ОРГАНИЗАЦИЯ ОБЪЕДИНЕННЫХ НАЦИЙ **E**

ЭКОНОМИЧЕСКИЙ И СОЦИАЛЬНЫЙ СОВЕТ

Distr.
GENERAL

ECE/CES/2006/29
3 April 2006

RUSSIAN
Original: ENGLISH

ЕВРОПЕЙСКАЯ ЭКОНОМИЧЕСКАЯ КОМИССИЯ СТАТИСТИЧЕСКАЯ КОМИССИЯ

КОНФЕРЕНЦИЯ ЕВРОПЕЙСКИХ СТАТИСТИКОВ

Пятьдесят четвертая пленарная сессия
Париж, 13-15 июня 2006 года
Пункт 6 предварительной повестки дня

СЕМИНАР ПО ПЕРЕПИСЯМ НАСЕЛЕНИЯ И ЖИЛИЩНОГО ФОНДА
ЗАСЕДАНИЕ III

Системы хранения данных для целей распространения результатов переписей[1]

Документ представлен Национальным статистическим институтом Испании

I. РЕЗЮМЕ

1. Применение информационных систем, использующих новый технологический подход - Корпоративный интеллект, - может значительно повысить эффективность использования данных, собранных в ходе крупномасштабных статистических операций, таких, как переписи населения. Основным элементом таких систем является использование специальных структур, в которых информация хранится в виде, удобном не для обработки, а для поиска конкретных данных. Для этих целей создаются системы - хранилища данных, проектируемые по нестандартным моделям для оптимизации работы.

[1] Настоящий документ был подготовлен по просьбе секретариата.

2. Пользователи взаимодействуют с такой системой через простой, но мощный интерфейс. Сложные структуры данных остаются скрытыми от пользователей. Чтобы получить нужную информацию, пользователи должны лишь выбрать соответствующие переменные и задать необходимые условия. Используя инструментарий аналитической обработки данных в режиме реального времени (технология OLAP), пользователи могут находить информацию по любым интересующим их аспектам. В связи с применением таких систем необходимо также пересмотреть процессы обеспечения конфиденциальности статистики и редактирования данных.

II. ПЕРЕПИСЬ 2001 ГОДА: ТЕХНОЛОГИЧЕСКИЙ ВЫЗОВ

3. На протяжении всего цикла переписи Национальный статистический институт Испании (НСИ) активно использовал технологический компонент. В этой связи можно отметить следующее:

a) в переписные листы были заранее внесены личные данные участников переписи, взятые из Регистра учета населения; они были намеренно составлены таким образом, чтобы облегчить перевод информации в электронный вид с применением быстродействующей считывающей системы, и в результате обработка основной части информации была завершена в течение трех месяцев после сбора данных;

b) все население имело возможность заполнять переписные листы, используя Интернет (Испания стала первой в мире страной, обеспечившей такую возможность для всего населения страны).

4. При организации распространения результатов переписи пришлось принять во внимание следующие соображения:

a) основным каналом распространения информации будет Интернет. Могут также распространяться книги и электронные продукты, поскольку части пользователей эти продукты известны. Доступ к Интернету будет бесплатным и не будет требовать от клиентов соблюдения особых условий в отношении скорости передачи данных или используемой технологии;

b) в результате распространения информации как обычные пользователи, так и пользователи-специалисты должны получать максимально возможный объем информации;

c) из опыта предыдущей работы известно, что поступает много специальных запросов о предоставлении информации. Это создавало проблемы, поскольку увеличивалась рабочая нагрузка на технических специалистов. Такие запросы следует свести к минимуму и обслуживать на платной основе.

5. Эти требования не могли быть удовлетворены при традиционной системе распространения некоторого набора таблиц данных. Новый подход предполагает "систему самообслуживания", при которой пользователи могут составлять таблицы, исходя из своих собственных потребностей.

