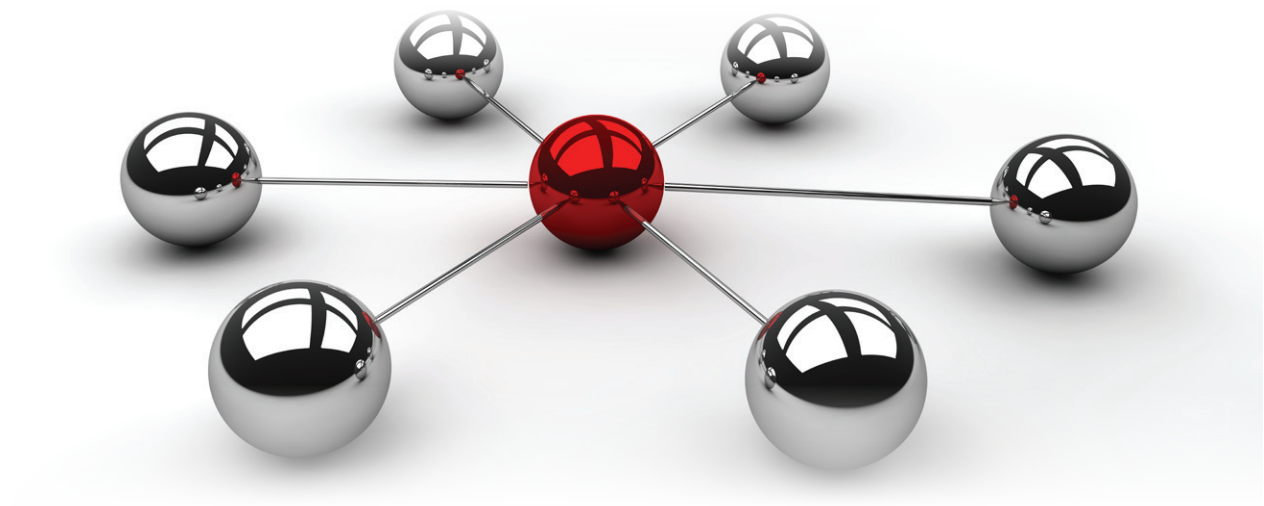
PREDICTIVE SOLUTIONS

Знакомьтесь —

IBM SPSS Statistics Server!

Преимущества клиент-серверной реализации Вашего статистического пакета

Довольно часто действующие или потенциальные клиенты задают нам такой вопрос: «Я слышал, что есть сервер IBM SPSS Statistics. Может быть, мне нужна серверная конфигурация?». Охотно рассказываем всем о преимуществах клиент-серверного варианта и помогаем принять решение.



Если коротко, IBM SPSS Statistics Server пригодится Вам, когда необходимо:

- Разгрузить рабочие станции сотрудников от рутинных вычислений;
- Повысить скорость доступа к корпоративным данным;
- Существенно повысить эффективность вычислений за счет использования специальных приемов, доступных только в серверной версии;
- Автоматизировать повторяющиеся задачи (например, построение регулярных отчетов в автоматическом режиме);
- Навести порядок с доступом к корпоративным источникам данных (организовать контролируемый и протоколируемый централизованный доступ к данным). Это относится как к данным, которые находятся в хранилищах (базах данных), так и к тем, которые хранятся в файлах, например, в формате Statistics *.sav.

В помощь желающим подробнее разобраться в вопросе, мы подготовили эту брошюру, в которой сравнивается архитектура и возможности обычной (настольной) и клиент-серверной реализаций статистического пакета IBM SPSS Statistics.

