УДК 004.652

А.Е. Петров, А.И. Лапуста

ВЫБОР МОДЕЛИ ПРЕДСТАВЛЕНИЯ ДАННЫХ В АВТОМАТИЗИРОВАННОЙ СИСТЕМЕ ПРОВЕДЕНИЯ ОПРОСОВ

Рассмотрена проблема хранения и обработки данных при проведении опросов в маркетинговых исследованиях. Для изучения проблемы рассмотрены варианты хранения структуры и результатов в реляционных и альтернативных базах данных, их преобразование для последующей аналитической обработки.
Ключевые слова: базы данных, слабо-структурированные данные, параллельные вычисления, электронные формы..

При проектировании баз данных разработчики часто сталкиваются с ситуацией, когда количество атрибутов, которыми описывается объект, является огромным, может меняться со временем или вовсе не известно заранее.

Электронные опросы и информация о клиентах в CRM, сведения о пациентах в медицинских автоматизированных системах, карточки товаров в интернет-магазинах – типичные примеры данных с динамической структурой. Как же хранить такие данные?

В реляционных системах управления базами данных известны три популярных подхода к решению такой задачи:

1. Entity-Attribute-Value (Сущность-Атрибут-Значение)
2. Динамическая таблица
3. Сериализованные данные

Схема «Сущность-Атрибут-Значение» [2] является одной из самых популярных за счет своей гибкости. В основе лежит принцип нормализации данных в БД, создаются отдельные таблицы для атрибутов, сущностей и таблица значений (табл. 1)

Поля в столбцах «сущность» и «атрибут» являются внешними ключами для соответствующих таблиц. Таким образом, анкета из 30 вопросов будет представлять собой 30 строк в данной таблице: это накладывает определенный «штраф» на производительность при чтении или записи большого количества сущностей, а аналитика на таких таблицах требует сложных запросов с объединением таблиц.

Наиболее очевидным решением является динамическая таблица, столбцы которой представляют все возможные атрибуты (табл. 2).

Таблица 1

id	сущность	атрибут	значение
1	A (анкета)	X (имя)	Вася
2	A (анкета)	Y (фамилия)	Петров
3	A (анкета)	Z (возраст)	32

Таблица 2

id	имя	фамилия	возраст	…
1	Вася	Петров	32	…
2	Анна	Иванова	18	…

Очевидным плюсом является скорость записи и чтения, однако из-за особенностей внутренней структуры современные БД накладывают ограничение на количество столбцов, а при их добавлении блокируют таблицу на время, зависящее от количества строк.

В подходе с сериализованными данными, сущность целиком сохраняется в формате XML или JSON (табл. 3).

Такой подход не позволяет использовать встроенные функции БД и каждый раз требует преобразовывать данные, что не делает его пригодным для аналитики.

Одним из подходов, набирающих популярность в последнее время, стал отказ от традиционных РСУБД и переход на нереляционные базы данных, основанные на модели хранения «ключ-значение» (табл. 4).

Такой подход позволяет хранить слабо структурированные данные. Его также часто называют NoSQL, потому что базы данных этого типа не поддерживают языка запросов SQL. Вместо этого используется принцип MapReduce [1], который позволяет параллельно обрабатывать большие объемы информации: на первом этапе (map) происходит разделение сущностей на группы и передача их в параллельные потоки обработки данных. На втором этапе (reduce) происходит свертка уже обработанных данных.

Таблица 3

id	data
1	<form> <name>Вася</name> <surname>Петров</surname> <age>32</age> </form>

Таблица 4

	значение
1	**ключ** **значение** имя Иван фамилия Петров возраст 32

Такие БД обладают определенными преимуществами:

1. параллельные вычисления;
2. горизонтальное масштабирование;
3. хранение слабо-структурированных данных.

Поэтому они получили большое распространение в крупных интернет-компаниях: Google (BigTable), Amazon (Dynamo) и Facebook (Cassandra). В частности Facebook открыл свою разработку для сообщества, разместив проект в сообществе Apache. Среди других открытых и бесплатных продуктов можно отметить проекты Apache Hbase, Apache CouchDB и MongoDB.

СПИСОК ЛИТЕРАТУРЫ

1. *Jeffrey Dean, Sanjay Ghemawat* «MapReduce: Simplified Data Processing on Large Clusters» - Google, 2004.

2. *Nadkarni, Prakash* «The EAV/CR Model of Data Representation» - Yale, 1999.

ГИАБ

КОРОТКО ОБ АВТОРАХ

Петров Андрей Евгеньевич – профессор, доктор технических наук, helen_pet@mail.ru
Лапуста Алексей Игоревичт – студент, lapusta@gmail.com
Московский государственный горный университет,
Moscow State Mining University, Russia, ud@msmu.ru