# Интеграция данных – трудности перевода


Форум

**Интеграция корпоративных прикладных систем 2011**

Ноябрь 2011 г., Москва

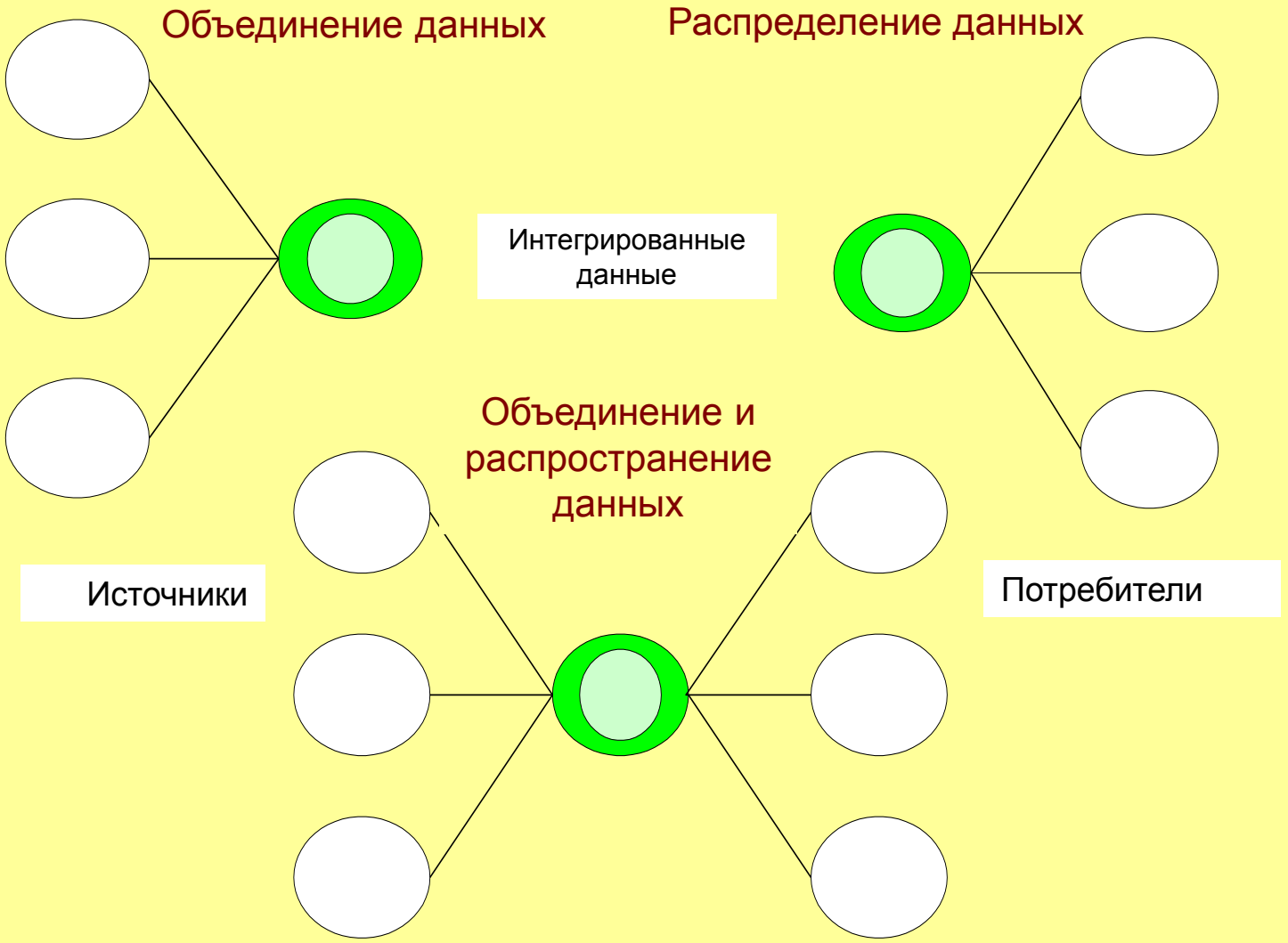
Артемьев Валерий Иванович (Банк России) 

# (1) Что такое интеграция данных?

*Интеграция данных* – обеспечение единого согласованного представления данных для ряда информационных ресурсов, объединенных общим смысловым содержанием, и/или на основе общего представления – частных представлений.