IBM SPSS Statistics Server: настольный вариант

Настольный вариант IBM SPSS Statistics знаком всем пользователям этого статистического пакета. Он идеально подходит для решения задач, связанных с анализом данных в рамках Ad hoc исследований или при подготовке специальных отчетов, когда все данные, нуждающиеся в обработке, легко могут быть собраны на одном персональном компьютере. Большое количество аналитических задач в сферах маркетинговых исследований, бизнес-анализа, а также в социологии, медицине, психологии имеют именно такую природу: сегодня Вы работаете с одним набором данных (одной анкетой, опросником, тестом), завтра – с другим набором, послезавтра — с третьим. Основной результат Вашей работы – аналитическое заключение, записка, статья или отчет. Последовательность анализа, результирующих таблиц и диаграмм уникальна и вряд ли потребует буквального повторения в будущем. Таким образом, все компоненты Вашей работы — данные, логика преобразований и анализа, сам процесс вычислений и результаты — находятся в Ваших руках, и пакет IBM SPSS Statistics, установленный на Вашем ноутбуке, позволит связать все это воедино.

Итак, в настольном варианте SPSS устанавливается непосредственно на рабочую станцию пользователя[1]. Чаще всего весь цикл работы с пакетом и замыкается в пределах рабочей станции: пользователь открывает локально сохраненный файл данных, обрабатывает его путем диалога с пакетом или с помощью синтаксиса с возможностью локального сохранения полученных данных (например, в формате *.spv или в «офисные» форматы). Вся вычислительная нагрузка по обработке ложится на рабочую станцию пользователя. Если задача оказывается вычислительно емкой, ресурсы Вашего компьютера могут оказаться ограниченными для других приложений.

В этой схеме есть свои исключения. Так, например, данным вовсе не обязательно располагаться непосредственно на рабочей станции пользователя:

- вполне подойдет сетевой диск, который распознается Вашим компьютером как локальный.
- если в операционной системе рабочей станции Вы настроите источник данных ODBC, то сможете работать и с данными в удаленно расположенной СУБД (в том числе выполнять как импорт, так и экспорт).

Работа с настольной конфигурацией IBM SPSS Statistics ничем не ограничивает востребованные новинки последних версий данного пакета. Так, например, Вы можете пользоваться программными расширениями, разработанными на Python или R, а также интегрировать программы на этих языках в командный синтаксис SPSS.

Неужели этого недостаточно? Действительно, в большинстве случаев настольная архитектура является абсолютно оправданной и не требует какого-либо масштабирования. Кому, в каких случаях, и чем может помочь клиент-серверный вариант этого замечательного статистического пакета?

[1] Работа с IBM SPSS Statistics на терминале, по сути, тоже является настольным вариантом использования этого пакета, для которого, однако, требуется установка сетевой (concurrent) лицензии на сервер терминалов. Concurrent-лицензия – не то же самое, что IBM SPSS Statistics Server.

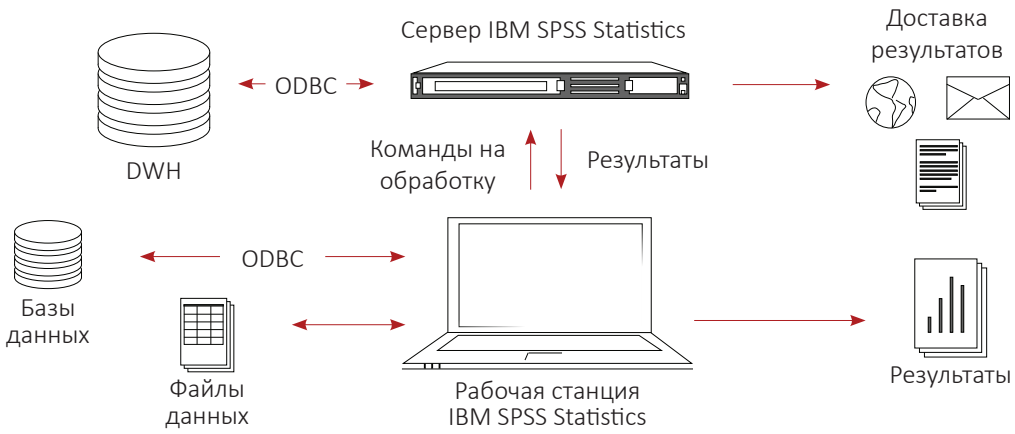
IBM SPSS Statistics + Server = IBM SPSS Statistics Server

IBM SPSS Statistics Server является организующим дополнением к уже существующим в организации настольным версиям IBM SPSS Statistics, централизованным вычислительным ядром, доступным для обслуживания аналитических процессов в режиме 24/7, и единой входной точкой к корпоративному хранилищу данных.