6. В этой связи характеристики системы распространения информации должны быть следующими:

a) "система самообслуживания" должна быть в состоянии дать ответ на большинство запросов пользователей в течение максимум 10 секунд, поскольку при более продолжительной задержке с ответом пользователи могут выйти из программы;

b) используемый интерфейс должен быть очень интуитивным и удобным для пользователя;

c) для пользователей, привыкших к традиционным методам распространения информации, одна из подсистем должна выдавать набор заранее заданных таблиц, содержащих основную информацию.

7. В связи с особыми характеристиками операции по проведению переписи (огромное число переменных, высокий уровень детализации географической информации) результаты переписи оказываются слишком сложными для размещения в традиционных реляционных базах данных. В контексте упомянутых требований и обстоятельств вызрело решение применить для создания эффективной системы технику многомерного моделирования данных и инструменты технологии OLAP. Теперь использование этих инструментов предусмотрено в концепции Корпоративного интеллекта. У НСИ не было опыта работы с такими системами, и поэтому их внедрение представляло собой трудную задачу.

III. ОЦЕНКА АЛЬТЕРНАТИВНЫХ РЕШЕНИЙ И ВЫБОР ТЕХНОЛОГИИ

8. Опыт использования таких информационных систем имеют главным образом субъекты частного сектора, где на постоянной основе анализируются различные аспекты

деловой среды. Характеристики таких систем отличаются от требований, предъявляемых к работе статистических учреждений, по крайней мере по трем следующим параметрам:

a) у этих систем имеются мощные механизмы обновления данных, поскольку они должны работать с весьма динамичной информацией (данные о продажах конкретного продукта, курсы акций) и быть постоянно готовыми дать управляющим информацию о текущем состоянии дел в компании;

b) у этих систем существует четко определенный набор коэффициентов и функций специальных докладов, и для анализа сохраняются лишь соответствующие этому набору аспекты информации;

c) пользователей информации можно обучить в конкретной технологической среде использованию систем и операционных данных.

9. С другой стороны, в процессе распространения статистических данных:

a) скорость обновления информации не является критическим фактором, поскольку статистические операции не дают новых результатов на непрерывной основе. В частности, переписи проводятся раз в десять лет;

b) приходится учитывать наличие множества типов пользователей и удовлетворять их различные интересы;

c) интерфейс, с которым имеет дело пользователь, должен быть интуитивным, поскольку обучить всех возможных пользователей применяемой системы нет возможности.

10. Так как необходимо обеспечить обслуживание всех пользователей, независимо от используемых ими технологий и без необходимости загрузки больших объемов программного обеспечения (таких, как апплеты, клиентские программы и т.д.), то имеющийся в распоряжении пользователя интерфейс должен быть приспособлен для работы именно в такой среде.

11. Таким образом, при выборе технических решений пришлось учитывать два следующих важных соображения:

a) время ответа на запросы: было решено, что задержка более чем в 5 секунд нежелательна;

b) наличие инструментов, позволяющих индивидуально регулировать структуру интерфейса, обеспечивая его интуитивность и широкие практические возможности.

12. Были выбраны технологические решения, предложенные Институтом SAS в виде компонентов Корпоративного интеллекта (КИ), на том основании, что:

a) в этих решениях программное обеспечение не привязано к конкретному аппаратному обеспечению и обеспечивает как решение всего спектра вопросов, так и доступ к разнообразным источникам данных;

b) эта технология известна НСИ, и значительная часть его сотрудников работают с инструментами SAS;

c) в ходе предварительных испытаний эта технология показала наилучшие результаты;

d) разработка и запуск в эксплуатацию системы хранения данных с использованием КИ представлялись легким и быстрым делом, поскольку имелся целый ряд хорошо разработанных инструментов.

13. До принятия решения по последнему из перечисленных пунктов было проведено испытание. НСИ обратился к различным компаниям с предложением представить свои идеи в отношении прототипа системы с пятью переменными. Через несколько дней консультанты SAS разработали полностью функциональную систему, работающую с высокой эффективностью.

