Борисенков М.С. Применение виртуализации при анализе и синтезе больших многомерных данных // Международный журнал информационных технологий и энергоэффективности. – 2023. –
Т. 8 № 7(33) с. 23–30

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004

ПРИМЕНЕНИЕ ВИРТУАЛИЗАЦИИ ПРИ АНАЛИЗЕ И СИНТЕЗЕ БОЛЬШИХ МНОГОМЕРНЫХ ДАННЫХ

Борисенков М.С.
ФГБОУ ВО «Технологический университет имени дважды Героя Советского Союза, летчика-космонавта А.А. Леонова», Королёв, Россия (141074, Московская область, город Королёв, ул. Гагарина, д.42), e-mail: borisenkov.matvey@mail.ru

В данной статье рассматривается технология виртуализации в контексте анализа больших многомерных данных. Приводятся ключевые понятия и объясняется взаимосвязь между появлением технологии и прорывом в сфере анализа данных. Приводятся концепции, с помощью которых в связке с виртуализацией можно получить наиболее точные результаты анализа данных и синтеза результатов анализа.

Ключевые слова: Виртуализация, OLAP, большие данные, многомерные данные, гипервизор.

APPLYING VIRTUALIZATION TO THE ANALYSIS AND SYNTHESIS OF LARGE MULTIVARIATE DATA

Borisenkov M.S.
"Technological University named after Twice Hero of the Soviet Union, Cosmonaut A.A. Leonov», Korolev, Russia (141074, Moscow region, Korolev, Gagarin str., 42), e-mail: borisenkov.matvey@mail.ru

This article discusses virtualization technology in the context of big multidimensional data analysis. The key concepts are presented and the relationship between the emergence of the technology and the breakthrough in the field of data analysis is explained. The methodologies that can be used in conjunction with virtualization to obtain the most accurate results of data analysis and synthesis of analysis results are presented.

Keywords: Virtualization, OLAP, big data, multidimensional data, hypervisor.

1. Виртуализация как фундаментальное понятие современной IT сферы.

В условиях лавинообразного развития технологий, увеличения потоков входных/выходных данных возникла потребность оптимизации использования вычислительных ресурсов.

В начале 1960-х годов зародилась концепция разделения времени (time-sharing) – распределение вычислительных ресурсов между несколькими пользователями: пока один вводит данные, машина занимается расчетами других. Но данная концепция небезопасна, сложна, не отличается нужной стабильностью, как следствие возникновение потребности в

создании новых технологий, которые смогут не только заменить концепцию разделения времени, но и предоставить пользователям новые возможности [1].

Командой инженеров IBM был предложен совершенно новый подход – в рамках одной ЭВМ предоставить каждому пользователю виртуальную машину со своей ОС. Пользователи подключались к гостевым ОС с помощью специальных устройств ввода-вывода – терминалов.

Виртуализация обладала существенными преимуществами над концепцией разделения времени:

- Увеличенные надежность и безопасность за счет изоляции пользователей.
- Запуск любых приложений (не только приспособленных к концепции разделения времени) за счет симуляции отдельного компьютера для каждого пользователя.
- Увеличенная производительность за счет использования легковесных гостевых ОС.

Виртуализация - создание абстракций над аппаратным обеспечением. Другими словами, это сокрытие реализации за универсальным методом обращения к ресурсам. Среди множества разновидностей виртуализации следует выделить три наиболее фундаментальные:

- Аппаратная виртуализация;
  Позволяет создавать изолированные и независимые виртуальные компьютеры с помощью программной имитации ресурсов физического/виртуального сервера.
- Виртуализация рабочих столов;
  Данная технология позволяет отделить логический рабочий стол от физической инфраструктуры.
- Контейнеризация;
  Позволяет запускать ПО в изолированных на уровне операционной системы пространствах [2-3].

Архитектура первого поколения виртуализации выглядит следующим образом (Рисунок 1).

