HPA-аналитика и большие данные: особенности и применения

Обзор последних инноваций компании SAS в области развития технологий анализа больших данных, управления рисками, совместной аналитики с использованием информации из сторонних баз данных и источников, что позволяет "видеть" общую картину в разных разрезах и принимать более правильные решения, минимизируя издержки и риски, увеличивая прибыль, а также оптимизируя внутренние бизнес-процессы.

Алексей Мещеряков – руководитель направления платформенных решений, компания SAS Россия/СНГ.

Введение

В последние два года отрасль бизнес-анализа переживает целую революцию. За счет появления новых высокопроизводительных специализированных аппаратных платформ (для SAS-приложений – это возможность развертывания на базе решений IBM Netezza, IBM DB2, Cloudera, EMC Pivotal (ранее Greenplum), Teradata Aster, Teradata, Oracle Exadata, SAS SPDS) скорость обработки BA/BI-запросов повысилась в десятки и сотни раз. На смену классической ETL-интеграции (Extract, Transform, Load) идет федеративный доступ к данным, позволяющий "на лету" без дублирования данных в DW получать нужную информацию из первичных источников. Появление новых технологий (аппаратных и программных) дало возможность работы с огромными массивами данных (уровня петабайт) в режиме, приближенном к реальному времени. Все эти инновации стимулировали ряд тенденций в отрасли:

- *ключевым элементом управления бизнес-процессами и принятия решений становятся аналитические OLAP-системы с максимально жесткой интеграцией DW, СХД для OLTP-приложений и других источников данных;*
- *возможность включать в анализ данные из гораздо бОльшего числа источников (социальные сети, hadoop-кластеры и др.), чем только из операционных OLTP СУБД, и очень больших объемов;*
- *"демократизацию" инструментов BA и BI – возможность использования аналитики на всех уровнях бизнес-менеджмента (включая совместные исследования) за счет упрощения интерфейса доступа к данным (без привлечения ИТ-специалистов) и поддержки концепции BYOD;*
- *принятие тактических и стратегических решений стало гораздо более реактивным – максимально приближенным к реальному времени;*
- *упрощение развертывания приложений BA за счет предоставления услуг на базе высокопроизводительных аналитических платформ, размещаемых в облачных средах.*

Реализация этих тенденций, в свою очередь, в значительной степени позволяет повысить качество анализируемых данных/моделей, конкурентоспособность бизнеса и его прибыльность, одновременно снижая его издержки и риски, позволяя более точно осуществлять его прогнозирование развития на долгосрочный период.

Для компании SAS 2013 г. практически стал прорывным в области развития средств углубленной (advanced) аналитики. В новой версии HPA Server 9.4, помимо более высокой доступности и отказоустойчивости, теперь появились возможности его развертывания как локально, так и в облаке, а также использования его High Performance компонент отдельно, что сделало его более демократичным с точки зрения ценовой доступности. Существенно был расширен функционал HPA Server 9.4 – для работы с большими данными, поддержки отраслевой высокопроизводительной аналитики и др. Среди основных нововведений SAS в 2013 г. можно отметить следующие:

- *SAS HPA Server 9.4 (с новыми процедурами для High Performance Statistics, High Performance Econometrics, High Performance Optimization, High Performance Data Mining, High Performance Text Mining, High Performance Forecasting – прогнозирование в энергетике, прогнозирование спроса, прогнозирование в ритейле);*
- *SAS Enterprise Guide 6.1;*
- *SAS Visual Analytics 6.2/6.3;*
- *SAS Enterprise Miner 12.3/13.1;*
- *SAS Enterprise Decision Management;*
- *SAS Model Manager 12.3;*
- *SAS Contextual Analysis;*
- *SAS Scoring Accelerator for Hadoop;*
- *SAS Text Analytics (SAS Content Analytics с поддержкой русского языка);*
- *SAS® Data Management (дополненное функциями для работы с большими данными);*
- *SAS Event Stream Processing (ESP).*

По результатам исследования компании Forrester Research, Inc. "The Forrester Wave™: Big data predictive analytics solutions, Q1 2013" ("Прогнозная аналитика на больших данных") SAS лидирует по всем трем показателям (рис. 1) и это без учета состоявшихся анонсов в 2013 г.

В соответствии с ежегодным отчетом исследовательской группы IDC "Worldwide Business Analytics Software 2013–2017 Forecast and 2012 Vendor Shares" (июнь 2013 г.), посвященным определению позиций основных игроков рынка бизнес-аналитики по итогам 2012 г. и глобальным прогнозам на 2013–2017 гг., компания SAS, по итогам 2012 г., не только сохраняет лидерство на мировом рынке углубленной аналитики (рис. 1) с долей 36,2% (что почти на 1% превышает показатели предыдущего года и в 2 раза – долю ближайшего конкурента), но и наращивает ее. При этом доля SAS на рынке решений углубленной аналитики по-прежнему растет быстрее, чем сам сегмент – 36,2 % в 2012 г., 35,3% в 2011 г. и 34,9 % в 2010 г.

По данным того же исследования IDC, SAS входит в пятерку крупнейших мировых по-

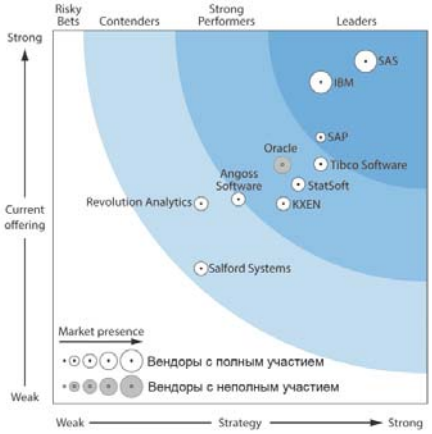

Рис. 1. Прогнозная аналитика на больших данных 1 квартал 2013 г.,Forrester Inc.

