte Daten nach Regionen, Produkttypen oder bestimmten Produkten untersuchen.
- Roll-Up: Der Data Cube ermöglicht es, aggregierte Daten in höheren Hierarchieebenen darzustellen. Roll-Up ermöglicht es Benutzern, aggregierte Daten auf höheren Ebenen darzustellen und einen Überblick über die Gesamtperformance zu erhalten.

Was ist das Star Schema, wie ist es aufgebaut und wie hängt es mit dem Data Cube zusammen?

Das Star Schema ist eine spezielle Art von Datenmodell, das für die Erstellung von Data Warehouses verwendet wird. Das Star Schema besteht aus einer zentralen Tabelle (Fact Table) und mehreren umliegenden Tabellen (Dimension Tables). Der Fact Table enthält die Kerninformationen, die für die Analyse benötigt werden, während die Dimension Tables die Details zu den Dimensionen der Analyse liefern, wie beispielsweise Zeit, Produkt, Region, Kunde und Verkaufskanal.

Der Fact Table enthält normalerweise numerische Messwerte wie z.B. Umsatz, Menge, Gewinn, Kosten und dient als Faktentabelle für die Analyse. Die Dimension Tables enthalten die Beschreibungen für die Faktentabelle und ermöglichen es, die Faktentabelle nach verschiedenen Dimensionen zu filtern und zu gruppieren.

Das Star Schema ist eng mit dem Data Cube verbunden. Der Data Cube basiert auf der gleichen multidimensionalen Datenstruktur wie das Star Schema und verwendet ebenfalls Dimensionen und Faktentabellen, um komplexe Datenanalysen durchzuführen. Der Data Cube nutzt jedoch eine andere Darstellung des Star Schemas, indem es die Daten in einer n-dimensionalen Matrix speichert.

Was sind die Unterschiede zwischen Facts und Measures?

Facts sind die konkreten Daten, die in einer Faktentabelle gespeichert werden, während Measures die numerischen Werte sind, die mit diesen Facts verbunden sind.

Measures sind Berechnungen auf der Grundlage von Facts. Eine Measure verweist auf einen Fact (z. B. die Bestellmenge) und gibt eine Aggregationsfunktion an (z. B. Durchschnitt oder Summe).

Facts können nicht aggregiert werden, da sie bereits konkrete Daten darstellen. Measures hingegen können aggregiert werden, um ein besseres Verständnis der Daten zu ermöglichen.

Facts sind der Ausgangspunkt der Analyse, während Measures als Indikatoren der Leistung dienen und es ermöglichen, Trends und Muster in den Daten zu erkennen.

Wie müssen die Daten für die Rollup-Funktion aufgebaut sein?

Es wird eine Hierarchie benötigt, entlang derer die Aggregation stattfinden kann. Um die Rollup-Funktion anzuwenden, müssen die Daten so strukturiert sein, dass sie in jeder Hierarchieebene aggregiert werden können.