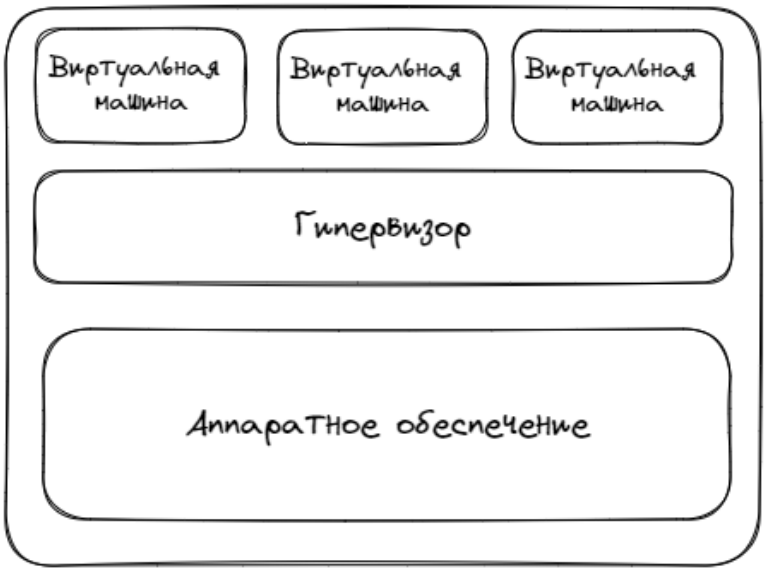

Рисунок 1 – Первое поколение виртуализации

Гипервизор выполняется как контрольная программа непосредственно на аппаратной части компьютера и не требует операционной системы общего назначения [4-5].

Архитектура современных технологий виртуализации (Рисунок 2, 3).

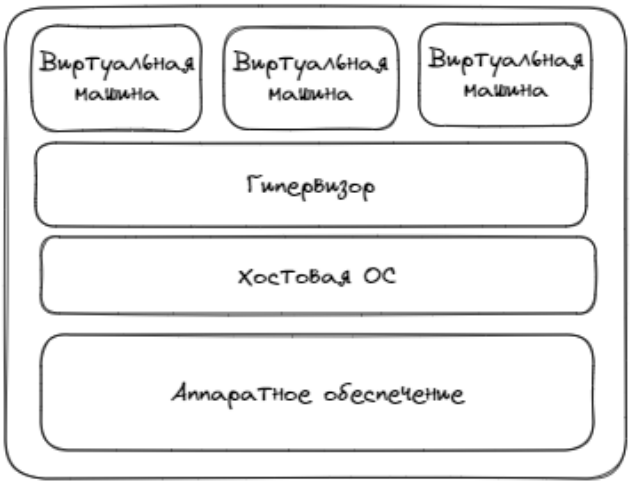

Рисунок 2 – Современная архитектура виртуализации №1

В данной схеме, гипервизор выполняется поверх хостовой операционной системы и управляет гостевыми ОС, а эмуляцией и управлением физическими ресурсами занимается хостовая ОС.

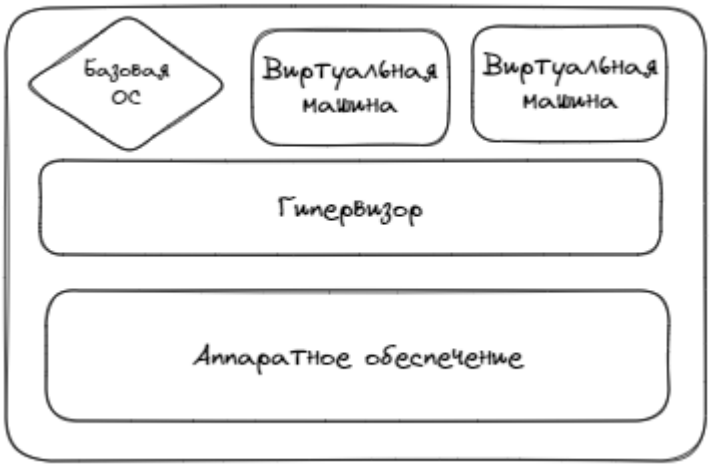

Рисунок 3 – Современная архитектура виртуализации №2

Данная модель сочетает в себе характеристики предыдущих двух архитектур. Здесь гипервизор выполняется поверх базовой операционной системы. После установки гипервизора ядро ОС переходит в режим поддержания виртуализации и передает управление ресурсами процессора и памяти гипервизору. Данная модель наиболее распространена по причине хорошей совместимости с оборудованием. Эта модель быстрее и эффективнее распределяет нагрузку и ресурсы между всеми модулями виртуализации, поэтому повсеместно используется для разработки/тестирования программного обеспечения и работы с большими данными [6-7].