Рис. 2. Chartis RiskTech Quadrant™ for Data Management and BI for Risk.

ставщиков аналитических решений и лидирует в таких областях, как BI и аналитическая отчетность, текстовая аналитика, хранилища данных, управление цепочками поставок, сервисные операции, аналитический CRM и клиентская аналитика, финансовый анализ и стратегическое управление.

По итогам 2012 г., SAS вновь стала одним из лидеров "магических квадрантов" Gartner – в категории средств для интеграции данных и в категории средств для управления качеством данных. В списке, состоящем из более чем 40 западных вендоров, участвующих в "магическом квадранте" Gartner, SAS является единственной компанией, которая самостоятельно локализовала инструментарий и позволяет работать с русским языком.

Еще одна область, где SAS добилась больших успехов, это риск-менеджмент для банков. В соответствии с исследованием "Chartis RiskTech Quadrant™ for Data Management and BI for Risk" (август, 2013) компания SAS с решением SAS ESP (Event Stream Processing) занимает лидирующие позиции в области корпоративных решений для этого сектора рынка (рис. 2).

Третье направление, которое активно развивает SAS, – технологии работы с большими данными – семейство решений SAS® DataFlux® Data Management. Gartner позиционирует SAS в квадранте Лидеров в исследовании "Magic Quadrant for Data Integration Tools". Также Gartner позиционирует решения SAS/DataFlux в квадранте Лидеров в своем отчете "Magic Quadrant for Data Quality Tools".

Высокопроизводительная аналитика для риск-менеджмента

Решения для риск-менеджмента являются ключевым фокусом для SAS и поддерживаются такими глобальными инициативами SAS, как: IM & A, IMM (Integrated Marketing Management), Risk Intelligence. Более 50% всего оборота SAS составляют решения для финансового сектора, в которых неотъемлемыми компонентами являются технологии управления рисками, вследствие чего решения по риск-менеджменту для банков имеют особое значение.

Текущий – 2013 г. – в области риск-менеджмента был ознаменован появлением новых более жестких требований во всех сферах, включая банковскую, страховую и нефинансовую. В первую очередь это коснулось банков вследствие появления двух документов:

- "Principles for effective risk data aggregation & risk reporting" (BCBS239 – Jan 2013), *Basel Comittee on Banking Supervision;*
- "Principles for An Effective Risk Appetite Framework" (FSB – 17 July 2013), *Financial Stability Board.*

Принципы по снижению рисков, изложенные в этих документах, должны быть реализованы банками в период с 2013 по январь 2016 г. Всего этих принципов 14:

- *Governance and Infrastructure:*

  1 – Governance; 2 – Data Architecture and IT infrastructure;

- *Risk Data Aggregation capabilities:*

  3 – Accuracy and Integrity; 4 – Completeness; 5 – Timeliness; 6 – Adaptability;

- *Risk Reporting:*

  7 – Accuracy; 8 – Comprehensiveness; 9 – Clarity and Usefulness; 10 – Frequency; 11 – Distribution;

- *Supervisory review:*

  12 – Review; 13 – Remedial actions and supervisory measures; 14 – Home/host cooperation.

Обобщение этих принципов позволяет выявить следующие тенденции в области решений риск-менеджмента:

- необходимость в более детализированном сборе данных и управлении;
- потребность в поддержке задач, которым требуются большие вычислительные мощности, например, расширенное стресс-тестирование, детальное прогнозирование, детальный мониторинг и др.;
- специализацию требований для отчетов и вычислений.

Это, в свою очередь, порождает следующие требования к разрабатываемым программным и аппаратным аналитическим платформам:

- необходимость в более эффективном и мощном управлении данными;
- необходимость в повышении скорости вычислений;
- необходимость в поддержке специфических моделей и алгоритмов;
- потребность в эффективных пользовательских аналитических инструментальных средствах.

Среди новых и уже зарекомендовавших себя решений SAS для управления рисками можно отметить следующие:

- CPnM (Capital planning and management) in Banking & Insurance;
- SAS RMB (Risk Management for Banking) в части развития нового функционала (Liquidity, ALM – Asset and Liability Management, CVA – Credit Value Adjustment);
- SAS CRMS в части развития нового функционала (CRD IV – European Union Capital Requirements Directive, Regulatory reporting);
- SAS Event Stream Processing (ESP).

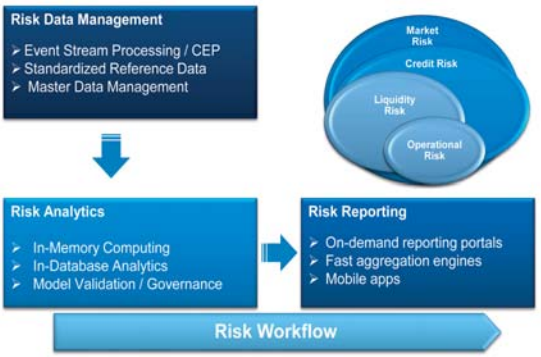

Рис. 3. Управление рисками (рыночными, кредитными, ликвидности, операционными) с использованием RDM-приложений, работающих на базе высокопроизводительных аналитических технологий SAS, поддерживаемых специализированными аппаратными платформами.

