Гибридная модель данных – будущее высокопроизводительных СУБД для кластерных вычислительных систем с многоядерными ускорителями*

К.В. Бородулин, Л.Б. Соколинский

Южно-Уральский государственный университет

Согласно прогнозам аналитической компании IDC, количество данных в мире удваивается каждые два года и к 2020 г. достигнет 44 Зеттабайт, или 44 триллионов гигабайт. Единственным эффективным решением проблемы хранения и обработки сверхбольших баз данных является параллельная обработка запросов на многопроцессорных вычислительных системах с распределенной памятью [1]. В последнее время значительной популярностью при обработке сверхбольших объемов данных стали пользоваться модели noSQL. Подобный подход демонстрирует очень хорошую масштабируемость, однако он уступает классической реляционной модели по возможностям оптимизации запроса. Нами предлагается *гибридная модель данных*, позволяющая соединить преимущества обоих подходов. База данных в *гибридной модели* представляется в виде объектов двух классов: *доменные индексы* и *отношения ключей*.

Доменным индексом $I_{\mathfrak{D}}$ домена $\mathfrak{D}$ называется отношение $I_{\mathfrak{D}}(Y,V)$, представляющее собой множество кортежей вида (y,v), где $y \in \mathbb{Z}_{\geq 0}$, $v \in \mathfrak{D}$. Атрибут Y отношения $I_{\mathfrak{D}}$ будем называть *ключом*, а атрибут V – *значением*. Ключ однозначно определяет значение, и большему ключу всегда соответствуют большее значение. Множество доменных индексов $\{I_{\mathfrak{D}_1},\ldots,I_{\mathfrak{D}_m}\}$ *индексирует* все неключевые значения базы данных $\mathfrak{B}$.

Отношение ключей Q, представляющее реляционное отношение R, – это отношение вида $Q(K^*,Y_1,\ldots,Y_n)$, $K \subset \mathfrak{D}_0 = \mathbb{Z}_{\geq 0}$ где K – первичный ключ, $Y_j \subset \mathbb{Z}_{\geq 0}$ $(j=1,\ldots,n)$, удовлетворяющее свойству

$$\forall r \in R\left(\exists q \in Q\left(\forall j \in \{1,\ldots,n\}\left(\exists x \in I_{\mathfrak{D}_{a_j}}\left(r.A_j = x.V \wedge x.Y = q.Y_j \wedge r.K = q.K\right)\right)\right)\right). \quad (1)$$

Другими словами, каждому кортежу r из R соответствует в точности один кортеж q из Q с совпадающим значением первичного ключа K, атрибуты которого через соответствующие доменные индексы указывают на соответствующие значения кортежа r. Поскольку K также является первичным ключом в Q, то из (1) следует:

$$\forall q \in Q\left(\exists r \in R\left(\forall j \in \{1,\ldots,n\}\left(\exists x \in I_{\mathfrak{D}_{a_j}}\left(r.A_j = x.V \wedge x.Y = q.Y_j \wedge r.K = q.K\right)\right)\right)\right). \quad (2)$$

Гибридная модель данных позволяет эффективно размещать данные в распределенной оперативной памяти за счет сжатия данных. При этом, предполагается размещать доменные индексы в сжатом фрагментированном виде в оперативной памяти многоядерных ускорителей, что обеспечивает эффективное использование последних. Гибридная модель сочетает возможность глубокой оптимизации запросов с эффективными методами обработки больших объемов информации на базе колоночного представления данных и использования модели «ключ-значение».

Литература

1. Соколинский Л.Б. Параллельные машины баз данных // Природа. 2001. № 8. С. 10–17.

* Работа выполнена при финансовой поддержке Минобрнауки РФ в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014—2020 годы» (Госконтракт № 14.574.21.0035).