2. Виртуализация данных как способ обработки больших многомерных данных.

С учетом постоянно растущего в геометрической прогрессии потока информации, обычные методы анализа становятся все менее и менее актуальны. Поэтому, было предложено использование технологии виртуализации, как основы для анализа больших многомерных данных. Эта идея оправдала себя на 200%.

Виртуализация данных - способ организации доступа к данным, при котором не требуется информация об их структуре и принадлежности к конкретной информационной системе.

Основой виртуализации данных является выполнение распределенных процессов управления данными, для запросов к многочисленным разнородным источникам данных и объединение результатов в виртуальные представления (Рисунок 4).

Особенности виртуализации данных

- Значительное ускорение создания объектов данных.
- Модель виртуализации данных обеспечивает практически полную сохранность и безопасность данных.
- Явное объединение не типизированных данных из различных источников.
- Гибкость в анализе многомерных данных и модернизация существующих технологий или добавление новых за счет модульной реализации технологии.

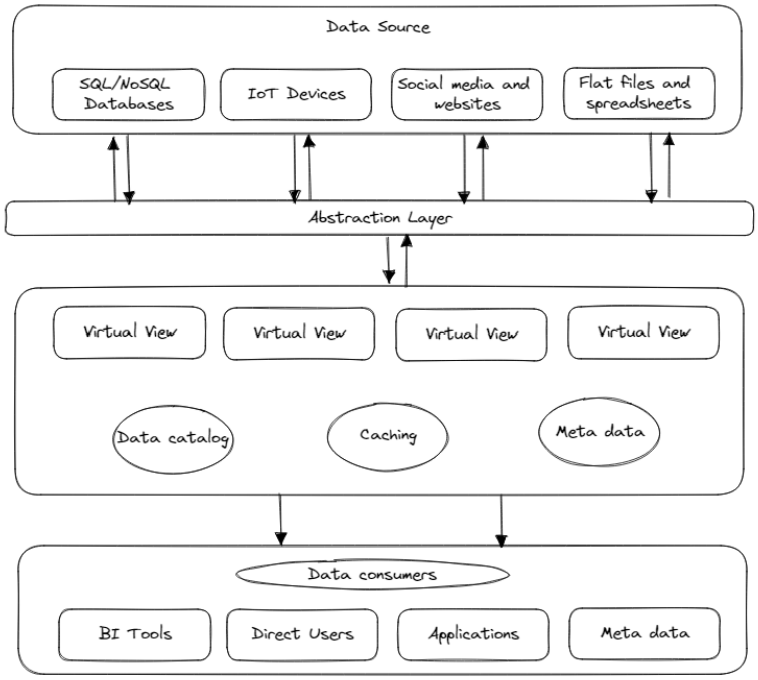

Рисунок 4 – Модель виртуализации данных

Первым уровнем в модели является уровень соединения (Data Source). На нем, с помощью соединителей и протоколов связи реализуется доступ к данным, рассредоточенным по многочисленным системам, которые содержат как структурированные, так и неструктурированные данные. Платформы для виртуализации данных могут работать с большинством современных СУБД и облачных хранилищ.

Вторым уровнем является уровень абстракции (Abstract Layer), по-другому виртуальный или семантический уровень. Он является связующим звеном между всеми источниками данных и всеми пользователями, образуя всю основу виртуализации. Данный уровень содержит все логические представления для доступа к источникам. Сложность базовых структур данных скрыта от конечного пользователя, таким образом пользователи видят только схематические модели данных, что упрощает работу с данными для потребителя.

Третьим уровнем является уровень потребления (Data consumers). Этот уровень является единой точкой доступа к данным. В зависимости от типа потребителя для предоставления абстрактных представлений данных используется несколько протоколов соединений. Они могут взаимодействовать с виртуальным уровнем (Abstract Layer), используя язык SQL, различные API (REST, SOAP) и стандарты доступа (JDBC, ODBC).

С помощью описанной выше модели, построенной на современной архитектуре виртуализации, появилась возможность достаточно комфортно работать с большими многомерными данными, за счет того, что данные быстро распределяются и структурируются, что значительно облегчает дальнейшую обработку и анализ. Но для наиболее эффективного анализа данных и синтеза результатов исследования к виртуализации данных добавляется идеология OLAP.

