

СЕТИ ДАННЫХ (DATA GRIDS) ДЛЯ ТЕРРИТОРИАЛЬНО-РАСПРЕДЕЛЕННЫХ СИСТЕМ АВТОМАТИЗАЦИИ КРУПНЫХ ОРГАНИЗАЦИЙ

В.В. Тутченко

Институт Программных систем, АСУ-ТП, ТОВ "Ер-Джи Дейта Украина"

e-mail: viktor@rgdata.com.ua

Введение.

Новое поколение высокоскоростных каналов связи позволило организациям создавать территориально-распределенные массивы данных, объединяя их посредством высокоскоростных сетей и Интернет-технологий в хранилища данных. Это явилось причиной появления новых понятий: большая группа пользователей, мультидоменная структура, географическое и территориально – распределенная структура, совокупность групп пользователей и ресурсов, координация, коллективная обработка. Множество этих понятий сформировало концепцию Сетей данных (Data Grids), обеспечивающей единую безопасность, распространенный доступ и сгруппированные возможности вышеназванных ресурсов [1]. Сети данных – вычислительная система построенная на основе глобальной сети, оперирующая с большим количеством распределенных данных (кластерами). Обычно Сети данных комбинируют с локальной вычислительной сетью. Таким образом, Сети данных [2][3][4] является сетями с доступом к распределенным ресурсам обработки данных.

Технология Сетей данных первоначально была использована как платформа для научной общности, которая нуждается в разделении ресурсов – процессорных мощностей и управление большими объемами данных, объединенных высокоскоростными глобальными сетями, высокопроизводительными процессорами и высокочастотными хранилищами данных, кластерами.

Основные фундаментальные принципы Сетей данных – обеспечение безопасности, высокопроизводительной протокола для передачи больших объемов данных расширяемый механизм репликации, обеспечивающий прирост и отказоустойчивость данных.

Для извлечения максимума преимуществ из инфраструктуры Сетей данных, необходимо организовать возможность поиска сквозь многочисленные доступные наборы данных; возможность находить соответствующий свободный вычислительный ресурс; возможность управления правами доступа собственниками ресурса. Таким образом, единая организация со строго определенными территориально распределенными ресурсами и архитектурой может получить выгоду от использования инфраструктуры Сетей данных.

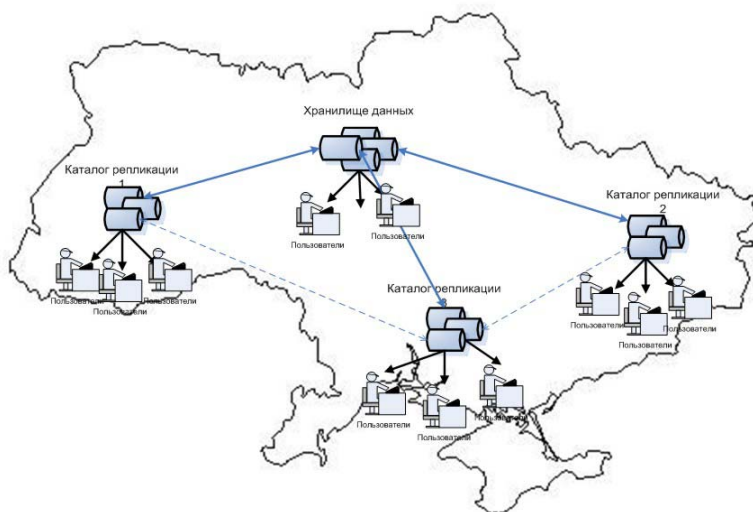
Рост популярности Сетей данных в научных и коммерческих сегментах привел к появлению разнообразных систем, предлагающих такие решения. К сожалению, это также привело к трудности в оценке этих решений из-за беспорядка в точном определении их точных целевых областей.

Кроме того, существуют несколько различных механизмов с подобными свойствами, поддерживающих распределенную инфраструктуру, требующую переработки большого объема данных или информации.

Главная цель этого доклада состоит в том, чтобы классифицировать Сети данных и построить их модели для территориально-распределенной системы автоматизации крупной организации.

Общее представление

На рисунке 1 показано представление высокоуровневой Сети данных, состоящей из хранилищ данных, которые связаны высокоскоростными сетями, охватывающими различные регионы Украины. Толстые линии показывают сети высоких полос пропускания, связывающие главные центры, более тонкие линии



– сети более низких полос пропускания, которые соединяют конечных пользователей со вспомогательными центрами.

Данные реплицируются из региональных центров посредством запроса синхронизации. Изменения поступают на центральное хранилище данных, откуда механизм ответа управляет копиями других участкам хранения во всех регионах.

Механизм ответа управляется в соответствии со стратегиями ответа, которые учитывают текущие и будущие требования на наборы данных, местоположение запросов и емкости запоминающего устройства, для создания копий данных. Пользователи в соответствующем регионе соединяются с их региональным центром, для получения требуемых данных, в соответствии с предоставленным правом доступа. Если эти данные не присутствуют, в их региональном центре, запрос формирует результат от центрального хранилища данных, либо от другого регионального центра с менее загруженным ресурсом.