Что меняется с приходом сервера IBM SPSS Statistics в организацию?

В первую очередь, успокоим противников изменений: все задачи, которые выполнялись в локальном режиме, по-прежнему могут выполняться локально. Клиентские реализации SPSS, установленные на рабочих станциях, полностью сохранили свою функциональность, и, как и раньше, могут работать независимо от сервера.

Даже при подключении клиента к серверу пользователь по-прежнему видит знакомый Редактор данных SPSS, окно вывода (Output), меню и синтаксис. Однако внешний вид — это лишь вершина айсберга…



1. Роли клиентов и сервера

Как только Вы подключились к серверу IBM SPSS Statistics, Ваше локальное приложение («клиент») становится пользовательским интерфейсом для удаленно расположенной мощной вычислительной машины. Привычным образом запуская аналитические процедуры через меню или синтаксис, Вы лишь отправляете команды удаленному серверу, не загружая вычислениями собственный компьютер. Разумеется, теперь Вы заказываете серверу обработку не локально расположенных данных, а тех источников, которые либо находятся на сервере, либо тех, к которым сервер имеет доступ (например, в корпоративном хранилище). По завершении обработки сервер возвращает на Ваш компьютер результаты в виде таблиц (Pivot Tables), диаграмм, текстовых сообщений, либо экспортирует результаты обработки в те же удаленные источники данных. Один сервер способен независимо обслуживать запросы неограниченного количества клиентов[2].

Итак, на время расчетов высвобождаются ресурсы клиентских рабочих станций. Пока сервер выполняет длительное задание, пользователь может даже отключить свою машину от сервера, а затем вновь подключиться и забрать результаты. Здорово, но это — далеко не основное преимущество!

[2] Лицензия сервера не ограничивает количество подключаемых клиентов, однако, разумеется, сервер должен быть соответствующим образом масштабирован, чтобы его мощность соответствовала создаваемой клиентами нагрузке. Стоимость лицензии сервера, в свою очередь, привязывается к мощности сервера.

2. Скорость доступа к данным

Если Вы обрабатываете данные из централизованного источника (например, строите регулярные отчеты по продажам нескольких филиалов из корпоративного хранилища), то в первую очередь данные должны как-то поступить на обработку. В случае настольного варианта Вы почти всегда загружаете необходимую выборку данных на локальный компьютер. Когда источник данных и Ваша рабочая станция находятся не в одной локальной сети (LAN), этот этап становится узким местом всего процесса. С сервером все иначе. Почти всегда инструмент обработки данных располагают «рядом» с централизованным источником (в профессиональной терминологии — co-location), таким образом, что сервер всегда имеет подключение к источнику по быстрой локальной сети. Часто это оказывается в 20-40 раз быстрее, чем доступ к данным из другого сегмента сети, например, посредством интернета. Скорость же канала от Вашей рабочей станции до сервера Statistics, по большому счету, не имеет значения, т.к. между этими двумя узлами передаются очень небольшие фрагменты данных (для отображения на Вашем экране), синтаксис (команды на обработку) и результаты (таблицы и диаграммы с высоким уровнем агрегации представленных в них данных).

3. Автоматизация и планирование выполнения регламентных заданий на надежной платформе

В связи с тем, что сервер является круглосуточно активной службой, он гораздо лучше, чем пользовательская машина (ноутбук или настольный компьютер), подходит для выполнения автоматизированных процессов, особенно выполняющихся регламентно.