- Основные цели – повышение оперативности принятия решений и улучшение их качества
- Единая модель интегрированных данных
- Виртуальное или материализованное представление
- Однородные или неоднородные модели и схемы
- Фиксированный или динамический состав источников
- Вытягивание или проталкивание данных
- Реляционные, объектные, объектно-реляционные БД, файлы данных, унаследованные системы, репозитарии, Web-сайты и сервисы как источники и потребители.
- Пересекается с интеграцией приложений

# Принципиальные схемы интеграции данных




**EDI**

**EDR**

**ETL**

**CDC**

**EII**

**ECM**

**MOM**

**ESB**

**MDM**

# Метафора перевода текста для интеграции данных

- Общее у интеграции данных и перевода текста – преобразование формы представления с максимальным сохранением смысла.
- Перевод посредством общего языка предметной области
- Бизнес-глоссарий, словари-справочники данных, НСИ
- Различие и отсутствие понятий
- Синонимы и омонимы
- Жаргонизмы и архаизмы
- Синекдоха
- Проблемы интеграции данных обусловлены, прежде всего, трудностями перевода.

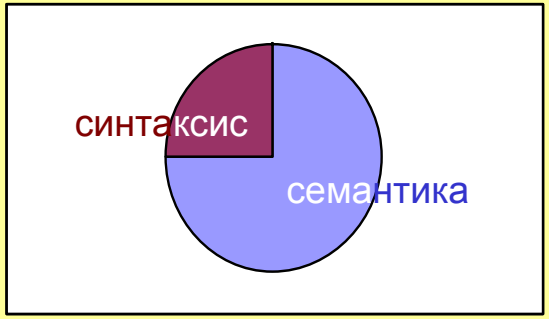
# Подходы к интеграции данных

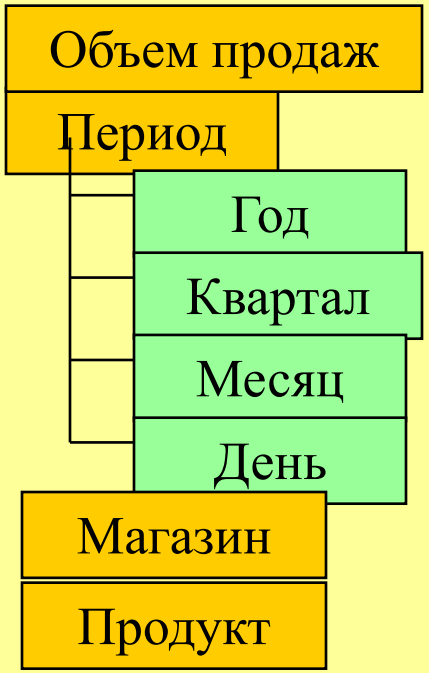
1. **Синтаксический подход** основан на внешнем сходстве интегрируемых данных, семантика второстепенна, своеобразный подстрочник.
- Фокус на технические метаданные – отображение и преобразование данных на основе физических, реже логических схем. Первые роли у ИТ.

2. **Семантический подход** основан на содержательном сходстве объединяемых данных, синтаксис второстепенен, профессиональный перевод.
- Фокус на бизнес–метаданные (онтологии, семантические сети, концептуальные схемы, таксономии, каталоги и модели показателей, глоссарии). Первые роли у бизнеса.
- Имеются удачные примеры (Home Credit Bank), когда методическая проработка при интеграции данных дала больший эффект, чем внедрение новых технологий.
- Интеграция данных – прежде всего, организационно-методическая, а не техническая задача.

# Зачем нужны метаданные?

- Изменчивость бизнес–среды привела к использованию метаданных для параметризации и настройки процессов сбора, хранения и обработки данных
- 75% успеха в обеспечении качества данных зависит от методической проработки семантики данных специалистами предметной области, т.е. от бизнес–метаданных (Gartner)
- Бизнес–аналитика как основной потребитель КХД основана на описаниях данных в терминах предметной области
- Важна фиксация проектных решений по КХД в виде электронных моделей и других видов метаданных для автоматизации его разработки и сопровождения.



# Примеры бизнес–метаданных

- Термины и категории предметной области, их классификации (глоссарий и концептуальная модель)
- Перечень и описание измерений и фактов, многомерные модели показателей (кубы)
- Связи терминов, описаний показателей, измерений, логических моделей
- Представление семантических слоев пользователей для витрин данных (метамодели, юниверсы)
- Точность и единицы измерения значений показателей
- Логические модели компонентов КХД и источников данных
- Бизнес-правила контроля и очистки данных
- Формулы (алгоритмы) расчета производных показателей

# Проблемы качества данных

- Многие проекты интеграции данных потерпели неудачу из-за недостаточного качества данных.
- Контроль и очистка данных для их согласования решают лишь малую часть проблем качества данных.
- Большая часть проблем заключается в отсутствии единого понимания предметной области и в разной интерпретации бизнес–правил, в фрагментарном покрытии предметной области и несогласованности уровней детальности, несовместимости справочников
- Качество данных зависит от их применения.
- Кроме усилий по очистке данных требуются методическая проработка, согласование терминов, структуризация и алгоритмизация каталога показателей на основе таксономии и/или многомерной модели.