Решение SAS ESP – одно из последних, появившееся в сентябре 2013 г., представляет собой форму технологии комплексной обработки событий (complex event processing – CEP), которая ориентирована для критичных данных и приложений принятия решений. Это решение поддерживает встроенный ESP-механизм в составе других решений, когда необходимо интегрировать возможности принятия решения в реальном времени с платформой управления принятием решений или/и другими решениями SAS (рис. 3).

Существующие пользователи SAS могут легко интегрировать ESP-механизмы в другие SAS-решения, например, такие как: SAS Visual Analytics, SAS High-Performance Risk и SAS Fraud Detection. ESP-механизм может быть встроен внутрь решения или находиться "in front-end". Например, ESP-механизм, находясь "перед" SAS High-Performance Risk, может вычислять значения возможных рисков за счет использования его массивного параллелизма (in-memory). Полученные результаты передаются обратно в ESP-машину, которая транслирует их, например, приложению принятия решений, где они оцениваются, и уточненные исходные данные снова возвращаются в ESP-машину. Такой цикл может повторяться до получения требуемых результатов.

Возможности ESP-механизма поддерживаются такими высокопроизводительными технологиями SAS, как: SAS® High-Per-

Рис. 4. Возможности ESP-механизма реализуются тремя высокопроизводетельными технологиями SAS: SAS® High-Performance Analytics, SAS® High-Performance Solutions, SAS® Visual Analytics.

formance Analytics, SAS® High-Performance Solutions, SAS® Visual Analytics (рис. 4).

Потребность в решении SAS ESP может возникнуть в случаях, когда:

– существуют непрерывные запросы, генерируемые итерационными алгоритмами;
– имеются очень высокие требования к задержкам обрабатываемых событий;
– имеется высокий поток событий (>100K событий/сек).

Использование технологий SAS ESP дает следующие преимущества:

– *при агрегировании данных* – возможность обрабатывать части портфолио независимо друг от друга, агрегировать риски для корпоративной визуализации, измерять инкрементальные риски;
– *при сравнении* – одновременно проводить множество анализов, сравнивать риски повседневно;
– *при проведении стресс-тестирования* получать ряд результатов в зависимости от исходных данных.

Особенностью SAS ESP, как и всего семейства решений DataFlux, является поддержка интерфейса, ориентированного на бизнес-правила, с которым может работать нетехнический специалист.

HPA-аналитика и большие данные

В конце октября 2013 г. компания SAS существенно дополнила функционал своего ключевого решения для управления и анализа данных – SAS® Data Management (SDM) – функциями для работы с большими данными и, прежде всего, с Hadoop-кластерами. SDM расширяет полезные функции в среде Hadoop-кластеров с помощью технологий MapReduce, Hive, Pig и других, предоставляя такие основные возможности, как управление метаданными, преобразование и защита данных. При этом пользователям не требуется каких-либо специализированных навыков для работы с Hadoop-кластерами и все запросы к ним готовятся непосредственно из SAS-приложений с привычной семантикой.

Новые интерфейсы на основе ролей и объединенное управление данными в SDM значительно снижают нагрузку на ИТ-персонал по управлению данными, позволяя бизнес-пользователям самостоятельно проводить анализ и устраняя трудоемкие ручные процессы. Помимо этого, бизнес-пользователи становятся гораздо более эффективными, когда они имеют доступ к достоверной информации именно тогда, когда в ней нуждаются.

Второе важное расширение в SDM: новый сервер федерации – SAS Federation Server, с помощью которого пользователи могут объединять/комбинировать данные из множества источников для создания виртуальных бизнес-витрин информации, отчетов, генерирования ответов и обработки данных на месте, не перемещая или дублируя их.

История

Развитие технологий больших данных, в частности, на основе распределенных файловых кластерных систем началось с популяризации социальных сетей и электронной торговли. Это такие проекты, как: LinkedIn, Facebook, Digg, Google+, Amazon, Ebay, Yahoo и др. Вследствие недостаточной развитости первых поколений распределенных файловых систем, при реализации выше названных проектов предъявлялись пониженные требования к данным: допускалась потеря части данных, пониженная их целостность и др. Последующие версии (с 2011–2012 гг.) файловых систем, используемых для подобных целей имели уже гораздо бОльшую устойчивость *(см., например, публ. ""Файловые системы: от дисков к облакам" в SN № 2/54, 2013; прим. ред.).* Среди основных особенностей, которые привнесли эти проекты – это возможность хранить и анализировать в течение длительного времени огромное количество данных, измеряемое петабайтами. Наиболее заметный из реализованных проектов – проект Hadoop (2006 г.) с открытым кодом, разработанный на базе стандартных дешевых серверов и распределенной файловой системы HDFS, и допускающий масштабирование до сотен и тысяч узлов.

В течение нескольких лет возможности Hadoop-кластеров использовались исключительно для решения узкого класса задач, но с 2010 г. данные из Hadoop-кластеров стали активно интегрироваться с другими корпоративными данными для выявления и анализа более общих бизнес-процессов. Работать с Hadoop-кластерами можно, используя традиционные компоненты ИТ-архитектуры и методы доступа, однако это требует больших программистских усилий и, соответственно, временных издержек. При этом и производительность обработки данных также перестает удовлетворять необходимым требованиям. Так стали развиваться более специализированные технологии для работы с большими данными.