Уходя с работы на выходные, аналитик по привычке выключит свой ноутбук, позабыв, что в 03:00 в воскресенье планировщик операционной системы должен будет запустить процедуру сборки еженедельного отчета о продажах и запасах в разрезе филиалов. С сервером такого не произойдет.

Повышенная стабильность работы сервера, а значит, стабильность выполнения всех запланированных аналитических процессов, обеспечивается сразу несколькими составляющими, выгодно отличающими его от настольного компьютера:

- отказоустойчивость аппаратной части за счет использования оборудования серверного класса;
- высокая надежность серверной операционной системы, в которой устанавливается IBM SPSS Statistics Server[3];
- квалифицированное администрирование сервера и мониторинг, выполняемое специалистами ИТ по внутренним регламентам;
- управление нагрузкой сервера (включая балансировку нагрузки в кластере при необходимости).

При планировании выполнения заданий с помощью сервера аналитик может выбирать известное время минимальной загрузки сервера и базы данных, таким образом, чтобы вычислительно нагруженное задание выполнилось быстрее.

4. Повышенная защищенность данных

При работе через сервер персональные данные клиентов и другие защищаемые корпоративные данные из центрального хранилища обрабатываются на сервере, не поступая на локальный компьютер аналитика. Это обеспечивает дополнительную защиту данных, тогда как постоянная загрузка данных на локальный компьютер повышает вероятность того, что данные будут скомпрометированы.

Кроме того, поскольку сервер Statistics теперь является посредником для доступа к корпоративным источникам, соответствующие каналы (порты) локальных компьютеров пользователей теперь можно закрыть.

При подключении удаленных пользователей SPSS Statistics Server поддерживает шифрование с помощью Secure Sockets Layer (SSL) между клиентом и сервером, а также протоколы туннелирования и NAT.

6. Специальные средства повышения производительности

Серверная версия IBM SPSS Statistics обладает рядом дополнительных функций повышения производительности вычислений, недоступных в настольных версиях данного ПО. К основным отнесем:

[3] Релиз IBM SPSS Statistics Server 23.0 совместим с большинством распространенных серверных операционных систем: Windows Server 2008–2012 R2, Red Hat Enterprise Linux (RHEL) 5–7, SUSE Linux Enterprise Server (SLES) 11–12, Solaris 10–11, AIX 6.1–7.1.

- Наличие SQL push back при работе с базами данных: запрошенные трансформации частично или полностью будут выполнены прямо в базе данных за счет автоматической генерации соответствующих SQL-запросов сервером SPSS. Это радикально повышает скорость вычислений, поскольку исключается время на выгрузку/загрузку данных из/в СУБД.
- Поддержка предварительной компиляции синтаксиса трансформации данных, в результате чего расчеты выполняются быстрее.
- Отсутствие ограничений на количество процессоров/ядер, которые может задействовать многопоточный алгоритм обработки[4].
- Поддержка сжатия временных файлов, а также поддержка сохранения больших файлов прямо по ходу сортировки, что исключает дополнительный проход по данным.

7. Новые аналитические процедуры по сравнению с настольной версией

IBM SPSS Statistics Server содержит процедуры, характерные для работы с большими объемами данных (и, в частности, с большим количеством потенциальных предикторов) в проектах Data Mining. Пользователи клиент-серверной конфигурации могут использовать:

- Классификационный алгоритм Naïve Bayes;
- Алгоритм ускоренного отбора предикторов для прогностической модели (Predictor Selection).

8. Дальнейшая интеграция и масштабирование

С платформой хранения и автоматизации аналитических процессов IBM SPSS Collaboration & Deployment Services управлении процессами на IBM SPSS Statistics Server приобретает дополнительную гибкость. Даже сложные многоэтапные процессы теперь можно будет связывать в единые задания с помощью наглядного интерфейса, планировать и осуществлять мониторинг их выполнения непосредственно в платформе автоматизации. Станет возможна не только централизованная обработка данных, но и централизованное хранение аналитических процессов и результатов с поддержкой версий всех аналитических процессов, публикация результатов на портале, авторизованный доступ к ним разных категорий пользователей, уведомления администраторов, аналитиков и бизнес-пользователей по результатам выполнения заданий и многое другое.