# Что существенно влияет на качество данных?

- Согласованный бизнес–словарь
- Устойчивые классификационные схемы и справочники
- Временная привязка показателей
- Охват наблюдаемых явлений
- Идентификация участников бизнес–процессов
- Подходы к определению показателей
- Полнота именования
- Трактовка содержания данных
- Однозначно определенные алгоритмы расчета для производных показателей
- Предварительный анализ качества данных источников
- Согласованное ведение НСИ
- Аудит и мониторинг качества данных

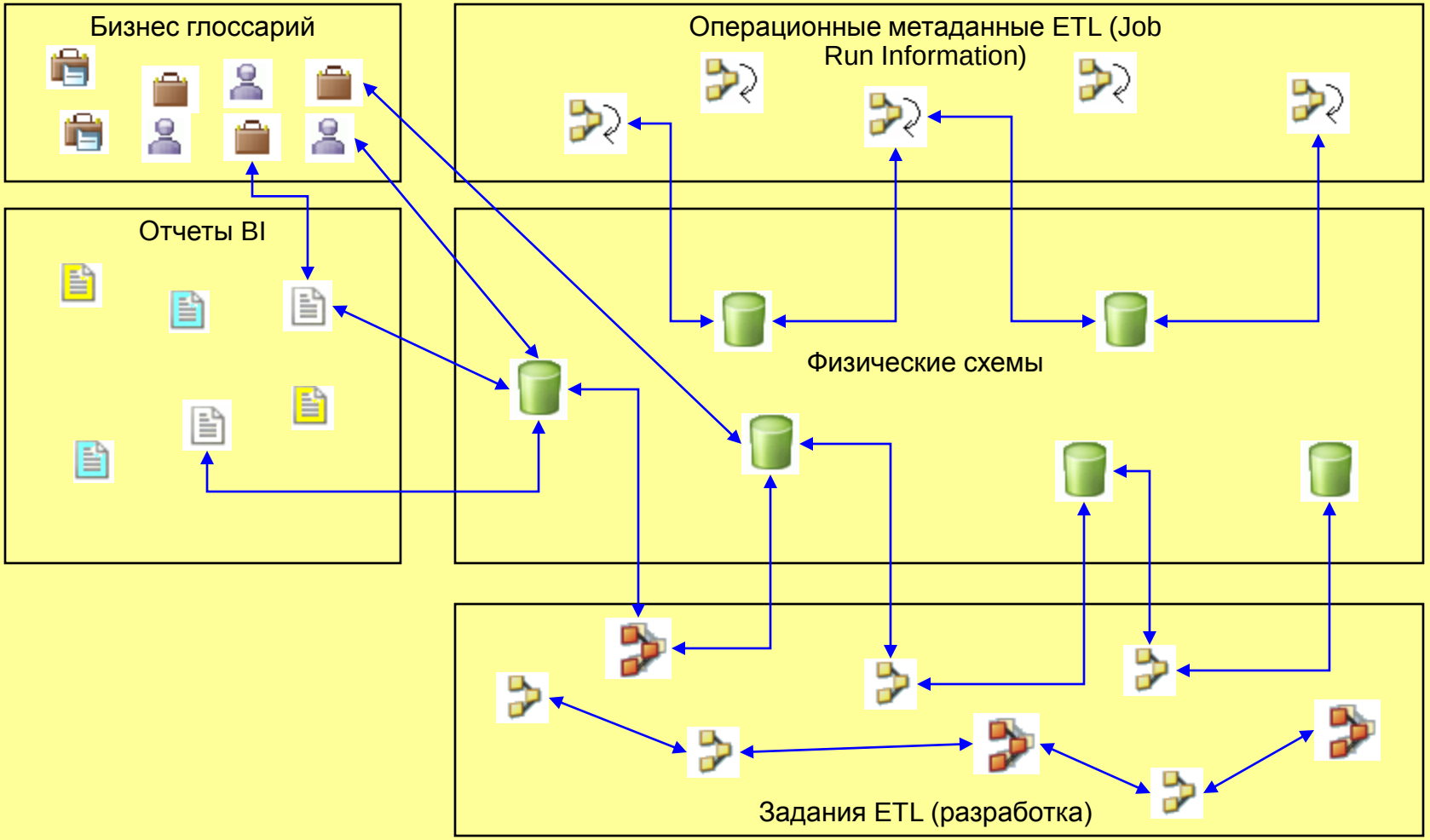
# С чем столкнулись при интеграции данных в Банке России?

