УДК 681.3.069

© Меньшикова Л.В.
Menshikova L.

НЕКОТОРЫЕ ПОДХОДЫ К РАЗРАБОТКЕ АРХИТЕКТУРЫ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКИХ СИСТЕМ

SOME APPROACHES TO THE DEVELOPMENT OF AN INFORMATION-ANALYTICAL SYSTEMS ARCHITECTURE

Аннотация. *Рассмотрена типовая архитектура современной информационно-аналитической системы. Даны определения ее составных частей. Рассмотрены подходы к проектированию современной информационно-аналитической системы, а также мероприятия, которые позволят реализовать эти подходы.*

Annotation. *The typical architecture of modern information-analytical system is considered. Definitions of its components are given. Approaches to designing modern information-analytical system, and as actions which will allow to realize these approaches are considered.*

Ключевые слова. *Информационно-аналитические системы, проекты информатизации крупномасштабных предприятий, разработка информационно-аналитических систем.*

Key words. *BI-systems, IT-projects of the large-scale companies, information and analytical systems development.*

Основная задача, решаемая при планировании новой информационно-аналитической системы, в которой планируется реализовать более одного алгоритма и в разных алгоритмах которой используются одинаковые данные в качестве не только входных и выходных, это провести синтез этих алгоритмов, то есть максимально все упростить, ничего не потеряв.

Бизнес-аналитика (англ. Business intelligence, сокр. BI) – это инструменты, используемые для преобразования, хранения, анализа, моделирования, доставки и трассировки информации в ходе работы над задачами, связанными с принятием решений на основе фактических данных. Идея BI и само название были предложены аналитиками Гарднер в конце 80-х.

При построении систем бизнес-аналитики возникает ряд проблем:

1. Обеспечение согласованности новых измерений.

2. Формирование непротиворечивого пространства данных, которое необходимо корректно не только сформировать в момент инициализации хранилища данных, но и корректно вводить данные из новых задач.

3. Недостаточно эффективное партнерство между представителями департаментов информатизации и бизнеса.

4. Отсутствие ясного понимания бизнес-пользователями, что именно они хотят получить от создаваемой системы. Всех пользователей принято делить на две категории: 1-я группа – те, кто могут создать Excel – таблицу с данными самостоятельно, и 2-я группа – те, кто этого сделать не могут, а таких по статистике 80 %. Для второй группы пользователей имеет значение степень сложности получения требуемого отчета.

5. Задержка получения данных, не устраивающая бизнес-пользователей. Здесь единственный путь – менять архитектуру с учетом наличия кризиса реального времени, а также напрямую получать данные из источника.

6. Несогласованные измерения и факты, присутствующие в информационной среде предприятия. Интеграция и дедубликация данных – это самый сложный процесс.

7. Недостаточная детальность данных, объем которых постоянно растет, причем бизнес постоянно стремиться добавлять все новую и новую деструктурированную информацию. Здесь выход – как можно больше де-

Меньшикова Лариса Валерьевна – кандидат физико-математических наук, доцент, экономический советник Департамента информационных систем Банка России, МИРЭА, тел./факс: +7(495)753-94-20.

Menshikova Larisa – candidate of physico-mathematical science, associate professor, economical adviser of Bank of Russia, MIREA, tel./fax: +7(495)753-94-20.

скрипторов для сложных понятий.

8. Неприемлемые форматы данных не только в части представления, что важно только для конечного пользователя, но и в части необходимой для расчета точности данных.

9. Медленная доставка данных. Здесь время должно быть меньше пяти секунд, иначе конечный пользователь будет избегать таких запросов.

10. Низкое качество данных. Управлять качеством данных можно при помощи построения схемы событий с ошибками, а также путем сверки данных по предыдущим измерениям, которые можно считать для них аудитом.

11. Хранение агрегированных данных вместо атомарных или вместе с атомарными. Это рано или поздно приводит к тому, что показатель рассчитан по новым данным, а промежуточные результаты отражаются старые или наоборот.

12. Наличие в организации нескольких корпоративных хранилищ данных, каждое из которых позиционируется как единый и единственный источник непротиворечивой информации, равно как и многослойных витрин данных при наличии единого хранилища.

На рисунке представлена типовая архитектура ИАС для организации, имеющей большое число территориально разбросанных филиалов.

Множество, представленное на рисунке – полное, то есть все уровни и подуровни, которые есть в архитектуре какой-либо информационно-аналитической системы, – есть на этом рисунке, а порой и в нескольких экземплярах.