В особенно нагруженных средах IBM SPSS Statistics Server может устанавливаться как вычислительный кластер, на несколько серверов, с балансировкой нагрузки между ними.

Некоторые факты

Выше мы много говорили о повышении производительности вычислений при использовании сервера IBM SPSS Statistics. Разумеется, дать универсальную численную оценку отмеченных выше преимуществ может быть непросто, так как производительность зависит от множества параметров (данные, мощность оборудования, специфика запросов и процедур, общая загруженность инфраструктуры и т.д.). Однако специально проведенные тесты дают некоторые индикативные оценки прироста скорости. Итак, например, что Вы можете ожидать от сервера IBM SPSS Statistics в сравнении с настольной реализацией:

- Процедуры преобразования данных (ADD FILES, AGGREGATE, MATCH FILES) выполняются в 6 раз быстрее;
- Процедуры сортировки — в 3,5 раза быстрее;
- Процедуры моделирования (регрессии, GLM, MIXED, NOMREG) — в 3 раза быстрее.

Тест на доступ к данным (последовательное чтение всех записей) файла размером в 50 Mb с удаленно расположенной настольной станции IBM SPSS Statistics по интернет-каналу (WAN со скорость 3 Mbit) показал результат около 2 минут. Та же самая операция, запрошенная через сервер Statistics, который находится в одной LAN с источником данных, заняла лишь 4 секунды. Разумеется, еще большая абсолютная экономия времени наблюдается при необходимости обрабатывать данные объемом в четверть гигабайта, гигабайт и т.д.

[4] Примеры процедур, поддерживающих многопоточную обработку в IBM SPSS Statistics 23.0: парные корреляции (Bivariate), линейная, порядковая, логистическая и мультиномиальная регрессии (Linear, Ordinal, Logistic, Multinomial), факторный анализ, регрессия Кокса, множественная импутация.

Чем IBM SPSS Statistics Server не является?

Распространено мнение, что сервер Statistics, фактически, является сетевым вариантом лицензирования пакета, исключающим необходимость приобретения отдельных рабочих мест. Это не так. Statistics Server является вычислительным ядром, который может функционировать отдельно от клиентских версий Statistics (например, запускать и выполнять по расписанию регламентные задания, как описано выше), но он не предоставляет интерфейса для работы клиентов. Если Вы хотите работать со Statistics Server в диалоговом режиме, Вам потребуется лицензия не только на Statistics Server, но и на клиентское рабочее место.

Вариант «сетевого» лицензирования Statistics (когда лицензия не привязана к физическому пользователю, а может использоваться разными пользователями с ограничением числа одновременно работающих пользователей) обеспечивается специальной лицензией Concurrent User License. Эта же лицензия используется для работы в SPSS на терминалах (иногда именно это и называют «сетевой лицензией»). В этом случае в сети присутствует так называемый «сервер лицензий» SPSS, который учитывает количество занятых и свободных лицензий рабочих мест SPSS Statistics, и выдает свободные лицензии во временное пользование. Но сервер лицензий и Statistics Server — не одно и то же.