Другой аспект Сетей данных должен поддерживать разделение собранных данных, в разрезе региональных областей. Эти центры должны работать с данными также независимо от основных систем хранения, т.е. поддерживать off-line работоспособность, что налагает требования к использованию устройств резервирования информации [5].

Присутствие большого количества пользователей, находящихся в различных региональных областях и использующих общие ресурсы формирует следующие особенности:

- Быстрый прирост данных: объемы данных могут превышать терабайты.
- Географическое распределение: Данные должны быть доступны всем пользователям. Для этого требуется инфраструктура организации, охватывающей различные регионы.
- Единственный Источник: Это означает, что все данные, внесенные в одном регионе, затем распространялись на другие. Таким образом,
- Объединенное пространство имен: данные в Сетях данных используют то же самое логическое пространство имен, как и в уникальных именах файлов.
- Ограниченные ресурсы: система может иметь ограничение по вычислительным, либо по полосе пропускания или другим ресурсам. При проектировании надо гарантировать эффективное разделение ресурсов для каждого региона.
- Местная автономия: В то время как различные региональные центры вносят данные в систему, они должны сохранять контроль над своими данными. В частности, только они должны решать, кто может получить доступ к данным регионального уровня.
- Ограничения доступа: В пределах Сетей данных должны существовать различные уровни доступа среди пользователей.
- Разнородность: окружающие среды Сети данных охватывают различные аппаратные средства и конфигурации программного обеспечения, которые потенциально используют различные протоколы. Система должна быть спроектирована для работы независимо от различных интерфейсов.

Модель Data Grids для территориально-распределенной организации.

Рассмотрим модели представления Сетей данных территориально-распределенной организации: централизованную (Рис.2) и децентрализованную (Рис.3).

Так же, как классические Сети данных для научных организаций, состоящие из нескольких уровней, так и для нашей территориально-распределенной системы электронного документооборота Фонда государственного имущества Украины, построим уровни для центрального хранилища документов – сервера центрального аппарата и Департаментов, распределим до уровня 27 региональных отделений Фонда (региональные центры в каждой области Украины), далее на уровне районов в каждой из областей. Такая централизованная модель представлена на рисунке 2. Верхним уровнем является центральное хранилище документов – центральный аппарат Фонда. Данные передаются снизу вверх от районных отделений к областным, а далее в центральный аппарат Фонда. Узким местом такой модели является полоса пропускания каналов связи.