14. Архитектура этой системы состоит из следующих основных элементов:

A. Программное обеспечение

15. Информационный сервер переменной мощности (ИСПМ): системная база данных, принадлежащая SAS. Она ориентирована на поиск ответов на запросы и способна выполнять параллельные операции и работать с расчлененными объектами. Эта система использует современные индексы "b-дерево".

16. Сервер OLAP: этот компонент системы управляет многомерными структурами в модели HOLAP (гибридная OLAP), т.е. он управляет не только таблицами данных, как в системе реляционной базы данных, но и гиперкубами или многомерными объектами.

Этот сервер обеспечивает простое логическое видение других процессов и перенаправляет запросы на минимальные уровни, где могут быть даны соответствующие ответы.

17. Администратор хранилища информации: инструмент для моделирования данных. Он позволяет определить все структуры в графической интуитивной среде и вырабатывает код SAS, необходимый для извлечения, преобразования и загрузки данных (процесс ETL).

18. Интеграционный пакет продуктов AppDev Studio: инструмент для разработки интерфейса. Он обеспечивает большое количество компонентных объектов Java, функционирующих в среде SAS.

19. Руководство для пользователей: инструмент, применяемый конечными пользователями для запроса информации и составления докладов.

B. Аппаратное обеспечение

20. Производственная среда: два компьютера Sun Fire 480 с операционной системой Solaris, объединенные в активно-пассивный кластер для обеспечения высокого уровня работоспособности. Эти компьютеры соединены с системой SAN - Symmetrics фирмы EMC, располагающей запасом памяти примерно 2TB для хранения детализированных и сводных объектов.

IV. ПРОЦЕСС МОДЕЛИРОВАНИЯ ДАННЫХ

21. Процесс моделирования данных начался с данных в плоских текстовых файлах, полученных в результате регистрации и редактирования первичных данных, собранных в период с ноября 2001 года по март 2002 года.

22. На основе этих данных был получен первый уровень агрегирования, на котором учитываются все имеющиеся комбинации изучаемых переменных и просчитываются различные параметры всех этих комбинаций. Этот уровень агрегирования соответствует так называемой N-таблице.

23. Поскольку объем файла очень большой и прямые запросы здесь затруднены, требуется выполнить агрегирование. Для выхода на следующие уровни агрегирования используются лишь отличающиеся друг от друга подмножества переменных.
В результате этого процесса получают группу "уменьшенных" объектов. Все они

объединены в логическую структуру, в рамках которой сервер OLAP может дать ответ на любой запрос, касающийся рассматриваемых переменных.

24. При моделировании данных важно определиться по следующим моментам:

a) способ агрегирования переменных: необходимо принять во внимание, каким образом соотносятся между собой переменные, о каких переменных информация будет запрашиваться чаще, существуют ли иерархические связи между переменными и т.д.;

b) разумное соотношение между уровнем агрегирования и используемыми ресурсами: более высокий уровень агрегирования позволяет более эффективно отвечать на запросы, но требует больше времени для загрузки или обновления структур хранилища данных, а также больший объем памяти для хранения информации.

25. При осуществлении проекта распространения данных переписи исходили из следующих основных посылок:

a) рассматриваются "естественные" иерархические структуры (например, возраст в группах - > возраст по годам, географические уровни);

b) для работы с переменными или пунктами переписного листа на разных уровнях детализации информации используются другие типы иерархической структуры (например, послевузовское образование, степень/звание, которое можно получить по окончании учебы, различные уровни экономической деятельности по Классификации видов экономической деятельности Европейского сообщества (КДЕС);