Многие факторы вносят вклад в увеличение объема данных. Примеры включают более новые корпоративные сценарии, которые основываются на транзакциях и аккумулируют во времени потоковые данные от/в социальных(е) медиа, или, например, данные, собираемые от многочисленных датчиков/видеокамер.

Данные, которые сегодня анализируются, могут представляться в различных формах. Это традиционные СУБД для OLTP- и OLAP-приложений, текстовые документы, почтовые сообщения, видео-/аудио-данные, логи, посты и др.

Однако потребность в использовании специализированных технологий больших данных может возникать уже сразу при решении ряда задач. Например, задачи типа моделирования и вычисления рисков, работающие с относительно малым объемом данных, могут генерировать объемы вычислений, которые могут выполняться в течение дней и более, существенно выходя за необходимое время принятия решения. Другой пример – бизнес-процессы могут требовать длительных ETL-процессов или процессов по преобразованию/подготовке данных для проведения анализа.

Технологии SAS для Hadoop

SAS предлагает ряд стратегий для развития и поддержки технологий больших данных:

Табл. 1. Технологии Hadoop и их назначение в Hadoop-инфраструктуре.

Hadoop Technology	Назначение
HDFS	Hadoop Distributed File System (HDFS) – распределенная, масштабируемая и переносимая файловая система написанная на Java для Hadoop framework. Пользователи загружают файлы в файловую систему, используя простые команды, а HDFS уже сама заботится о создании множества копий блоков данных и распределении их по множеству узлов в Hadoop-системе с целью поддержания параллельных операций, надежности и избыточности
Map-Reduce	Ключ к программированию и алгоритм обработки в Hadoop. Алгоритм делит работу на две ключевых стадии: Map и Reduce. Не все вычисления и анализ могут быть написаны в рамках MapReduce-подхода, но для анализа, который может быть проконвертирован, становится доступна обработка с высоким параллелизмом. MapReduce-программы пишутся на Java. Все другие языки, доступные в Hadoop, в конечном счете компилируют к программам MapReduce.
Pig	Pig Latin – процедурный язык программирования, доступный для Hadoop. Он обеспечивает способ реализации ETL-процедур и базовый анализ без написания MapReduce-программ. Он идеален для процессов, в которых последовательные шаги осуществляют операции над данными. "Пример Pig Latin программы: A = load 'passwd' using PigStorage(':'); B =foreach A generate $0 as id; dump B; store B into 'id.out';"
Hive	Hive – другой альтернативный язык для Hadoop. Hive – декларативный язык, очень похожий на SQL. Hive включает HiveQL (Hive Query Language – язык запросов Hive) для того, чтобы декларировать исходные таблицы, целевые таблицы, объединения (joins) и другие функции, подобные SQL, которые применяются к файлу или набору файлов, доступных в HDFS. Hive позволяет использовать структурные файлы, типа файлов с разделительными запятыми, которые могут быть определены как таблицы, к которым HiveQL может сделать запрос. Программирование на Hive подобно программированию для базы данных. Вот пример программы Hive:"INSERT OVERWRITE TABLE pages SELECT redirect_table.page_id, redirect_table.redirect_title, redirect_table.true_title, redirect_table.page_latest, raw_daily_stats_table.total_pageviews, raw_daily_stats_table.monthly_trend FROM redirect_table JOIN raw_daily_stats_table ON (redirect_table.redirect_title =raw_daily_stats_table.redirect_title);"

– решения для прозрачного использования hadoop-кластеров в составе аналитических приложений SAS;
– специализированные решения/платформы для анализа больших данных – SAS LASR и SAS High-Performance Analytics;
– имплементацию виртуализации данных – SAS Federation Server;
– решения по управлению обработкой больших данных – SAS Data Management.

В табл. 1 представлены технологии Hadoop и показано их назначение в Hadoop-инфраструктуре. SAS Data Integration Studio дает возможность использовать Hadoop следующими четырьмя способами (рис. 5):

– как файл-ориентированное внешнее хранение данных, использующее возможности SAS file I/O;
– как данные окружения/среды, использующие механизм SAS/ACCESS®;

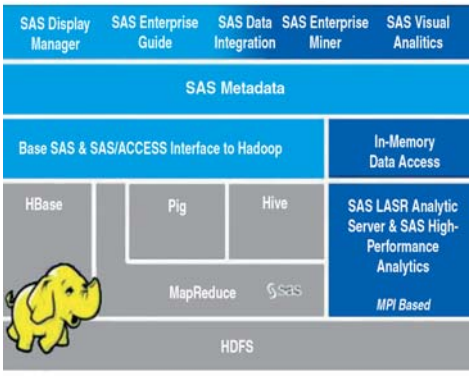

Рис. 5. Интеграция различных компонентов SAS и Hadoop *(Hadoop-технологии выделены серым, традиционные SAS-технологии – голубым, новейшие SAS-технологии – синим).*

– как вычислительную среду, используя Hadoop-запросы в SAS Data Integration Studio для Pig-, HIVE- и MapReduce-программирования;

– как данные для SAS LASR Analytic Server, используя конверсионные возможности к SAS LASR Hadoop storage.