3. OLAP как фундаментальная концепция анализа данных

Концепция OLAP кубов приобрела всеобщее признание так как позволяла получать достаточно неочевидные выводы при анализе данных, с помощью перебора двух многомерных множеств параметров относительно друг друга, другими словами, необходимо было проанализировать данные на стыке всех пар параметров каждого множества, что синтезировало наиболее точные результаты (Рисунок 5).

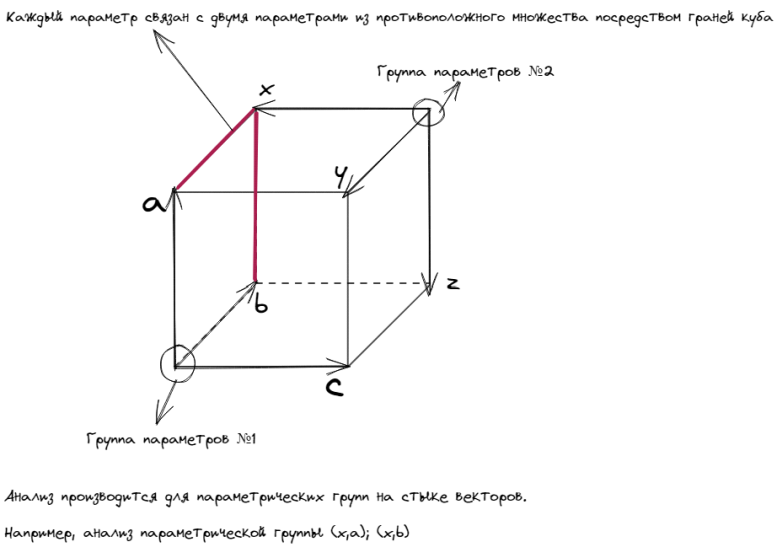

Рисунок 5 – Концепция OLAP кубов

С таким подходом сложные многомерные данные обретают понятную человеку структуру, что позволяет выделить ключевые параметры для анализа и получить результаты максимально точные и за минимально возможное время.

Методология OLAP предоставляет следующие возможности:

- Возможность любого логического, семантического и статистического анализа многомерных данных и сохранение результата в удобном для пользователя виде;
- Многомерное концептуальное представление данных, включая полную поддержку иерархий и множественных иерархий;
- Возможность обращения к данным независимо от их места расположения и объёма;
- Минимальное время синтеза результатов анализа, как правило время не превышает пяти секунд;

4. Преимущества виртуализации для анализа больших многомерных данных и синтеза результатов

Виртуализация данных обеспечивает доступ исходным данным в режиме реального времени и управление ими через логический уровень (Abstract Layer), исключая физическое перемещение больших объемов информации.

Внедрение виртуализации данных требует меньше затрат, чем разработка консолидированного хранилища с теми же задачами.

Отсутствует необходимость перемещать данные, а уровни доступа можно контролировать.

Независимо от типов и объемов данных пользователь может проводить любые анализы при минимальном уровне погружения в сферу науки о данных.
С добавлением методологии OLAP многомерный анализ становится бесконечным полем для экспериментов над многомерными данными, что способствует выявлению новых зависимостей и развитию технологий анализа и синтеза.

5. Практическое применение виртуализации при анализе большого потока данных

При непрерывном потоке данных возникает проблема с ее интерпритацией и хранением. Есть два варианта обработки потока данных. Первый заключается в том, чтобы просто записывать данные в том виде, в котором они передаются. Второй вариант более сложный, он заключается в преобработке данных, то есть в первичной структуризации данных по каким-либо параметрам, возможно так же формирование формы хранения, например в виде OLAP-кубов. Подход с преобработкой данных реализован в современных BI-системах (Рисунок №6). Используя BI-системы не нужно программировать sql-запросы, так как интерфейс позволяет извлекать данные без знания технических подробностей о них. Платформа формирует логических двойников данных и физические обращения к ним.

