Аналитика на базе SAS HPA, EMC Greenplum и EMC Chorus

Обзор новых инициатив корпорации EMC для развития своего нового направления высокопроизводительной аналитики больших данных на базе интеграции с решениями компании SAS и нового инструментария для совместной работы аналитиков – EMC Chorus.

Денис Серов – руководитель направления технического консультирования, EMC Россия и СНГ.

Введение

В первой половине 2012 г. EMC сделала три важных анонса, связанных с продвижением своего направления высокопроизводительной аналитики для больших данных на базе решения EMC Greenplum UAP и СУБД Greenplum. Во-первых, была анонсирована (17 января 2012 г.) интеграция EMC Greenplum UAP с топовым решением SAS для высокопроизводительной аналитики – SAS HPA (High Performance Analytics – HPA), прежде всего, на базе технологий (интерфейсов) SAS In-Memory Analytics и SAS In-Database Analytics. Это стало возможным благодаря заключению в 2011 г. долгосрочного договора EMC с SAS о технологическом партнерстве.

Во-вторых, EMC объявила о выпуске Greenplum Chorus – платформы, аналогичной Facebook, для социального взаимодействия команд исследователей данных, позволяющей создавать наборы данных, обмениваться ими, обсуждать результаты и обеспечивать оперативное предоставление наиболее актуальных результатов анализа данных бизнес-пользователям. И все это – в рамках единой рабочей среды. Greenplum Chorus это не только "социальная", но и открытая платформа. Одновременно с ее анонсом EMC выступила с инициативой OpenChorus (openchorus.org), направленной на ускорение инновационного процесса и внедрение приложений для совместной работы с наборами данных на основе социальных сетей на платформе Greenplum Chorus. Исходный код Greenplum Chorus станет доступным по лицензии open source во втором полугодии 2012 г. Обновления проекта будут доступны на сайте openchorus.org.

В-третьих, EMC объявила о приобретении компании Pivotal Labs, одного из законодателей стандартов для динамичной разработки ПО с помощью современных методологий и сред программирования (таких, как Ruby on Rails и Pivotal Tracker). Этот инструментарий уже используется более чем 350 тысячами разработчиков в тысячах компаний. Приобретение Pivotal позволит EMC ускорить процесс подготовки и внедрения новых приложений для больших данных на предприятиях.

Интеграция решений SAS HPA с EMC Greenplum UAP

Интеграция флагманского решения для высокопроизводительной аналитики SAS с унифицированной аналитической платформой EMC Greenplum UAP позволяет пользователям SAS вывести существующую бизнес-аналитику с минимальными усилиями на качественно новый уровень, придавая бизнесу совершенно новую динамику развития. Пакет SAS HPA включает 3 базовые технологии: SAS In-Memory Analytics (аналитика в памяти), SAS In-Database Analytics (аналитика в СУБД) и SAS Grid Computing (распределенные вычисления). Интеграция осуществляется только с первыми двумя технологиями на базе двух соответствующих продуктов: SAS HPA for Greenplum и SAS Scoring Accelerator for Greenplum.

Необходимость в этих технологиях связана с тем, что традиционные программные и аппаратные средства бизнес-аналитики, включая корпоративные хранилища данных (EDW) и операционную инфраструктуру для данных, давно уже не в полной мере отвечают современным требованиям бизнес-анализа, требующих возможности высокой скорости обработки многотерабайтных и петабайтных объемов разного типа данных. Использование старых подходов приводит к тому, что:

- снижается производительность приложений для прогнозирования. В результате замедляется развитие бизнес-процессов, уменьшается число моделей и одновременно работающих приложений, которые могут быть созданы, протестированы и развернуты;
- снижается прикладная точность моделей. Поскольку с ростом объемов данных развитие и проведение тестирования замедляются, то уменьшается и число разрабатываемых моделей, а также учитываемых в них параметров, что снижает их инновационность, увеличивает их сегментированность и приводит к снижению точности предсказания за счет того, что формируются сверхобобщенные модели;
- снижается качество тестирования. Когда вычисления становятся проблемой, тестирование проводится в условиях жесткого лимита времени. Это приводит к неполному описанию модели, что, в свою очередь, может привести к непредсказуемым результатам;
- конфликт из-за ресурсов. Поскольку инфраструктура с разделением ресурсов деградирует с ростом объемов данных, то и борьба между группами аналитиков за их использование усиливается. Это снижает возможности сотрудничества и приводит к конфликтам и снижению общей результативности.