- Отчетная дата (периодичность сбора) или период (периодичность) наблюдения?
- Фрагментарность наблюдений по времени (7 уровней)
- Десятидневка (декада) не состоит из пятидневок
- Размыто понятие «кредитная организация»
- Территория РФ или территориальное учреждение Банка России?
- «Арест платежа» в одной УОС отсутствовал в другой
- Индивидуальный предприниматель – юридическое лицо?
- Нет уникального идентификатора КО (рег.№, БИК, ОГРН)
- Неустойчивые частные классификации, неполные группы
- Много новых классификаторов и справочников вне НСИ
- Нормативный расчет или мотивированное суждение бухгалтера
- Неполные наименования, относительные этикетки
- Элемент данных может содержать разные домены значений
- Сбор аналитических отчетов, а не данных для построения отчетов
- Из-за безопасности не проводился анализ качества источников

# Подходы к управлению метаданными

- **Федеративная схема управления метаданными** – децентрализованное ведение частных метаданных, централизованное их хранение
- **Консолидация метаданных** – импорт и связывание частных метаданных в общем репозитории метаданных
- **Интеграция метаданных** – согласованность описаний и свойств данных как основа семантической интеграции данных
- **Фокус на бизнес–метаданные** – ведение и использование бизнес–метаданных для понимания данных и результатов анализа в терминах предметной области
- **Множество представлений и связность метаданных** – основа для преобразования метаданных, отображения одних данных на другие, анализа зависимости и происхождения
- **Унификация форматов обмена метаданными**
- **Историчность и версионность метаданных**
- **Коллективное ведение метаданных**
- **Ведение каталога показателей банковской статистики (КПБС)**

# Связывание областей метаданных



# Готовые модели КХД для банков на основе IBM Banking Data Warehouse

**Взаимосвязанные модели КХД**

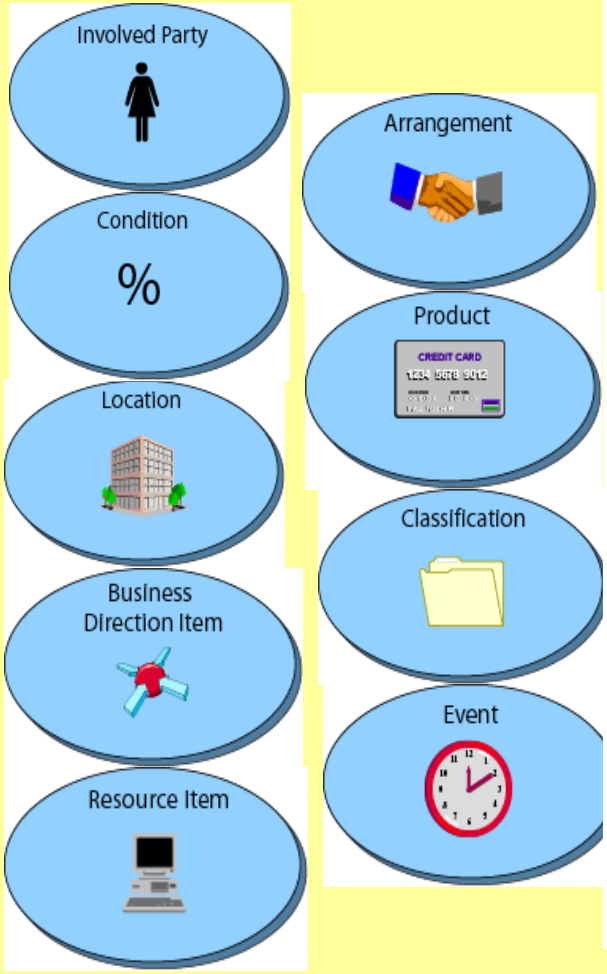
**для западного коммерческого банка**

- Концептуальная модель - модель терминов и категорий (FSDM)
- Модели аналитических требований - многомерные модели показателей для бизнес–аналитики (BST)
- Логическая модель данных хранилища (BDWM)
- Физическая модель данных хранилища