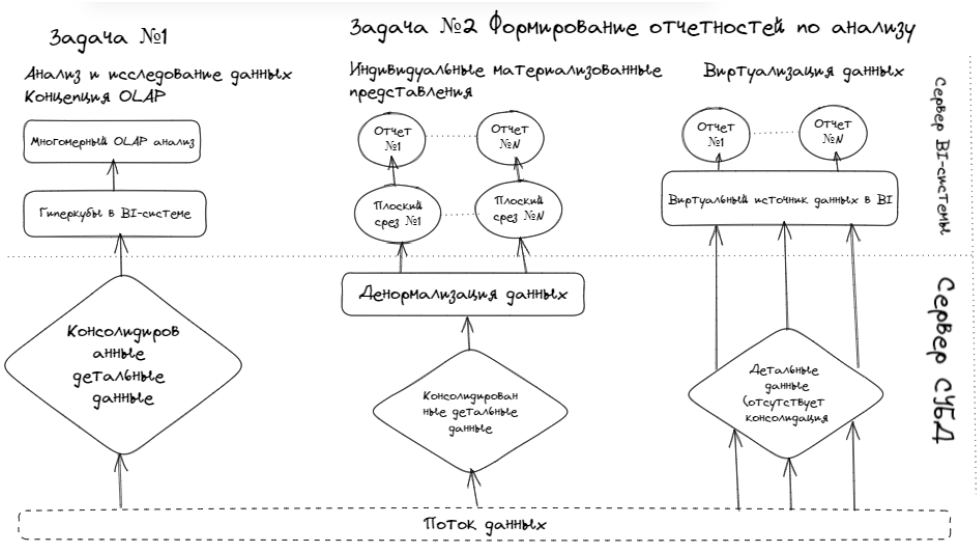

Рисунок 6 – Принцип работы BI-системы

На первом шаге в BI могут быть сформированы ROLAP-кубы (в которых указана связь с физическими данными). Затем структурированные многомерные OLAP-кубы трансформируются под шаблоны отчетов.

Заключение

В заключении необходимо подчеркнуть, что технология виртуализации дала огромный толчок направлению анализа данных, увеличив в разы вычислительные мощности без значительных трат на новое оборудование. Благодаря этому, все больше компаний внедряют виртуализацию в свои системы анализа данных, значительно увеличивая прибыль при минимальных вложениях.

Сама технология и архитектура виртуализации не стоит на месте, постоянно ведутся работы над оптимизацией распределения ресурсов между виртуальными модулями и над доработкой/изменением принципов работы самой технологии.

Список литературы

1. Турулин И.И. Виртуализация (изоляция вычислительных процессов) - учебное пособие. – Таганрог: ТТИ ЮФУ (бывший ТРТИ, ТРТУ), 2012. – 40 с.
2. Гультяев А.К. Виртуальные машины: несколько компьютеров в одном - СПб.: Питер, 2006. — 224 с. — ISBN 5-469-01338-3
3. Виртуальные машины [Электронный ресурс]. – Режим доступа: http://all-ht.ru/inf/vpc/p_0_0.html
4. Виртуальные машины [Электронный ресурс] // Информатизация и образование. – Режим доступа: http://hotuser.ru/server/1808-2009-12-02-08-28-47
5. Кулаичев А.П. Методы и средства комплексного анализа данных: Учебное пособие / А.П. Кулаичев. — М.: Форум, 2018. — 160 с.
6. Макшанов А.В. Технологии интеллектуального анализа данных. — М.: Лань. 2019. 212 с.

7. Симчера В.М. Методы многомерного анализа статистических данных / В.М. Симчера. — М.: Финансы и статистика, 2018. — 400 с.

References

1. Turulin I.I. Virtualization (isolation of computing processes) - a tutorial. - Taganrog: TTI SFU (former TRTI, TRTU), 2012. - p.40.
2. Gultyaev A.K. Virtual machines: several computers in one - St. Petersburg: Peter, 2006. - p. 224— ISBN 5-469-01338-3
3. Virtual machines [Electronic resource]. – Access mode: http://all-ht.ru/inf/vpc/p_0_0.html
4. Virtual machines [Electronic resource] // Informatization and education. – Access mode: http://hotuser.ru/server/1808-2009-12-02-08-28-47
5. Kulaichev A.P. Methods and means of complex data analysis: Textbook / A.P. Kulaichev. — M.: Forum, 2018. — p.160
6. Makshanov A.V. Data Mining Technologies. — M.: Lan. 2019. p.212
7. Simchera V.M. Methods of multidimensional analysis of statistical data / V.M. Simcher. — M.: Finance and statistics, 2018. — p.400