Доступ к данным в HDFS при использовании SAS® File I/O

Приложения SAS могут получать доступ к данным, записанным в HDFS, несколькими способами. Первый использует файловый ввод/вывод, который дает возможность прямого обращения к данным HDFS. Его использование в комбинации с Hadoop добавляет ряд уникальных возможностей, которые не являются частью Hadoop-языка непосредственно, и усиливают возможности программам Hadoop:

– HDFS – распределенная файловая система, поэтому компоненты любого файла могут храниться на многих узлах Hadoop. Функционал SAS при работе с файлом дает полное абстрагирование от его физического размещения в HDFS и взаимодействует с ним как если бы это был объединенный/единый файл;

– HDFS не обеспечивает метаданными о структуре данных, сохраненных в HDFS. Используя методы доступа к данным SAS, можно применять форматы SAS и автоматически обнаружить структуру любых данных, содержавшихся в HDFS.

Доступ к данным в HDFS при использовании SAS/ACCESS® интерфейса к Hadoop

Эта техника позволяет взаимодействовать с файлами в HDFS, использует новый SAS/ACCESS механизм для Hive. Он обеспечивает т.н. libname-доступ к любым данным, сохраненным в Hadoop, используя SAS Metadata Server для обеспечения дополнительного контроля и управляемости ресурсов в Hadoop. При-

Рис. 6. Меню SAS Data Integration Studio выбора способа транформации данных в Hadoop.

Табл. 2. Способы трансформации данных в Hadoop и их описание.

Тип трансформации	Функция
Hadoop Container	Удобная трансформация, позволяющая множество программ Hadoop связать в одно преобразование.
Hadoop File Writer	Перемещение структурированного файла в локальной системе в файл в HDFS
Hadoop File Reader	Перемещение файла в HDFS в структурированный файл в локальной системе
Pig	Выбор из набора доступных шаблонов программ на языке Pig, которые помогают писать ETL-программы в Pig, и/или писать собственные Pig Latin, чтобы управлять и обрабатывать данные в Hadoop
Hive	Выбор из набора доступных шаблонов программ на языке Hive, которые помогают писать ETL-программы в Hive и/или писать собственный код Hive, чтобы делать запросы, фильтры или иначе обрабатывать данные в Hadoop, используя язык Hive
MapReduce	Выбор Java jar файла, содержащего программы MapReduce, которые будут представлены Hadoop системе
Transfer from Hadoop	Передача одного или более файлов из HDFS в локальную систему
Transfer to Hadoop	Передача одного или более файлов из локальной системы в HDFS

меняя Hive, SAS может обработать отделенные запятыми структурированные файлы как таблицы. И в дальнейшем с ними можно работать как со стандартными объектами при написании ETL в SAS Data Integration Studio или других пользовательских интерфейсах.

Вычисления в Hadoop

SAS Data Integration Studio обеспечивает ряд преобразований, показанных на рис. 6, которые полезны для работы с данными в Hadoop. Более детальное описание этих трансформаций представлено в табл. 2.

Аналитика на больших данных для "масс": SAS® Data Management, SAS® LASR™ и SAS® Visual Analytics

SAS LASR Analytic Server поддерживает в оперативной памяти (in-memory) распределенную вычислительную систему, подобную Hadoop. SAS LASR наиболее соответствует аналитическим алгоритмам, которые MapReduce-парадигма поддерживает не в полной мере. Доставку данных SAS LASR из Hadoop обеспечивает SAS Data Integration Studio, которая в значительной степени упрощает этот процесс. Для этого необходимо только зарегистрировать таблицы к SAS LASR, используя новый механизм SAS/ACCESS, и затем SAS Data Integration Studio может применяться для выполнения разнообразного набора задач с Hadoop-данными на SAS LASR, точно так же, как это было бы для других источников данных.

SAS LASR в настоящий момент (в декабре 2013 г. выйдет версия, которая позволит загружать в память данные в виде star-схемы, *прим. ред.*) не поддерживает объединения (joins) или pushdown-оптимизацию (магазинную), как это делается для других баз данных. Поэтому, если необходимо соединить или модифицировать данные, то это нужно сделать до загрузки данных в SAS LASR. Это можно сделать, используя стандартные процедуры SQL; или, если ваши данные находятся в Hadoop, необходимо непосредственно выполнить объединения в Hadoop. Можно создать объединения, используя один из способов трансформации данных SAS Data Integration Studio. Есть доступные примеры и шаблоны, упрощающие создание кода. Как только завершена стадия подготовки данных в Hadoop, можно конвертировать Hadoop-файлы или таблицы к формату SAS LASR, используя шаблон "SAS LASR Analytic Server Loader", доступный в библиотеке трансформаций. После того, как данный процесс разработан и протестирован, его можно поставить на регламентное выполнение, обеспечивая тем самым автоматизацию всего процесса доставки информации от источника до распределенной оперативной памяти.

SAS® LASR Analytic Server является ядром более высокоуровневого решения – SAS® Visual Analytics, которое появилось на рынке в 2012 г. и позволило поменять устоявшуюся годами парадигму – "BI и BA – инструмент принятия решений топ-менеджментом" и сделать инструментарий бизнес-анализа доступным для совместного использования широкому кругу потребителей в режиме, приближенном к реальному времени.

SAS® Visual Analytics предоставляет сотрудникам инструменты для самостоятельного исследования и визуализации данных. Благодаря им, не только аналитики, но и бизнес-пользователи смогут исследовать данные и проводить анализ с целью выявления закономерностей, наглядно представлять результат и делиться полученной информацией с коллегами, использующими как веб-интерфейс, так и мобильные устройства. Традиционные системы отчетности изначально реакционны: требуют знания данных, которые используются; предварительного понимания, какие дополнительные преобразования хотелось бы провести, и как представить результат.