EMC и SAS выдвигают на рынок интегрированное решение для скоростной бизнес-аналитики, которое позволяет избежать перечисленных проблем и позиционируется для всех вертикальных рынков. Однако наиболее востребовано это решение будет там, где перечисленные проблемы в сочетании с большим объемом данных, проявляются наиболее выпукло:

- потребности пользователей BI-систем упираются в ограничения масштабирования объемов данных;
- пользователи BI-систем сталкиваются с задачами, которые они не могут решить, потому что они слишком сложны или тяжелы для выполнения на имеющейся BI-системе;
- потребители BI-систем нуждаются в предикативном анализе, приближенном к реальному времени;
- пользователи BI-систем нуждаются в отчетах, которые включают слишком много факторов/параметров, которые не могут быть включены в существующий процесс моделирования;

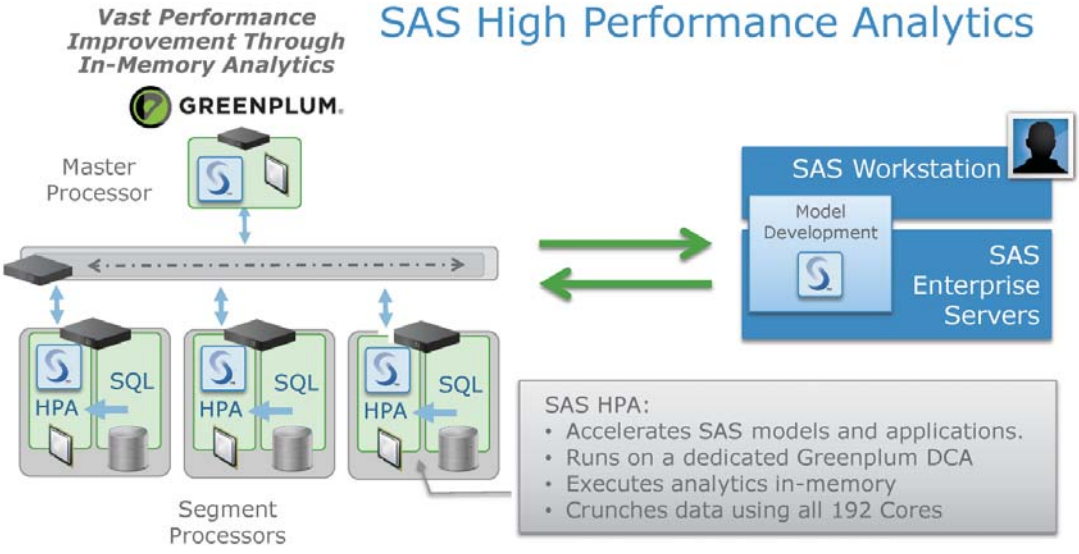

Рис. 1. SAS HPA прозрачно для SAS-приложений, выполняет те же самые функции, которые выполняются на SAS Server, но делает это внутри Greenplum DCA.

– предикативный анализ требует более гранулированного уровня данных, при переходе на который возникают ошибки/сбои;

– требуется массивная выборка переменных, что вызывает необходимость сортировки тысяч переменных для определения тех из них, которые оказывают наибольшее влияние на прогнозирование;

– пользователи BI-систем не хотят использовать неоптимальные методы моделирования;

– нет возможности быстро протестировать различные техники моделирования, чтобы найти наилучший способ повышения точности прогнозирования.

Практическими примерами использования данного решения для высокопроизводительной аналитики могут служить:

– снижение уровня ошибок кредитования и невозвратности кредитов;

– обнаружение и предотвращение мошенничества в близком к реальному времени;

– оптимизация розничных цен на уровне товаров в каждом магазине, чтобы идентифицировать самые лучшие и выгодные сегменты клиентов и генерировать соответствующие предложения.