Наличие в организации многослойных витрин данных при наличии единого хранилища, то есть попытка построить третью витрину поверх хранилища, а также объединение некоторых данных из двух витрин в третью – имеет смысл, если в витринах содержалась информация, которой нет в хранилище, то есть если пользователи обогащают витрину своими расчетами и данными. Но при этом возникает множество вопросов, которые в рамках хранилища уже решены:

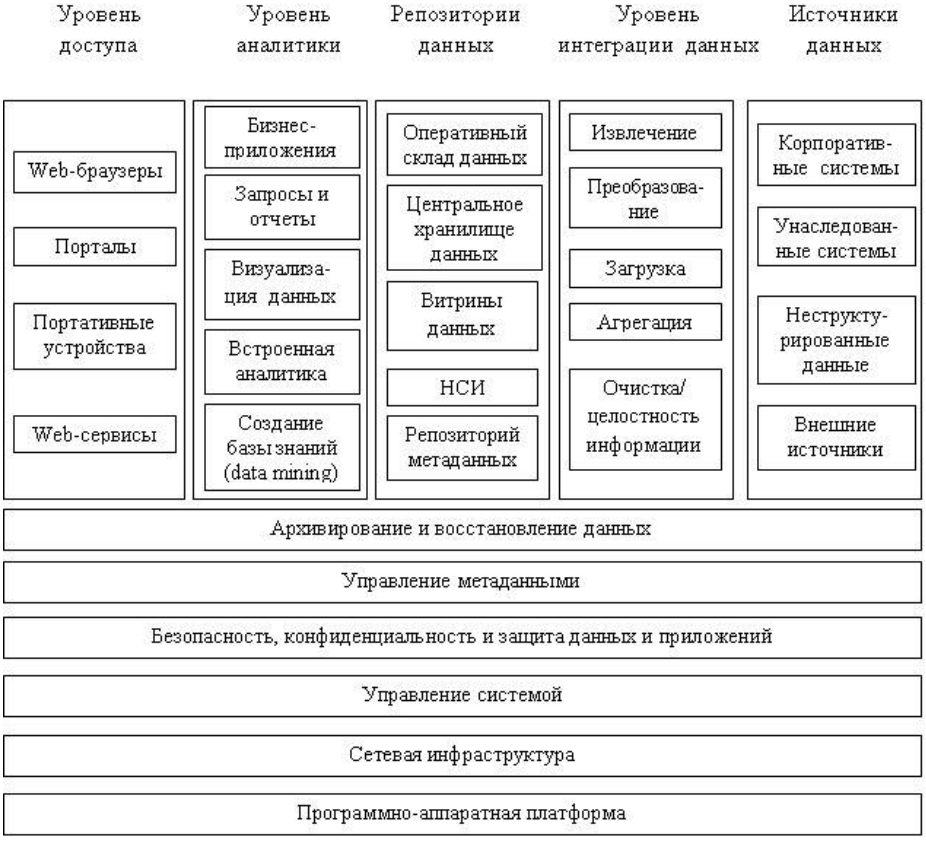
• какова ценность обогащенных данных по сравнению с теми, что прошли очистку в соответствии с корпоративными правилами?

• Кто отвечает за качество этих данных?

• Источник этих данных и правила их загрузки в систему?

Остановимся подробнее на каждом уровне архитектуры ИАС.

Источники данных - это транзакционные и аналитические унаследованные системы, архивы, разрозненные файлы числовых данных известных форматов, тек-



Типовая архитектура информационно-аналитических систем

стовые файлы различных форматов, а также любые иные источники структурированных и неструктурированных данных.

ETL (Extract, Transformation and Load) - извлечение, преобразование и загрузка данных, а также ELT (Extract, Load and Transformation) – извлечение, загрузка и преобразование данных, основная задача которых извлечь данные из разных систем, привести их к согласованному виду и загрузить в хранилище. Программно-аппаратный комплекс, на котором реализованы эти функции, должен обладать значительной пропускной способностью. Но еще важнее для него – это высокая вычислительная производительность. Поэтому лучшие из них способны обеспечивать высокую степень параллелизма вычислений и даже работать с кластерами и вычислительными гридами.

Системы ведения метаданных и нормативно-справочной информации (НСИ) – я также освещу в отдельной главе. Все метаданные (то есть «данные о данных» – форматы данных, категорированные определенным образом в данной системе, а также место и способ их хранения) можно разделить на четыре категории – технические метаданные, операционные метаданные, бизнес – метаданные и проектные метаданные, которые, вообще говоря, также можно отнести к техническим. Наиболее важными метаданными – являются метаданные организации бизнес-уровня.