SAS® Visual Analytics предоставляет различным группам пользователей инструменты для самостоятельного исследования и визуализации данных, рассматривать их характеристики, выявлять взаимосвязи, итерационно работая в реальном времени. Простые средства администрирования позволяют ИТ-специалистам создать единую рабочую среду для бизнес-пользователей, аналитиков и потребителей информации. ИТ-специалисты смогут загрузить и подготовить данные для всех пользователей, описать структуры данных в бизнес-терминах, а также единые правила их применения. Гибкое управление правами доступа позволяет ИТ-отделу обеспечивать безопасность и целостность данных, не влияя на производительность и не ограничивая возможности по исследованию данных. ИТ-отдел освобождается от необходимости обрабатывать потоки запросов пользователей на создание новых представлений данных или разовых отчетов, в результате его специалисты могут сосредоточиться на решении стратегических задач.

SAS® Visual Analytics позволяет:

– быстро и без лишних усилий анализировать любые объемы данных;

– выявлять взаимосвязи, которые раньше было невозможно обнаружить;

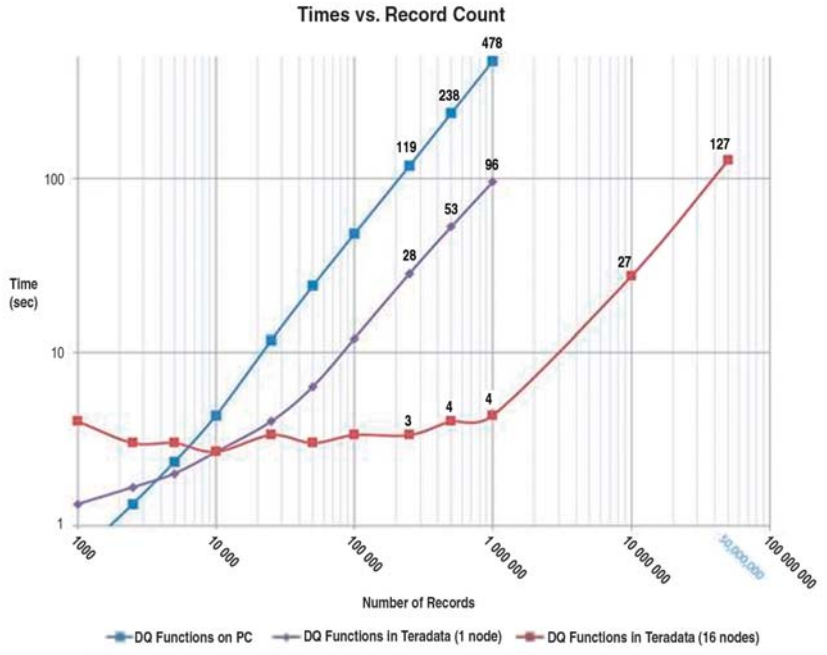

Рис. 7. Время выполнения DQ-процедур в зависимости от числа записей и платформы, на которой проводилось тестирование.

– понимать аналитические взаимосвязи (даже тем пользователям, которые не являются экспертами по анализу данных);

– определять направления дальнейшего анализа;

– использовать функции прогнозирования и сценарного анализа для предварительной оценки возможных вариантов развития событий;

– создавать интерактивные отчеты и информационные панели, которые после публикации будут доступны с различных устройств;

– просматривать отчеты в веб-браузере, на мобильных устройствах, в PDF.

Сочетание визуального интерфейса и высокопроизводительных технологий обработки данных сокращает время проведения аналитических исследований, позволяя организациям извлекать максимальную пользу из своих данных: на основе достоверной информации решить сложные проблемы, повысить производительность, найти новые способы повышения эффективности, увеличения дохода и оптимизации затрат.

SAS® In-Database Data Quality

Технологии SAS/ACCESS и продукты акселерации позволили перенести процесс обработки данных на уровень их хранения (включая базовые процедуры такие как: SORT, TABULATE и др. операции), что дало возможность существенно сократить общее время обработки запросов. Технологии SAS Scoring Accelerator и SAS Code Accelerator обеспечивают дополнительные преимущества, поддерживая встроенные процессы (SAS Embedded Process).

Помимо этого, на уровень баз данных SAS перенес технологии повышения качества данных. SAS Data Quality Accelerator for Teradata дает возможность выполнения следующих операций повышения качества данных на уровне хранения: Parsing, Extraction, Pattern Analysis, Identification Analysis, Gender Analysis, Standardization, Casing, Matching.

Пример улучшения производительности показан на рис. 7 (обе шкалы логарифмические). Из графика видно, что DQ-функции были выполнены на 50 млн записей чуть более, чем за две минуты на кластере Teradata с 16-ю узлами, в то время как PC был способен обработать приблизительно только 200 тыс. записей за то же самое время. Тестирование также показывает нецелесообразность использования специализированных DW при размерности таблиц < 10 тыс. записей. В то же время тестирование проиллюстрировало преимущество использования специализированных DW при больших таблицах. Так, для поддержания одного и того же времени обработки при изменении числа записей от 25 тыс. до 1 млн (в 40 раз) число узлов кластера Teradata было увеличено в 16 раз. При этом при использовании PC время обработки растет практически линейно с увеличением числа записей (см. рис. 7).