Программные продукты SAS HPA for Greenplum и SAS Scoring Accelerator for Greenplum представляют собой набор аналитических и функциональных инструментов, которые охватывают исследование данных, моделирование ситуации/развития и модели развертывания.

SAS HPA комбинирует массивно-параллельное выполнение с обработкой in-memory, позволяя клиентам готовить, исследовать и моделировать множество сценариев, используя объемы данных, ранее никогда не доступные.

Способность быстро обрабатывать большие объемы данных означает обработку не бОльшего числа подмножеств, агрегатов/совокупностей или выборок, а, скорее, использование более полных детальных данных для генерирования очень точного и своевременного понимания процессов и принятия хорошо взвешенных решений.

Массивный параллелизм, обработка в оперативной памяти и устранение ненужного перемещения данных позволяют пользователям решать большие и сложные аналитические проблемы со сверхвысокими скоростями.

Интегрированное решение SAS&EMC не заменяет статистических испытаний, а, скорее, позволяет проводить исследование на более полном наборе данных и более “тонком” уровне детализации, например, осуществлять аналитику на уровне транзакций (операционном уровне) вместо аналитики на клиентском уровне или уровне счета. Оно делает возможным:

– ускорение понимания и разрешения чувствительных ко времени проблем;

– получение сложных аналитических подходов для решения наиболее острых (дающих наибольшую отдачу) проблем;

– более быструю и уверенную реакцию на новые возможности; более оперативное управление рисками; более оптимизированный выбор правильного решения в наиболее удобный момент.

Решение SAS&EMC улучшает вычислительные возможности, повышает объемы обрабатываемых данных и снижает время выполнения запросов для пользователей SAS, прежде всего за счет поддержки т.н. аналитики in-memory, реализуемой для SAS-приложений на базе интерфейса и технологии SAS In-Memory Analytics (рис. 1). В решении SAS&EMC вся in-memory обработка осуществляется на уровне EMC Greenplum UAP, т.е. самого хранилища данных, ускоряя обработку запросов на порядки за счет существенного улучшения трех параметров обработки:

– *вычисление*. SAS HPA выполняется на всех 192 ядрах Intel-процессоров со многими терабайтами ОП на 16 серверах DCA;

– *емкость*. Масштабирование традиционных SAS-окружений часто ограничивается не объемом данных, а сетевой производительностью, когда данные перемещаются в SAS-серверы из баз данных и файловых серверов. Использование SAS HPA снимает это ограничение, перемещая вычисления вплотную к данным: в сегменты DCA. Это дает возможность SAS-аналитику исследовать на много большие объемы данных без издержек на перемещение;

– *доступность*. В традиционной архитектуре SAS данные сначала загружаются в базу данных, а затем из баз данных извлекаются в серверы SAS. Это занимает время и снижает доступность данных в реальном времени. Greenplum DCA предлагает самую быструю загрузку данных – более 10 Тбайт/час/шкаф, что существенно приближает аналитику к реальному времени.

Использование модуля SAS Scoring Accelerator for Greenplum позволяет приблизить к реальному времени выполнение таких операций, созданных в рамках продукта SAS Enterprise Miner, как: регрессионный анализ, прогнозирование, скоринговые оценки и др. Ускорение достигается за счет применения специального кода, оптимизированного для параллельного выполнения на узлах EMC Greenplum UAP и устранения перемещения данных от хранилища к SAS-серверу (рис. 2).