**Инструменты для ведения моделей**

**Методические руководства**

# Концептуальная модель IBM BDW

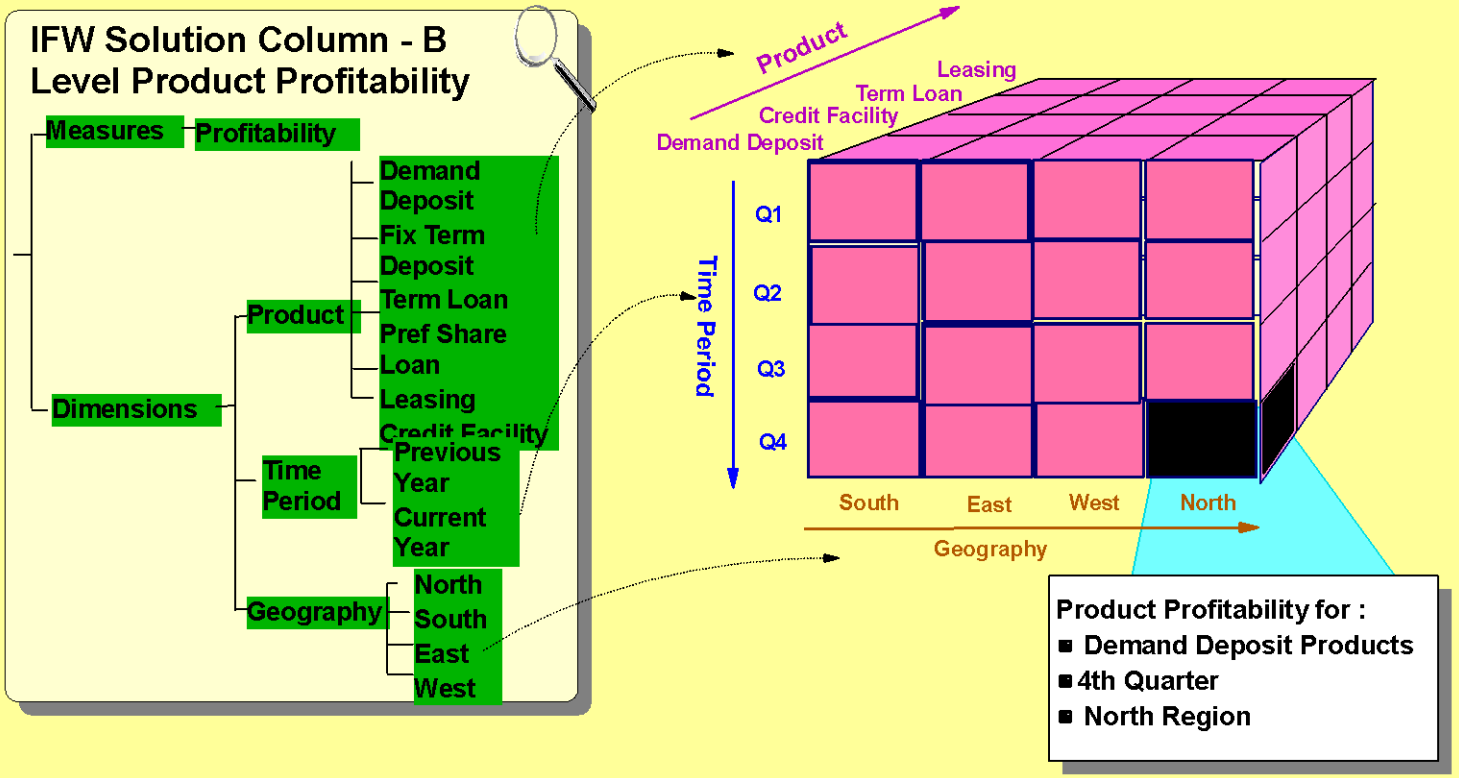



- **Концептуальная модель** задает рамки для предметной области и логической модели на основе **«концепций»**:
- Вовлеченная сторона (Involved Party) – клиент, банк, физическое лицо, …
- Местоположение (Location) – местоположение банка или клиента банка
- Единица учета (Accounting Unit) – учетные данные
- Классификация (Classification) – понятия для классификации информации
- Продукт (Product) – проданные/приобретенные товары или услуги
- Условие (Condition) –требования к финансовой деятельности
- Элемент ресурсов (Resource Item) – материальные и нематериальные стоимостные позиции
- **Классификационные иерархии**:
- реализует декомпозицию самой концепции на более «узкие» категории
- классификации атрибутов, относящихся к данной концепции (Descriptor)
- связи между сущностями разных «концепций» (Relationship)

# Модель аналитических требований

**Profitability**
**Risk**
**Compliance**

**Asset & Liability Management**



# (2) Что даст Банку России создание КХД?

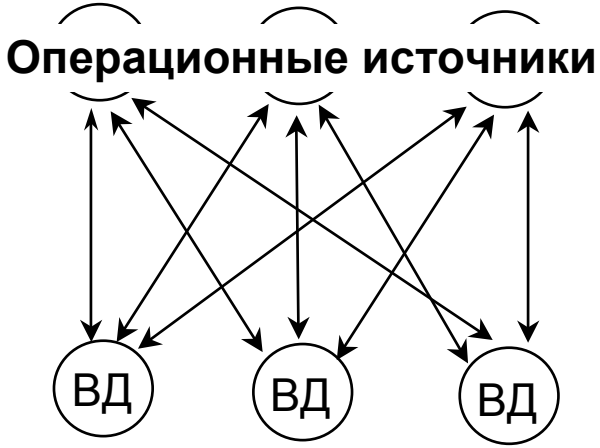
- **Повышение качества принятия решений** на основе непротиворечивых фактов
- **Повышение оперативности принятия решений** за счет сокращения времени подготовки и получения данных
- **Единая информационная модель предметной области** на основе согласованной терминологии, таксономии, справочной информации
- **Снижение нагрузки по сбору отчетности** путем устранения дублирования и производных данных
- Основа для корректного **применения методов и средств бизнес–аналитики**
- **Анализ зависимости данных** при внесении изменений
- **Анализ происхождения данных** при анализе данных
- **Синхронизация метаданных** (форматов сбора данных, структур источников данных, моделей КХД, метаданных бизнес–аналитики)
- **Ускорение внесения изменений** в КХД, аналитические приложения и сбор данных
- **Сокращение затрат на сопровождение** КХД, аналитических приложений и сбора данных
- **Сокращение затрат на эксплуатацию и модернизацию инфраструктуры**