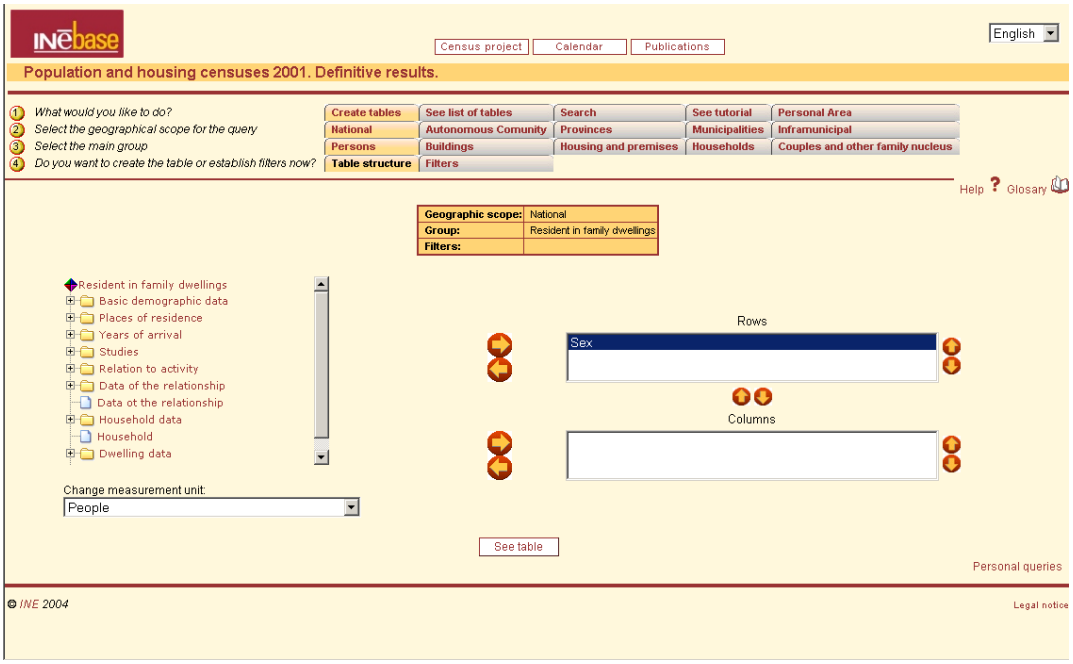
c) на этапе моделирования данных были сформированы тематические группы переменных по следующим двум соображениям: с учетом машинной мощности, необходимой для оценки всех возможных комбинаций всех переменных в случае рассмотрения их в качестве независимых переменных, и для того, чтобы обеспечить возможность совместного рассмотрения переменных, связанных по своему значению (например, обследование родителей: мать и отец рассматриваются в модели вместе).

26. При построении многомерной модели необходимо разработать прикладную программу, позволяющую использовать всю информацию, имеющуюся в данной модели.

V. РАЗРАБОТКА ИНТЕРФЕЙСА

27. Для того чтобы получить мощный, но простой в использовании инструмент, необходимо иметь в виду два обстоятельства:

a) во-первых, с помощью элементарных вопросов пользователю необходимо помочь вступить во взаимодействие с системой. Например, в связи с запросом: "Я хотел бы знать/размеры территории/к которой относятся данные по конкретной группе (люди, здания, единицы жилья)…" - перед пользователями открывается экран, на котором они могут выбрать одну или несколько переменных, относящихся к конкретным строкам или столбцам таблицы; еще один вспомогательный экран используется для определения фильтров (условий), которые необходимо применить при ответе на данный запрос;



b) при определении запроса наиболее важную функциональную роль играет экран, на который выводятся результаты. Во-первых, пользователи могут проверить, правильно ли понят их вопрос системой, поскольку на экране показаны все команды, сформулированные на предыдущих экранах. На экране имеется также панель инструментов, с помощью которой пользователи могут экспортировать результаты для представления их различных известных форматах, представлять выведенные на экран данные в графическом виде (демографические пирамиды, карты, гистограммы и т.д.), добавлять географическую переменную, изменять анализируемые переменные, возвращаться на этап задания параметров таблицы и т.д. В таблицах данных динамические элементы присутствуют во всех заголовках строк и столбцов.

