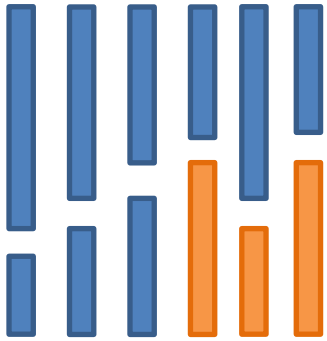


Южно-Уральский государственный университет
6-я научная конференция аспирантов и докторантов

Методы обработки запросов с использованием распределенных колоночных индексов

Иванова Елена Владимировна

Научный руководитель:
доктор физ.-мат. наук,
профессор Л.Б. Соколинский

Актуальность исследования

- Согласно отчету аналитической компании IDC к 2020 г. количество данных в мире достигнет 40 Зеттабайт. Современные технологии баз данных не могут обеспечить обработку столь крупных объемов данных. Из всего объема потенциально полезных данных в 2012 г. всего лишь 3% данных были проиндексированы и только 0.5% были подвергнуты анализу.
- Технологии решения:
  - параллельные системы баз данных
  - многоядерные ускорители
  - поколоночное хранение данных
- В соответствие с этим актуальной является задача разработки новых эффективных методов параллельной обработки и анализа сверхбольших объемов структурированных данных на кластерных вычислительных системах, оснащенных многоядерными ускорителями, с использованием колоночного представления и сжатия данных.

Адресная функция

- Пусть имеется отношение R, состоящее из n кортежей. Пусть на множестве кортежей $R = \{r_0, r_1, \dots, r_{n-1}\}$ задано отношение линейного порядка:

$$r_0 < r_1 < \cdots < r_{n-1}\,.$$

- ***Адресной функцией*** будем называть целочисленную функцию

$$@: R \to \{0, \dots, n-1\},$$

определяющую для кортежа $r \in R$ его порядковый номер $@(r)$ в упорядоченной последовательности.

Примечание. Для функции @ также используется операторная форма ее применения к кортежу r: $@r$.

Колоночный индекс

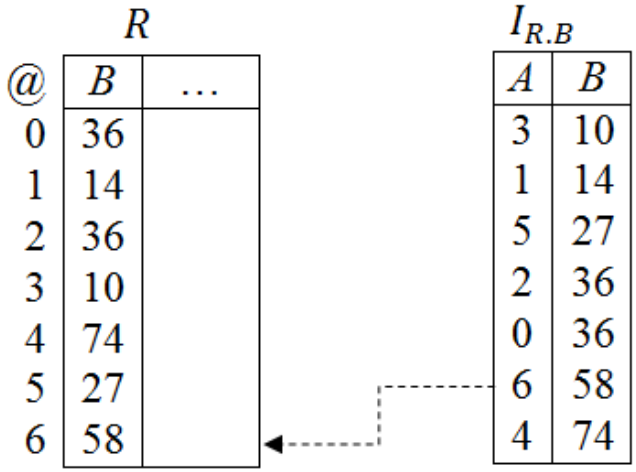
- Пусть задано отношение $R(B, \dots)$. Обозначим через $\mathfrak{D}_B$ домен атрибута B. Пусть на множестве $\mathfrak{D}_B$ задано отношение линейного порядка. ***Колоночным индексом*** $I_{R.B}$ атрибута B отношения R будем называть упорядоченное отношение $I_{R.B}(A, B) \in (\mathbb{Z}_{\geq 0}, \mathfrak{D}_B)$, удовлетворяющее следующим требованиям:

$|I_{R.B}| = n$ и $\pi_A(I_{R.B}) = \{0, \dots, n-1\}$;

$\forall x_1, x_2 \in I_{R.B}(x_1 \leq x_2 \Leftrightarrow x_1.B \leq x_2.B)$;

$\forall r \in R(\forall x \in I_{R.B}(@r = x.A \Rightarrow r.B = x.B))$;

где $n = |R|$.



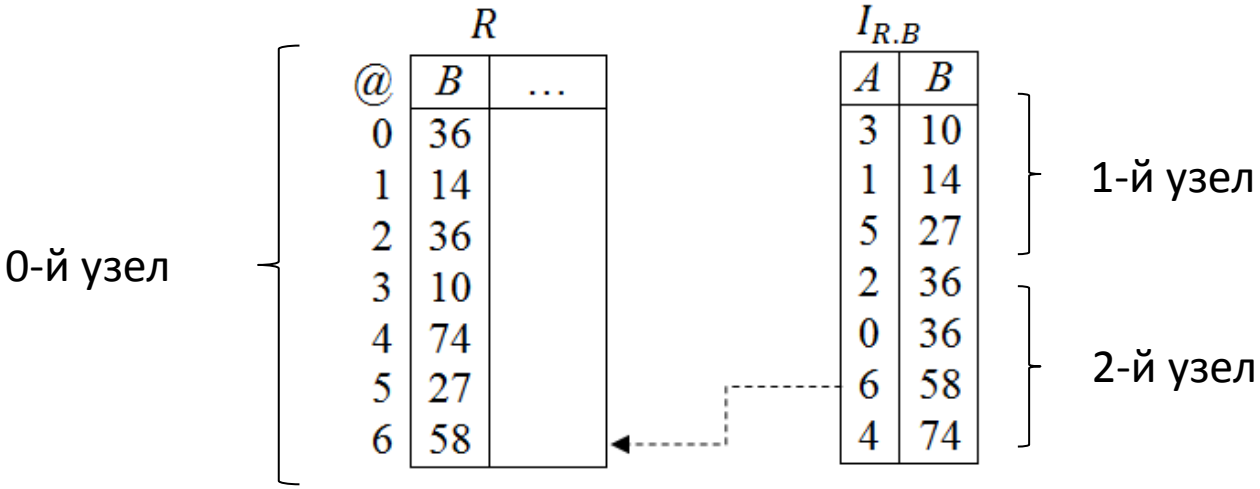
Фрагментация колоночного индекса

- Пусть необходимо фрагментировать $I_{R.B}$ по k узлам многопроцессорной системы. Функция фрагментации $\varphi: I_{R\ B} \to \{0,1,\dots,k-1\}$. Обозначим фрагментированный индекс как $I^i_{R.B}$, где i – номер узла многопроцессорной системы, на котором хранится фрагмент.

$$I_{R.B} = \bigcup_{i=0}^{k-1} I^i_{R.B},$$

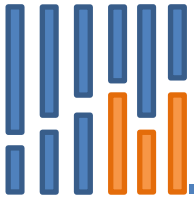
$$I^i_{R.B} \cap I^j_{R.B} = \emptyset, \text{ при } i \neq j$$

Колоночный распределенный индекс



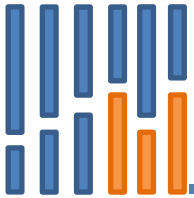
Если два атрибута имеют один и тот же домен, то их одинаковые значения в результате фрагментации гарантированно попадут на один и тот же узел кластерной системы.

Колоночный распределенный индекс хранится и обрабатывается в оперативной памяти узла кластерной системы в сжатом виде.



Реляционные операции для запросов с распределенным колоночным индексом

- Разработаны следующие операции:
  - естественное соединение по одному атрибуту
  - тета-соединение по одному атрибуту
  - пересечение



Заключение

- Разработана модель распределенного колоночного индекса для обработки запросов к сверхбольшим базам данных
- Разработаны некоторые реляционные операции для выполнения запросов с использованием распределенного колоночного индекса
- Планируется реализовать распределенный колоночный индекс в виде сопроцессора СУБД.