# Подходы к созданию КХД

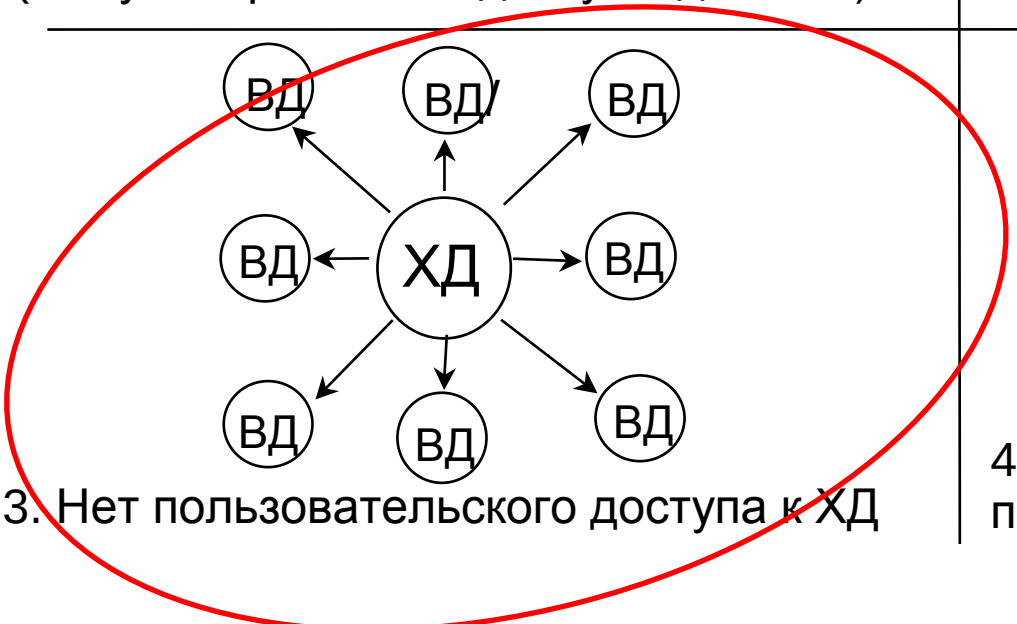
- **Что входит в состав КХД?**
  - ✓ Данные для анализа и системы управления данными
  - ✓ Средства интеграции и консолидации данных
  - ✓ Метаданные и системы управления метаданными
  - ✗ Средства бизнес–аналитики
- **Одно или несколько хранилищ данных?**
  - ХД по основной деятельности
  - ХД по административной, финансовой и внутрихозяйственной деятельности
  - ХД или витрины данных по эксплуатации АС и инфраструктуры
- **Многоуровневое ХД или набор согласованных витрин данных**
- **Централизованное или распределенное ХД**
- **Традиционное или динамическое ХД**
  - «чистые» ретроспективные данные для анализа
  - «сырые» оперативные данные для анализа или совместно с ретроспективными данными при транзакционной обработке
- **Применение апробированных решений и средств ведущих производителей**
  - соответствие МСФО, правилам Базель II
  - готовые отраслевые модели хранилища данных
  - апробированные методики, технологии и средства

# Варианты топологии “хранилища данных”

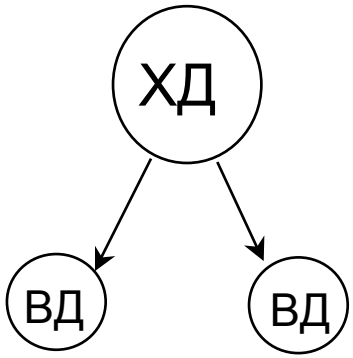

1. Виртуальное хранилище данных (т.е. универсальный доступ к данным)