10 ключевых улучшений, достигаемых с помощью IBM SPSS Statistics Server

Настольный вариант	Клиент-сервер
1. Ускорение обработки данных за счет использования ресурсов хранилища	
Все операции обработки после первичного запроса к БД выполняются на клиентской рабочей станции.	Часть ресурсоемких операций (сортировка, агрегирование) автоматически транслируется в SQL- код (push-back) для выполнения его внутри БД до передачи на сервер Statistics.
2. Скорость трансфера обрабатываемых данных из хранилища	
В вероятном случае нахождения рабочей станции IBM SPSS Statistics и сервера базы данных в разных сетях скорость передачи данных существенно ограничена каналом, связывающим эти сети.	Нахождение клиентской рабочей станции за пределами локальной сети сервера базы данных не влияет на скорость передачи данных для обработки за счет ко-локации сервера Statistics, осуществляющего обработку, и сервера базы данных.
3. Масштабируемость вычислений	
Все вычисления выполняются только на клиентской рабочей станции. Для увеличения производительности в масштабах организации требуется апгрейд всех рабочих станций с клиентами Statistics.	Вычисления выполняются на сервере (клиентская станция только отправляет инструкции и принимает результаты). Увеличение производительности в масштабах организации достигается апгрейдом одного сервера, в том числе переводом его в режим кластера.
4. Ресурсы рабочей станции	
В процессе любых расчетов ресурсы рабочей станции заняты текущими вычислениями.	При работе на сервере ресурсы рабочей станции свободны. Постоянное подключение к серверу на время расчетов не требуется. Есть возможность выполнять другую работу и запускать параллельные задания.

Настольный вариант	Клиент-сервер
5. Безопасность корпоративных данных	
Необходимо наличие доступа к данным по ODBC со всех рабочих станций, работающих с корпоративной БД. Доступ к клиенту Statistics не требует ввода пароля. Сохранить данные на неавторизованный носитель так же легко, как нажать *Файл…Сохранить как.*	Необходимо наличие доступа к данным по ODBC только для сервера. Клиент может находиться вне локальной сети организации и подключаться по VPN, с поддержкой SSL, PPTP, L2TP, NAT. Доступ к серверу – после ввода пароля. Доступ к файловой системе в процессе работы ограничен лишь файловой системой сервера и может быть разграничен средствами ОС.
6. Дополнительные статистические процедуры	
—	Процедура классификации Naïve Bayes и процедура автоматизированного отбора предикторов Predictor Selection доступны только пользователям клиент-серверной версии.
7. Администрирование	
Базовое журналирование событий и журнал синтаксиса.	Дополнительно: журналирование сессий пользователей, управление сессиями пользователей и процессами сервера, приоритезация пользователей, выделение индивидуального диска для хранения временных файлов.
8. Автоматизация	
Поддержка предварительно сформированных заданий (Productions jobs) на основе синтаксисов для регулярного запуска через планировщик.	Дополнительно: Прямой batch-режим с параметризацией запуска — возможность инициировать расчет на сервере без клиентского рабочего места Statistics. Повышенная надежность и доступность для выполнения процессов в производственном режиме за счет совместимости с серверными операционными системами и оборудованием.
9. Дополнительные средства повышения производительности вычислений	
—	Поддержка предварительно скомпилированных преобразований. Сохранение и сжатие больших массивов данных в процессе сортировки. Возможность использовать для вычислений более 4 параллельных потоков (multithreading). Асинхронное чтение данных для повышения скорости выполнения процедур.
10. Интеграция	
Интеграция с внешними системами посредством интеграционных плагинов, в том числе, для интерпретаторов Python, R.	Дополнительно: интеграция с системами масштаба предприятия посредством платформы IBM SPSS Collaboration & Services: репозиторий для хранения и планирования выполнения заданий, веб-портал, интеграция с data-mining платформой IBM SPSS Modeler Server.

Информационная брошюра подготовлена на основе собственного опыта компании Predictive Solutions во внедрении клиент-серверных решений на основе IBM SPSS Statistics, а также на основе информационных проспектов производителя ПО (IBM): Understanding the Benefits of IBM SPSS Statistics Server, IBM Corporation, 2010 (код публикации YTW03038USEN-00); IBM SPSS Statistics Server, IBM Corporation, 2013 (код публикации YTD03020-USEN-05).

ЗАО «Прогностические решения»
107031, г. Москва, ул. Кузнецкий Мост, 21/5
www.predictive.ru, info@predictive.ru
+7 (495) 229-56-11

PS