Необходимо заметить, что процедуры повышения качества данных крайне важны при управлении современными EDW, позволяя осуществлять:

– очистку “грязных” данных (очистка текстовых данных – грамматики, словарей, фонетики и др.; очистку числовых данных – статистика, выбросы, пропущенные значения; дедупликация данных и др.);

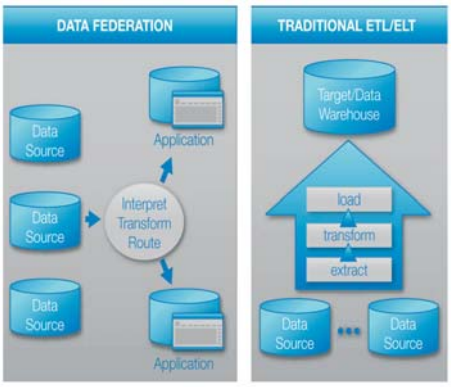

Рис. 8. Отличия между федерацией данных и традиционными ETL/ELT-методами.

– мониторинг и контроль качества данных (проверка Real-Time на этапе ввода в транзакционные системы);

– обогащение данных и аналитика (обогащение данных из внешних источников; определение пола, родственных связей и др.; интеграция с аналитическими решениями – SAS (Scoring, Anti-Money Laundering и др.).

Пример инсталляции. После реализации проекта по внедрению системы SAS Data Quality в банке Тинькофф Кредитные системы база контактов маркетинга, которая на старте проекта состояла из 16 млн клиентов, после очистки уже составляла 9 млн.

По результатам исследования Gartner “Magic Quadrant for Data Quality Tools”, 2012, SAS занимает первое место по представленности на рынке, а по стратегии развития входит в тройку лидеров.

Федерация данных и большие данные

Федерация данных это возможность интеграции данных, которая позволяет управлять множеством таблиц данных, как если бы они были единственной/одной таблицей, при сохранении их существующей автономии и целостности. Это отличается от традиционных ETL/ELT-методов, потому что при федерации из исходной системы перемещаются только необходимые данные (рис. 8). Именно поэтому она идеально подходит при работе с большими данными.

Сервер федерации (SAS Federation Server) является основой новых возможностей и

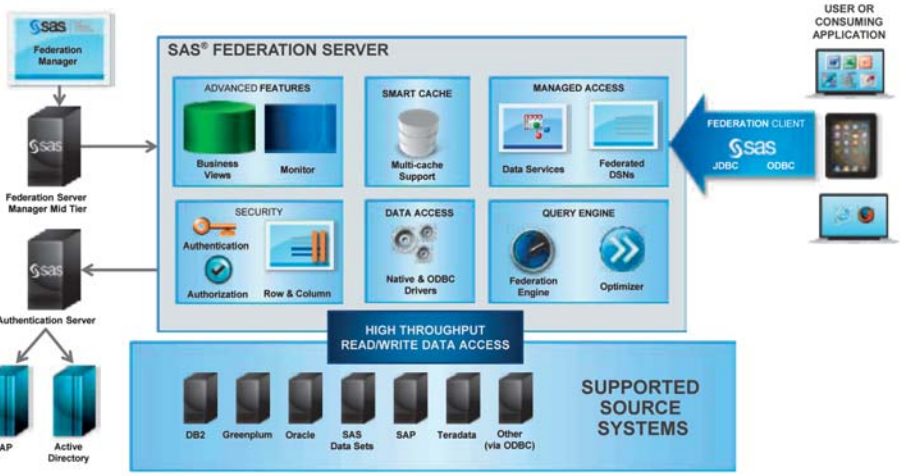

Рис. 9. Архитектура SAS Federation Server.

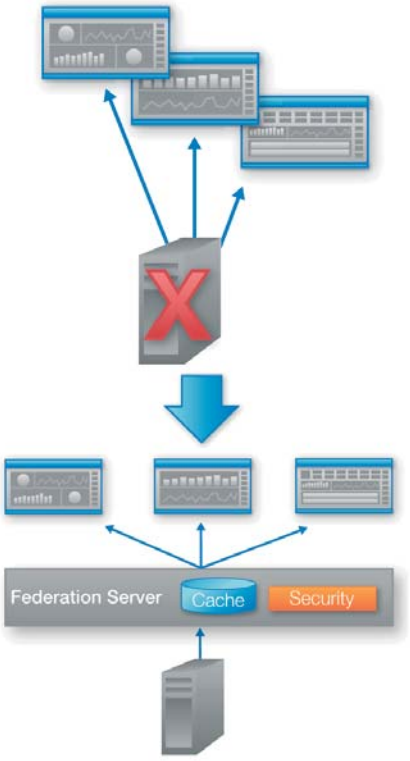

Рис. 10. Пример использования сервера федерации для защиты чувствительных данных.

позволяет комбинировать данные из множественных источников, управлять чувствительными данными через его особенности защиты и производительностью манипулирования данными через in-database оптимизацию и кэширование данных. Сервер имеет полностью связный ввод/вывод, поддержку push-down (магазинную) оптимизации, in-database кэширование, множество особенностей защиты (включая защиту уровня строки), интегрированный планировщик для того, чтобы управлять регенерацией кэша, множество "родных" механизмов доступа к данным для доступа к базе данных, полную поддержку SAS наборов данных, аудит и мониторинг возможностей/характеристик/SLA и другие особенности. Используя SAS Federation Server, можно получить централизованное управление всеми основными данными из множественных источников (рис. 9).

На рис. 10 представлено использование SAS Federation Server для защиты чувствительных данных. В этом сценарии данные являются собственностью организаций, которые не хотят предоставлять прямой доступ к их таблицам. И в каждом случае доступа к таблицам требуются свои политики доступа. Сервер федерации позволяет в полной мере решить эту задачу, кэшируя все запросы и данные, получаемые по этим запросам, а также устанавливая индивидуальные политики безопасности в каждом случае и полностью централизуя управление этими политиками.