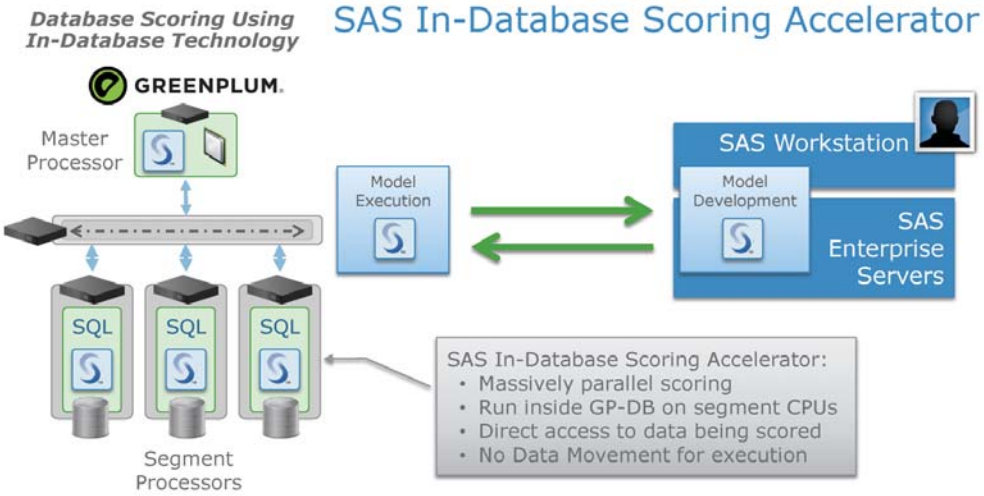

Рис. 2. SAS Scoring Accelerator устраняет необходимость перемещения данных из хранилища на уровень SAS-сервера и всю обработку запросов производит на уровне узлов Greenplum Database.

SAS HPA является новым флагманским решением, и оно будет востребовано не всеми пользователями SAS. Поэтому EMC будет продолжать поддерживать интеграцию Greenplum UAP, DCA и модулей DIA с другими продуктами SAS, включая:

– SAS Grid и SAS Enterprise: интеграция с UAP или с БД Greenplum обычно осуществляется через SAS Access;

– продукты для интеграции данных SAS, инсталлированные на модули DIA, в составе Greenplum DCA.

SAS HPA устанавливается в дополнение к полнорэковой модели Greenplum DCA с развернутой БД Greenplum. В результате получается законченное специализированное решение (appliance), готовое для использования с любым рабочим местом SAS или серверной платформой. Кроме того, решение может дополняться продуктами SAS или третьих фирм для интеграции данных, включая ETL-продукты от Informatica, Hadoop-кластерами, платформами данных (с файловыми системами и приложениями).

SAS и Greenplum совместно разработали SAS High-Performance Analytics for Greenplum как сайт-ориентированный апплайнс для аналитики. SAS High-Performance Analytics включает возможность выбора процедур от следующих продуктов SAS: Base SAS, SAS/STAT, SAS/ETS, SAS/OR и SAS Enterprise Miner. SAS 9.3 session также необходим для поддержки клиентского интерфейса. SAS-сессия (или клиент) управляет представлением запросов для их обработки и выполнения SAS HPA appliance. Полный список возможностей для формирования запросов можно получить на сайте SAS в разделе In-Memory Computing.

Миграция существующих SAS-проектов на SAS HPA for Greenplum потребует минимальных усилий, и полного переписывания каких-либо кодов не потребуется. Например, пользователь SAS, который исторически ориентировался на PROC LOGISTIC, может использовать HP-версию – PROC HPLOGISTIC для работы с апплайнсом SAS&EMC. Однако минимальные модификации могут потребоваться, т.к. не вся функциональность данной процедуры может поддерживаться на высокопроизводительной инфраструктуре.

Платформа для коллаборации аналитиков – EMC Greenplum Chorus

Первое решение Greenplum Chorus (Directly Available с 23 марта 2012 г. обеспечивает аналитическую высокопроизводительную платформу, которая позволяет группе аналитиков искать, исследовать, визуализировать и импортировать данные отовсюду, где бы они ни находились, в организации. Оно поддерживает широкие возможности социальных сетей, которые используются для наборов данных, понимания методов и бизнес-процессов, позволяя аналитикам данных, ИТ-персоналу, администраторам баз данных и другому персоналу сотрудничать в анализе больших данных.

Пользователи развертывают Chorus, чтобы создать гибкие самообслуживаемые

Рис. 3. Chorus обслуживает требования аналитики с итерационным циклом в четыре шага, основываясь на концепции совместной аналитики.

