Обзор EMC Greenplum – решения для бизнес-аналитики

ОСНОВНЫЕ ТЕМЫ

- Высокопроизводительные системы
- ИТ-бизнес
- Дата-центры
- Серверы и системы хранения

Бизнес-аналитика – относительно новое направление обработки данных. Она позволяет извлечь прибыль из данных, которые раньше считались «мертвым» грузом с высокими затратами на хранение. Этим оно напоминает производства по добыче золота. Только темпы развития отличаются на порядки.
А проблемы и подходы – очень похожи.
В обоих случаях перерабатывается огромное количество “сырья”. Но в случае добычи редких металлов мы получаем огромные горы обеднённой породы, а в случае бизнес-аналитики – петабайты данных, из которых сложно извлечь какую-либо прибыль.
Но появляются новые технологии обогащения, и всё повторяется заново.

EMC Greenplum – «новое слово» в технологии аналитической обработки массивов накопленных данных.
Это высокопроизводительная многопоточная система аналитической обработки структурированных или неструктурированных данных, гибкая и бесшовная при масштабировании. Внедрение такой технологии позваляет оперативно анализировать накопленные данные для выявления дополнительных источников дохода.

Постепенно информация приобретает всё большую ценность. Причём с течением времени люди учатся добывать полезные данные из информации, которая раньше считалась совершенно бесполезной. Это чем-то напоминает добычу золота. То, что раньше шло в отвалы – сегодня успешно перерабатывается и превращается в полезный продукт. Проблемы этих двух (казалось бы, не связанных) отраслей тоже схожи. Если в девятнадцатом веке золотой рудник чаще всего представлял собой шахту и небольшой заводик по переплавке руды рядом, то сегодня рудники представляют собой огромные комплексы с множеством ступеней обогащения руды, горами отвалов и сотнями (а то и тысячами) людей, которые всё это обслуживают.

Тенденция развития бизнес-аналитики схожа с процессом эволюции золотодобычи. Только идёт это развитие не века, а годы, то есть в сотни раз быстрее. Там, где еще какие-то сорок лет назад сидел человек с бумагой и ручкой, сегодня стоят специальные автоматические комплексы анализа данных, которые умеют быстро перерабатывать гигантские объёмы данных и извлекать из них ценную информацию.

Проблема в том, что количество обрабатываемых данных растёт очень быстро и отнюдь не линейно. Соответственно системы для хранения и обработки этих данных должны быть мощными и легко масштабируемыми.
Именно к таким системам относится программно-аппаратный комплекс EMC GreenPlum.

Архитектура ПАК Greenplum состоит из трёх типов объектов:

1. **Мастер-сервера** – принимают и распределяют запросы пользователей. Для отказоустойчивости используются 2 сервера в конфигурации Active/Passive с синхронной репликацией метаданных.

2. **Сегментные сервера** – именно тут хранятся обрабатываемые данные и происходит их обработка. Для повышения отказоустойчивости каждый сервер хранит не только свои данные. Часть ёмкости заполнена данными, скопированными с других сегментных серверов.

3. **Сетевое соединение (gNet)** – обеспечивает взаимодействие серверов, составляющих систему друг с другом. Для отказоустойчивости gNet состоит из двух независимых ЛВС со скоростью передачи данных 10 Гб/с каждая.

Коммутаторы, входящие в состав gNet, соединяют сегментные сервера с мастер-серверами и друг с другом. К ним может быть также подключена система резервного копирования. Важной частью gNet является логика поэтапного выполнения запросов, которая позволяет осуществлять обмен данными между сегментами, а мастер-серверу при этом возвращается окончательный ответ на запрос.

Преимуществами такого подхода к построению ПАК являются:

1. **Высокая скорость загрузки и обработки запросов**
   Так как EMC Greenplum построен по идеологии MPP (Massive Parallel Processing), то даже сложные запросы обрабатываются с максимально возможной скоростью. Это происходит благодаря тому, что мастер-сервер, на который пришел запрос, анализирует его и разбивает на множество мелких подзапросов, которые могут выполняться не зависимо друг от друга на разных узлах ПАК.

2. **Неограниченная масштабируемость решения**
   Рост полезного объёма хранения и производительности ПАК происходит посредством добавления новых сегментных серверов. Архитектура решения не ограничивает количество этих узлов. Так что размеры EMC Greenplum ограничены только потребностями бизнеса. На сегодняшний день крупнейшая инсталляция ПАК Grenplum в мире может хранить более 6 ПБ данных.

3. **Полиморфное хранение обрабатываемых данных**
   В отличие от традиционных СУБД, хранящих данные строго построчно, EMC Greenplum может хранить обрабатываемые данные, как в строках, так и в колонках. За счёт этого радикально снижается нагрузка на дисковую подсистему ПАК в процессе статистического анализа данных.

4. **Обработка больших объёмов неструктурированной информации**
   EMC Greenplum полностью совместим с решением Apache Hadoop, реализующем парадигму MapReduce. Это позволяет быстро обрабатывать петабайтные объёмы данных. Примерами применения Apache Hadoop можно назвать параллельную индексацию больших объёмов текста, анализ логов посещения, статистический анализ пользователей с разными атрибутами, поиск ключевых фраз.

5. **Встроенная аналитика**
   EMC Greenplum умеет выгружать часть математической логики ведущих аналитических приложений на уровень хранилища. За счёт этого уменьшаются объёмы передаваемых данных, что в итоге даёт рост производительности системы в целом для таких задач, как кластерный и регрессивный анализ, нейронные сети и прочее.

6. **Стандартная архитектура**
   Программное обеспечение EMC Greenplum может быть развёрнуто на любых серверах стандартной архитектуры. Это даёт компаниям возможность выбора: либо самим закупать аппаратную платформу для EMC Greenplum, либо приобрести уже протестированный ПАК производства EMC.

На сегодняшний день и в мире, и в России уже есть большое количество пользователей EMC Greenplum. Greenplum используют такие компании как TMobile, NASDAQ, NYSE, Walmart, Skype, «Тинькофф Кредитные Системы» и ещё ряд других компаний.

Задача установки и интеграции EMC Greenplum в существующую инфраструктуру решается квалифицированными специалистами, в некоторых случая с привлечением экспертов в области Business Intelligence (BI).
Такие услуги вам может предложить компания CompTek, являющаяся официальным дистрибутором EMC. Мы не только поможем подобрать конфигурацию оборудования, наиболее подходящую именно к вашим задачам, но и привезём, установим и настроим купленное вами оборудование.