Табл. 3. Потенциал выгод от использования Больших данных в государственном секторе Евросоюза: экономия от 150 до 300 млрд евро.

		Общая база, млрд. евро	Целевая выборка, %	Потенциал сокращения, %	Итого, млрд. евро
Повышение операт. эффективности	Операц. издержки	4 000	20-25	15-20	120-200
Сокращение ошибок и растрат	Платежи	2 500	1-3	30-40	7-30
Повышение сбора налогов	Налоговые сборы	5 400	5-10	20-10	25-110
					150-300 и более

Применения HPA-аналитики на больших данных

Согласно исследованию специалистов из MGI, европейский госсектор может сократить административные расходы на 15–20% (а это порядка 150–300 млрд евро) только за счет увеличения открытости данных и использования технологий продвинутой аналитики на больших данных (*ист.: McKinsey Global Institute, 2011,* табл. 3).

Примеры решения задач Big Data с высокопроизводительной аналитикой в России уже есть. Так, после 2-х месяцев с момента запуска пилотного проекта по внедрению SAS Visual Analytics в банке ВТБ24, на базе проведенного в данном решении анализа были построены портреты и определены профили клиентов, получены данные о распределении чистой прибыли и доходности по клиентским профилям, а также изучена специфика сегментов клиентов. При этом было обработано более 10 млн. строк – в 70 категориях и по 80 показателям. Работа с аналитической витриной по клиентским показателям велась в оперативном режиме.

Заключение

Появление новой версии HPA Server 9.4 в значительной степени сделала доступнее высокопроизводительную аналитику на больших данных, приближенную к реальному времени, более широким слоям бизнеса и госсектора, как по цене, так и по внедрению и использованию.

HPA Server 9.4 теперь включает в себя функции, которые позволяют дробить и параллельно обрабатывать аналитические задачи. Причем, как бы быстро ни росли объемы данных, эти высокопроизводительные алгоритмы будут автоматически масштабироваться для работы в распределенной среде. От пользователей и администраторов не потребуется никаких дополнительных действий.

За счет включения в визуальный анализ данных сценарного анализа и деревьев решений, бизнес-пользователи по клику мыши могут расширять возможности по отслеживанию изменений, анализу ситуации и выявлению трендов развития организации. Появились новые варианты получения отчетности из привычных систем SAS BI на мобильные устройства, в частности, на iPhone, iPad и планшеты на Android. Новые высокопроизводительные решения по оценке рисков могут существенно улучшить ситуацию в финансовом секторе.

Оперативный совместный анализ данных (с учетом данных, которые ранее были вообще не доступны для анализа вследствие их объема и невозможности их обработки в разумное время) позволяет активно воздействовать на множество бизнес-процессов, оптимизируя на всех уровнях управления, вовлекая в эти процессы тех менеджеров, от которых они зависят. А это обеспечивает повышение реактивности бизнеса на происходящие изменения, увеличивает его конкурентоспособность и прибыльность.

Алексей Мещеряков,
SAS Россия/СНГ.

Intel реализует преимущества HPC для анализа больших данных

Ноябрь 2013 г. – В рамках Supercomputing Conference (SC'13) Intel представила инновации в области высокопроизводительных вычислений и объявила о выпуске программных инструментов, которые помогут компаниям и исследователям в анализе неструктурированных данных.

Ускорение внедрения и развитие инноваций

Компания сотрудничает с партнерами для использования преимуществ технологий, доступных сегодня, и создания нового поколения высокоинтегрированных решений с более простым программированием и более высоким уровнем энергоэффективности. В рамках этого сотрудничества Intel также планирует предложить адаптированные продукты на основе высокопроизводительных вычислений для удовлетворения потребностей различных заказчиков. Эта инициатива призвана расширить преимущества масштабируемых платформ Intel, созданных на базе отраслевых стандартов, для включения дополнительных оптимизаций, которые лягут в основу новых прорывов в области науки, производства и образования.

В рамках конференции SC'13 представители корпорации рассказали о том, как новое поколение продукции Intel Xeon Phi (кодовое наименование Knights Landing), доступной в виде хост-процессоров, будет использоваться в стандартных стоечных архитектурах и позволит напрямую запускать приложения, а не переносить данные в сопроцессоры. Это поможет значительно снизить избыточную сложность программирования и отказаться от необходимости “разгрузки” данных, повышая производительность и уменьшая задержки, вызываемые подсистемой памяти, сетевыми и PCIe-соединениями.

Новая продукция также позволит разработчикам оптимизировать производительность готовых решений. В отличие от других эксафлопсных концепций, требующих от программистов создания специального кода для разных машин, новые процессоры Xeon Phi предлагают простоту и удобство использования стандартных моделей программирования.

Intel и Fujitsu также недавно объявили о новой инициативе, которая в будущем, возможно, позволит заменить электрические провода в компьютерах волоконно-оптическими каналами, по которым будет передаваться трафик Ethernet или PCI Express с помощью соединения Intel® Silicon Photonics. Это позволит устанавливать сопроцессоры Xeon Phi в специальный блок расширения, отделенный от хост-процессоров Xeon. При этом они будут работать так, как будто они установлены на системной плате. Такой подход даст возможность увеличить плотность размещения сопроцессоров и повысить емкость ресур-

(продолжение – стр.19)