По заголовкам пользователи могут ориентироваться в имеющейся информации. Выбирая курсором соответствующую категорию, они могут разукрупнять, расширять или укрупнять представленные в таблице параметры данных. В поисках информации о каком-либо представляющем интерес аспекте пользователи могут также изменять заданные переменные. Метаданные: чтобы лучше понять информацию, пользователи могут посмотреть в глоссарии значение какой-либо выведенной на экран переменной или категории переменных, а также в некоторых случаях, например в случае малых районов, запросить определение района или карту местности, для того чтобы получить представление о его местонахождении в городе или провинции;

c) ограничения - хотя система может справляться с очень сложными запросами, все же пришлось предусмотреть некоторые ограничения: в строках и столбцах таблицы должно быть не более трех встроенных переменных. Таблицу с бóльшим числом переменных трудно читать; возможности интерфейса ограничены показом таблиц, в которых насчитывается менее 10 000 ячеек. Таблицы большего размера представляются непрактичными, поскольку для их передачи и показа по Интернету требуется слишком продолжительное время. В таких случаях пользователям предлагается заполнить бланк заявления и загрузить результаты интересующего их запроса с FTP-сайта; ответы на некоторые вопросы не могут быть предоставлены в интерактивном режиме: чтобы защитить систему от выполнения слишком сложных расчетов, в тех случаях, когда для ответа на поставленный вопрос требуется слишком много времени, пользователям предлагается возможность доступа к файлу, который можно загрузить с FTP-сайта. Запросы, объем которых превышает 1 миллион ячеек, не принимаются к исполнению. Это правило установлено для защиты системы от слишком больших по объему запросов, с которыми может обратиться отдельный пользователь, с тем чтобы избежать злоупотреблений возможностями этого вида бесплатного обслуживания. Специальные запросы: когда пользователи не знают, каким образом сформулировать конкретный запрос, они могут изложить свой вопрос описательно. Эти запросы обрабатываются группой экспертов, использующих в качестве базового инструмента Enterprise Guide. Ограничения на информацию по соображениям конфиденциальности статистических данных: эта тема более подробно раскрывается ниже.

d) Другие функции: как упоминалось, рассматриваемая система может также предложить огромную подборку из более чем 10 000 таблиц данных, иерархически классифицированных по географическому признаку и по темам в соответствии с запросами традиционных пользователей нашей системы распространения информации. В режиме поиска можно легко найти отдельно взятую переменную или даже оговоренную категорию.

VI. КОНФИДЕНЦИАЛЬНОСТЬ ДАННЫХ

28. В отличие от традиционного распространения готовых к использованию таблиц, при котором содержание каждой таблицы можно проконтролировать с точки зрения конфиденциальности, рассматриваемая система позволяет объединять информацию по двум, трем и более переменным и делать повторные запросы, с помощью которых можно выявить характеристики отдельных объектов.

29. Поэтому был установлен ряд принципов, обеспечивающих защиту конфиденциальности данных. Эти принципы не предполагают очень жесткого толкования конфиденциальности, поскольку при таком подходе могут быть сильно сужены возможности распространения данных переписи по малым географическим районам, спрос на которые особенно высок.

30. Итак, применяются следующие принципы соблюдения конфиденциальности:

a) одномерные распределения можно распространять на любом уровне географической разбивки, поскольку они не раскрывают индивидуальной информации, за исключением тех случаев, когда заранее известны другие характеристики соответствующих объектов;

b) идентификация отдельных единиц обследования рассматривается как нарушение конфиденциальности только в том случае, если раскрывается точная индивидуальная информация. Предположения в отношении того, что некоторые опубликованные данные могут относиться к какому-то объекту, не рассматриваются как идентификация;

c) в распространяемой информации непременно присутствует элемент неопределенности в связи с тем, что некоторые данные отсутствуют, имеет место статистическое вменение и т.д.