Прежде всего, необходимо определить, где находятся требуемые данные из всех вышеперечисленных источников. Поскольку, как правило, эти данные хранятся в различных форматах, необходимо их привести к единому виду с использованием ETL в хранилище данных.

Эта работа не может быть выполнена без сопутствующего анализа метаданных и нормативно-справочной информации (НСИ), которая, по сути дела, является метаданными бизнес-уровня. Более того, практика внедрения хранилищ данных показала, что метаданные, созданные и импортированные из различных источников, фактически управляют всем процессом сбора данных.

Оперативный склад данных (Operational Data Store) необходим тогда, когда требуется как можно более оперативный доступ к пусть неполным, не до конца согласованным данным, доступным с наименьшей возможной задержкой. Зоны временного хранения (Staging area) нужны для реализации специфического бизнес-процесса, например, когда перед загрузкой данных контролер данных должен просмотреть их и дать разрешение на их загрузку в хранилище.

Иногда зонами временного хранения называют буферные базы данных, необходимые для выполнения внутренних технологических операций, например, ETL выбирает данные из источника, записывает их во внутреннюю БД, обрабатывает и передает в хранилище. Под зонами временного хранения будем понимать области хранения данных, предназначенные для выполнения операций внешними пользователями или системами в соответствии с бизнес-требованиями обработки данных. Выделение зон временного хранения в отдельный компонент ХД необходимо, так как для этих зон требуется создание дополнительных средств администрирования, мониторинга, обеспечения безопасности и аудита.

Информационные системы на уровне распределения данных все еще не имеют общепринятого названия. Они могут называться просто ETL так же, как и система извлечения, преобразования и загрузки данных на втором уровне. Или, чтобы подчеркнуть отличия от ETL, их иногда называют ETL-2. При этом системы уровня распределения данных выполняют задачи, значительно отличающиеся от задач ETL, а именно выборку реструктуризацию и доставку данных (SRD – Sample, Restructure, Deliver) ETL извлекает данные из множества внешних систем. SRD выполняет выборку из единого хранилища данных. ETL получает несогласованные данные, которые надо преобразовать к единому формату. SRD имеет дело с очищенными данными, структуры которых должны быть приведены в соответствие с требованиями различных приложений. ETL загружает данные в центральное хранилище. SRD должно доставить данные в различные витрины в соответствии с правами доступа, графиком доставки и требованиями к составу информации.

Для разделения функций хранения и функций обслуживания различных задач предназначены витрины данных, которые должны иметь структуры данных, максимально отвечающие потребностям обслуживаемых задач. Поскольку не существует универсальных структур данных, оптимальных для любой задачи, витрины данных следует группировать по территориальным, тематическим, организационным, прикладным, функциональным и иным признакам [1–4].

Бизнес-приложения можно по формальному признаку разделить на системы сценарных расчетов, системы статистического анализа, системы многомерного анализа, средства планирования и подготовки отчетности.

При разработке ИАС в настоящее время с учетом наличия современных средств телекоммуникаций, линий связи и возможностей вычислительной техники можно порекомендовать использовать следующие подходы для системы, у которой большое число пользова-

телей, а также большое число источников информации (к которым мы относим и датчики автоматизированного сбора данных мониторинга земной поверхности для различных систем подобных широко известной Международной аэрокосмической системы мониторинга опасных для цивилизации природных явлений и техногенных катастроф МАКСМ, пользователями которой будут правительства различных стран мира):

1. Интеграция и консолидация данных

Разрозненные данные должны быть интегрированы в едином централизованном корпоративном хранилище данных, которое представляет собой предметно-ориентированный, интегрированный, неизменяемый, поддерживающий хронологию набор данных, организованный для целей поддержки принятия решений, и должно стать единым источником непротиворечивых и согласованных данных для всех подсистем системы.

2. Централизованное ведение метаданных и нормативно-справочной информации (НСИ)

Все подсистемы должны использовать единые, ведущиеся централизованно метаданные и НСИ, обеспечивать возможность формирования локальных справочников, поддерживать историчность (версионность) метаданных и НСИ для обеспечения возможности проведения анализа с использованием данных за предшествующие временные периоды.

3. Обеспечение возможности формирования произвольных аналитических запросов и отчетов пользователями

Необходимо обеспечить возможность формирования произвольных аналитических запросов и отчетов пользователями, не обладающими навыками программирования, в терминах своей предметной области.