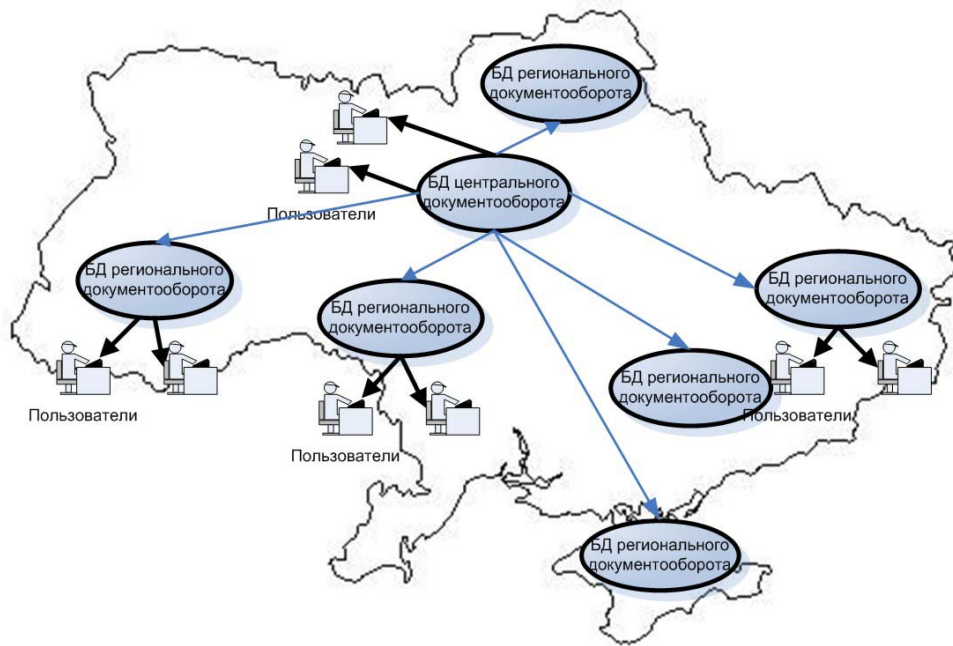


Рисунок 2. Централизованная модель Сетей данных организации

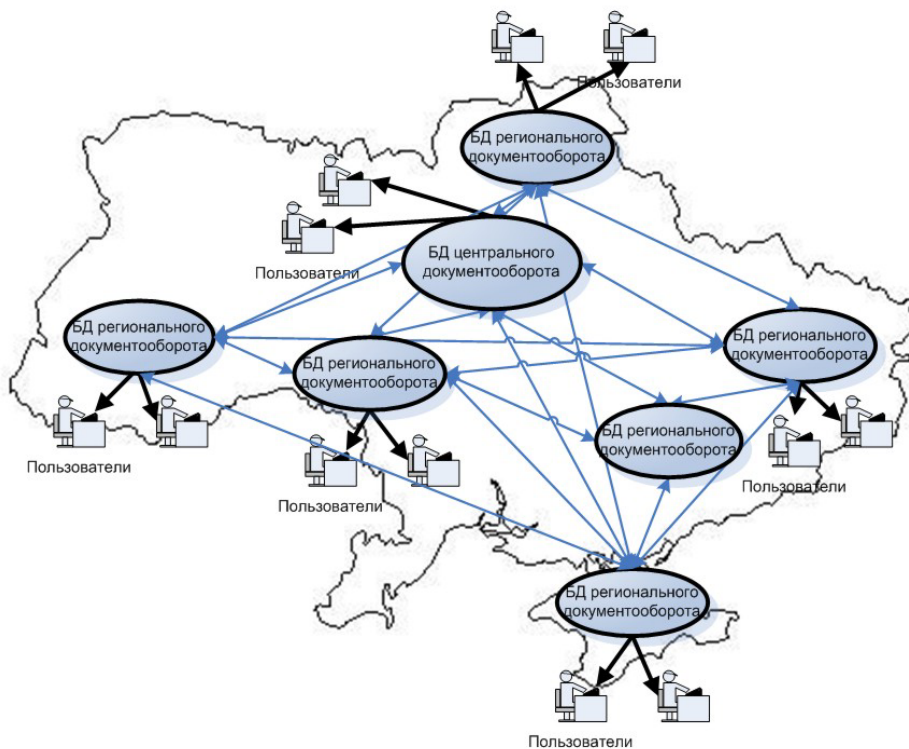


Рисунок 3. Децентрализованная модель Сетей данных организации

ФГИУ состоит из центрального аппарата, 27 областных отделений и более 100 представительств в районах.

Оценка потоков и объема электронного документооборота за 2006 год ФГИУ представлена в таблице.

Категория документов	Центральный аппарат ФГИУ	Региональное отделение ФГИУ по г. Киеву
Входящие документы	26 482	10 823
Исходящие документы	19 492	10 668
Распорядительные документы	2 594	--
Внутренние документы	1 643	--
Объем БД всех документов за 2000 – 2006 г. включительно.	299 391 16 180 МБ	124 925 5 520 МБ
Процентное отношение сканированных электронных документов к общему кол-ву документов в базе	18,3%	6,7%

Децентрализованная модель [6] территориально-распределенной организации показана на рисунке 3. Такая модель характерна тем, что пользователь, посылая запрос из любого районного отделения, могут получить данные не только из центральной базы данных документов, но из хранилищ данных других районных отделений даже своего уровня, если они будут иметь соответствующие права доступа. Каждое региональное и районное отделение управляет своей базой данных и определяет свои ограничения на данные.

Существуют также две другие модели – сенсорная и гибридная [4].

Сенсорная (или восходящая снизу) характерна тем, что данные собраны на центральной базе данных и подача информации идет от низших уровней наверх.

Гибридная модель совмещает вышеназванные модели.

Заключение

Предложенные модели Сетей данных могут быть прототипами для построения в недалеком будущем территориально-распределенных систем автоматизации крупных организаций. Классификация, рассмотренная в этой работе, основана на публикациях, содержащих основные расчеты стоимостных, временных характеристик, вычислительных мощностей, распределение ресурсов и управление данными и кластерами [2], [4], [5]. За пределами данного доклада осталось рассмотрение таких важных аспектов, как транспортная система, системы отказоустойчивости и резервирования данных, временных задержек, расчет объемов данных для представленных моделей территориально-распределенной организации, что может послужить продолжением работ по этой тематики в дальнейшем.

Литература

1. I. Foster and C. Kesselman, The Grid: Blueprint for a Future Computing Infrastructure. Morgan Kaufmann Publishers, 1999.
2. A. Chervenak, I. Foster, C. Kesselman, C. Salisbury, and S. Tuecke, "The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets," Journal of Network and Computer Applications, vol. 23, no. 3, pp. 187–200, 2000.
3. P. Avery, "Data grids: a new computational infrastructure for data- intensive science," Philosophical Transactions of the Royal Society of London Series a-Mathematical Physical and Engineering Sciences, vol. 360, no. 1795, pp. 1191–1209, 2002.
4. W. Hoschek, F. J. Jaen-Martinez, A. Samar, H. Stockinger, and K. Stockinger, "Data management in an international data grid project," in Proceedings of the First IEEE/ACM International Workshop on Grid Computing (GRID '00). Bangalore, India: Springer-Verlag, Berlin, December 2000.
5. R. Moore, A. Rajasekar, and M. Wan, "Data Grids, Digital Libraries and Persistent Archives: An Integrated Approach to Publishing, Sharing and Archiving Datas," Proceedings of the IEEE (Special Issue on Grid Computing), vol. 93, no. 3, 2005.
6. A. Rajasekar, M. Wan, R. Moore, and W. Schroeder, "Data Grid Federation," in Proceedings of the 11th International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2004). Las Vegas, USA: CSREA Press, June 2004.