31. Исходя из этих принципов, для обеспечения конфиденциальности данных при обработке запросов применяются следующие правила:

a) число затрагиваемых в запросе переменных зависит от размеров самой маленькой территории, которой этот запрос касается прямо или косвенно. Установлено четыре уровня численности населения, в отношении которых возможно представление информации по ограниченному количеству переменных:

Численность населения	**Максимально допустимое число переменных в одном запросе**
До 100 жителей	Одна переменная
От 101 до 5 000 жителей	Две переменные
От 5 001 до 20 000 жителей	Три переменные
Свыше 20 000 жителей	Без ограничений

b) По мере сужения географического охвата запросов производится группировка детальных значений некоторых переменных. Например:

Код занятий до двузначного уровня	Применительно к любой численности населения
Код занятий до трехзначного уровня	При численности населения свыше 20 000 человек

c) значения некоторых переменных ниже определенного порогового уровня не сообщаются. Например, страна гражданства указывается только на уровне районов, в которых проживает не менее 100 человек.

VII. ОПЫТ ИСПОЛЬЗОВАНИЯ СИСТЕМЫ РАСПРОСТРАНЕНИЯ ДАННЫХ

32. Обработано свыше 1 млн. запросов, включая как индивидуально сформулированные запросы, так и запросы готовых таблиц. С самого начала функционирования этой системы она была признана удобной для пользователей, и с течением времени большинство из них предпочитает формулировать свои собственные запросы, а не искать информацию в уже готовых таблицах. На приводимом ниже графике показано распределение количества запросов по месяцам.

33. Данные, касающиеся малых районов (подрайонов муниципальных образований), были опубликованы в декабре 2004 года, но только в Интернете. Поскольку выхода этой информации ожидали, система оказалась под большим давлением. Были созданы конкретные продукты, отвечающие потребностям пользователей, интенсивно работающих с такими данными.

34. Время ответа на запросы удалось выдержать в расчетном диапазоне: на 75% всех запросов ответы были даны менее чем за 4 секунды, а на 90% запросов - менее чем за 11 секунд.

VIII. ЗАДАЧИ НА БУДУЩЕЕ

35. В настоящее время ведется загрузка в рассматриваемую систему данных переписи 1991 года. Теперь можно будет получить значительно больше информации, чем было опубликовано до сих пор.

36. В НСИ в настоящее время ведется работа по включению временно́го измерения в имеющиеся данные, с тем чтобы можно было производить сравнение между данными различных переписей. С этим связаны некоторые трудности, поскольку для сравнения одних и тех же характеристик, обследованных в ходе различных переписей, необходимо найти общие метаданные. К маю 2006 года можно будет сравнивать данные переписей 1991 и 2001 годов.

37. Поскольку эта новая система распространения данных оказалась удачной, в данную платформу могут быть включены и другие статистические операции.

IX. УРОКИ, ИЗВЛЕЧЕННЫЕ ИЗ ЭТОГО ПРОЕКТА

38. Для продвижения проекта важно иметь группу энтузиастов, которые могли бы предусмотреть полный набор функциональных решений, отвечающих запросам пользователей, и имели надежную технологию, позволяющую добиться поставленных целей в пределах отведенного срока. В этом проекте принимали непосредственное участие три департамента: департамент переписи, а также департаменты распространения информации и компьютерной обработки данных.

39. Перепись является самым крупным проектом, который может осуществить управление статистики. Этот проект располагает колоссальным бюджетом, и в результате этого велико давление, направленное на получение быстрых и детальных результатов. Возможно, было бы лучше начать работу с более мелких проектов, однако большие

масштабы данной системы позволили исследовать многие технологические возможности систем КИ.

40. В рассматриваемой системе процессы редактирования данных имеют гораздо большее значение, чем в традиционной системе распространения статистической информации на основе таблиц. Требуется обеспечивать целостность и согласованность всех имеющихся данных, поскольку может поступить любой запрос. Когда же публикуется ограниченный набор таблиц, внимание можно концентрировать лишь на публикуемых данных.

41. Очень важно установить четкие правила, обеспечивающие защиту конфиденциальности статистической информации, поскольку существует много методов анализа такой информации.

* * * * *