4. Унификация взаимодействия с кредитными организациями

Наряду с унификацией внутренних технологических процедур необходимо исходить из принципов унификации процессов взаимодействия с пользователями системы из различных стран мира в части единой транспортной системы, форматов данных.

5. Использование НСИ и каталога бизнес-терминов как основы для построения метаданных, обеспечивающей их единообразное формирование

Метаданные должны строиться на основе каталога бизнес-терминов, определяющего функциональный смысл показателей.

6. Унификация и типизация проектных решений

Использование унифицированных технологических решений на уровне привелигерованных пользователей и типовых проектных решений на уровне обычных пользователей.

7. Открытость и эволюционность

Архитектура системы должна обеспечивать возможность поэтапной разработки и внедрения. Следствием этого является возможность практически неограниченного расширения функционального наполнения системы без принципиальной замены системно-технической платформы.

8. Масштабируемость

Необходимость обеспечения существенного роста потоков данных, количества рабочих мест и количества задач без существенного изменения прикладного программного обеспечения. Масштабирование должно обеспечиваться средствами администрирования и настройки, а также, за счет увеличения мощности технических ресурсов и перераспределения нагрузки административными средствами.

9. Надежность и безопасность

Система должна быть надежной и защищенной, обеспечивать бесперебойную работу, получение достоверных результатов и защиту от несанкционированных действий.

10. Обеспечение катастрофоустойчивости системы

Архитектура должна предусматривать возможность создания территориально распределенных комплексов (особенно центров обработки информации) – основных и резервных. Резервные комплексы должны выполнять функцию замещения основных комплексов при возникновении чрезвычайных ситуаций, ведущих к выходу последних из строя. Резервные центры должны размещаться на объектах с заранее подготовленной инфраструктурой и информационно-телекоммуникационными ресурсами.

11. Преемственность

При создании системы необходимо максимально полно использовать уже существующие технологические системы и инфраструктурные элементы. При необходимости используемые системы должны быть доработаны по требованиям, сформированным в процессе проектирования системы.

Для реализации вышеуказанных подходов требуется провести следующий комплекс мероприятий:

• спроектировать и создать единое корпоративное хранилище данных;

• централизовать процессы сбора, хранения и аналитической обработки данных с использованием каталога бизнес-терминов из различных информационных источников, с учетом требований к обеспечению

информационной безопасности, контроля и разграничения прав доступа к информации с различным уровнем конфиденциальности;

• унифицировать информационные системы и обеспечить требуемое развитие их функциональных возможностей путем применения современных информационных технологий;

• разработать и внедрить единые стандарты информационного взаимодействия между структурными подразделениями, эксплуатирующими ИАС, и с внешними абонентами на базе единых методических и нормативных требований;

• обеспечить возможность ретроспективного и прогнозного анализа информации системы на основе разнородной информации с использованием единого интерфейса пользователя;

• обеспечить возможность взаимодействия ИАС со смежными системами.

Проведение вышеуказанных мероприятий позволит:

• обеспечить методологическую совместимость существующих смежных информационных систем и проектируемой ИАС;

• повысить уровень защиты, конфиденциальности и целостности коллективных информационных ресурсов системы;

• уменьшить трудоемкость и стоимость разработки, эксплуатации и сопровождения средств сбора, обработки и предоставления отчетной информации и аналитических приложений.

Выводы

При разработке информационно-аналитической системы с большим числом пользователей и большим числом источников информации в настоящее время с учетом наличия современных средств телекоммуникаций, линий связи и возможностей вычислительной техники можно порекомендовать централизованную архитектуру с единым хранилищем данных, отвечающим концепции Инмана.

Литература

1. Меньшикова Л.В. Принципы построения корпоративного хранилища данных // Двойные технологии. – 2007.-№4(41).-С.73-76.

2. Menshikova L.V. "Development of Informational and Analitical Systems with Asquisition of Feedback from Users in the Framework of Large-scale Projects//Abstracts of First International Specialized Symposium Space and Global Security of Humanity"/Limassol-Cyprus: International Academy of Astronautics, Russian Academy of Cosmonautics, International Association "Znanie", 2009.- p.57.

3. В.Ю.Пирогов. Информационные системы и базы данных. Организация и проектирование //С.Петербург: БХВ-Петербург, 2009.– 528 с.

4. Банковские информационные системы /Москва: Маркет ДС, 2010. – 816 с.

Материал поступил в редакцию 12. 08. 2011 г.