аналитические платформы, где команды аналитиков могут "на лету" создавать свою рабочую инфраструктуру с данными и немедленно проводить необходимые исследования и оценки.

Chorus обслуживает требования аналитики с итерационным циклом в четыре шага, основываясь на концепции совместной аналитики (рис. 3).

Chorus поддерживает прямую интеграцию с корпоративным LDAP или Microsoft Active Directory для управления пользовательскими паролями, устраняя потребность в дополнительном инструментарии и решениях.

Chorus автоматически назначает права по доступу к данным в соответствии с корпоративными политиками безопасности, и пользователи видят только те данные, к которым имеют доступ.

Chorus обеспечивает федеративный поиск во всех активах данных предприятия. Chorus индексирует все метаданные, комментарии, SQL и наборы данных для создания "живого" словаря данных, доступного в форме подсказки при поиске.

С помощью Chorus аналитики могут просмотреть и исследовать наборы данных всего предприятия. После того, как данные импортированы в рабочую область,

Табл. 1. Системные требования к рабочему месту Chorus.

COMPONENT	REQUIREMENTS
CPU	• Quad core Intel Pentium Pro compatible (P3/Athlon and above)
Memory	• 8 GB RAM
Storage	• 500 GB free storage
Operating System	• Red Hat Enterprise Linux 5.x • SuSE Linux Enterprise Server 10
Internet Browser	• Mozilla Firefox 8+ • Google Chrome 14+ • Internet Explorer 9+ or Internet Explorer 6+ with Google Chrome Frame installed
Other Software and Utilities	• Bash shell, GNU tar, GNU zip, GNU runtime, GNU sed, GNU awk • Oracle JDK 1.6.0_24
Greenplum Database	• Greenplum Database Version 4.1+
HDFS	• Greenplum HD Version 1.1+ • Greenplum MR Version 1.0 and 1.2 • Apache Hadoop Version 0.20.2 and 0.20.205

Chorus будет управлять потоком данных, их зависимостями и обновлять их копию, когда источник модифицирован новыми данными. Аналитики могут оперировать данными и совместно использовать результаты их анализа как новый набор данных (витрина данных), который может быть предоставлен для исследования другим пользователям.

Данное решение с открытым кодом ПО Chorus будет поставляться на рынок в коммерческом и некоммерческом вариантах. В последнем случае будет отсутствовать какая-либо поддержка со стороны EMC на развертывание и сопровождение, а также возможны ограничения по масштабированию инфраструктуры и томов.

Системные требования к рабочему месту Greenplum Chorus представлены в табл. 1.

Данный инновационный продукт с 23 марта 2012 г. доступен пока всего для 10 клиентов в Северной Америке в режиме Directly Available, а открытая доступность кода OpenChorus по лицензии open source запланирована на 2-ю половину 2012 г.

Заключение

В настоящее время платформа Greenplum UPA предоставляет широкие возможности для интеграции с имеющимися приложениями аналитики, существующими СУБД, хранилищами данных, а также инструментами для совместных аналитических исследований. Они включают (но не исчерпываются):

– *интерфейс доступа и встраиваемые функциональные библиотеки, которые в полной мере позволяют использовать вычислительный потенциал GP UPA на базе технологий обработки запросов in-memory и многопоточных вычислений. В первом случае ускорение производительности по сравнению с традиционными решениями может составлять тысячи раз, во втором случае – сотни раз;*

– *решения на базе продуктов компании Informatica по интеграции с наиболее популярными СУБД и КХД;*

– *непосредственную интеграцию GP UAP с hadoop-кластерами и скоростной СХД EMC Isilon для больших данных, которая также поддерживается соответствующими встроенными программными модулями, упрощающими доступ к таким ресурсам;*

– *платформу Chorus для проведения совместной аналитики.*

Новые возможности представляют большой интерес для компаний, выбирающих оптимальную платформу для работы с большими данными. Однако окончательный ответ можно получить, лишь проведя концептуальное тестирование в сочетании с техническими консультациями специалистов компаний EMC и поставщика аналитического программного обеспечения, например, компании SAS.

Денис Серов,
руководитель направления технического консультирования, EMC Россия и СНГ