2. Много витрин данных



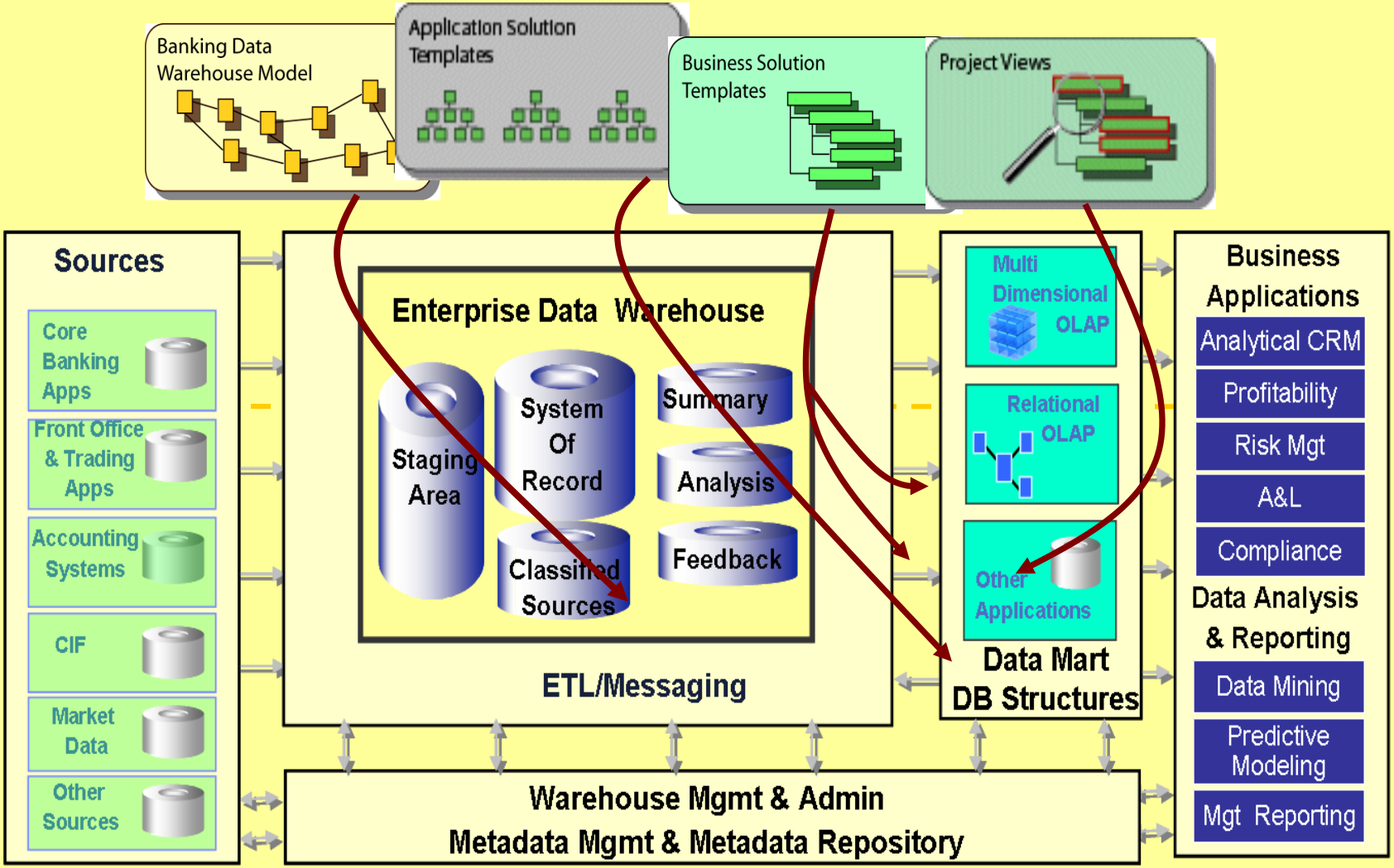

3. Нет пользовательского доступа к ХД



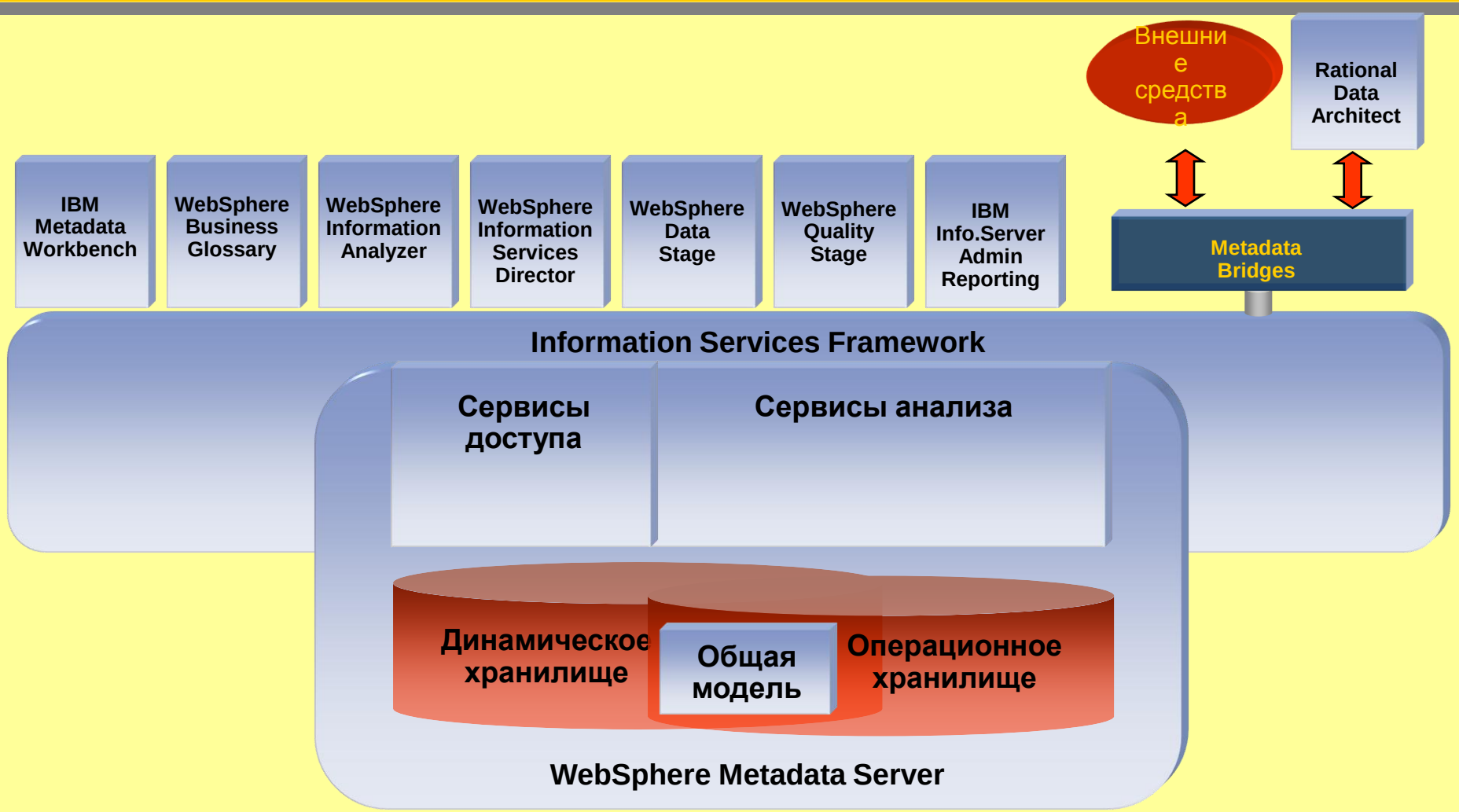

4. Избранные витрины данных и прямой доступ пользователя к ХД

Gartner

# Архитектура КХД на основе IBM BDW



# Архитектура подсистемы управления метаданными



# Используемые программные средства

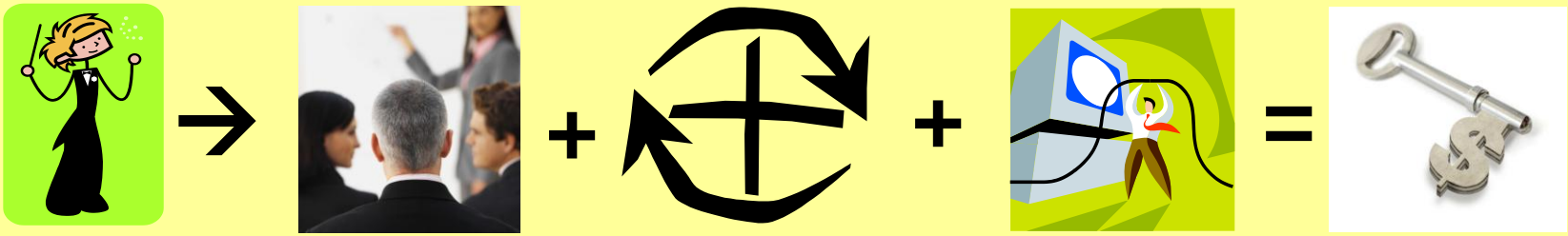
- Rational Data Architect (RDA) – средство проектирования моделей данных
- IBM Enterprise Model Extender for RDA – инструменты для работы с моделями BDW в RDA
- WebSphere Metadata Server – сервер метаданных
- WebSphere Information Analyzer – анализатор качества данных
- FastTrack – средство описания бизнес–правил и преобразований для ETL
- WebSphere DataStage & QualityStage – средства для формирования ETL-процедур, контроля и очистки данных
- WebSphere Business Glossary – глоссарий терминов
- WebSphere Metadata Workbench – инструмент администратора метаданных

# (3) Организационные проблемы создания КХД

- Отсутствие координации подразделений при расширении состава собираемых данных
- Слабо вовлечены бизнес–подразделения в управление метаданными
- Недостаток компетенции по концептуальным моделям, CASE и другим новым средствам – требуется расширенное обучение
- По соображениям ИБ затруднен анализ качества реальных данных (профилирование)
- Нужна оргструктура и новые роли для корпоративного управления данными (Data Governance)
- Нужны стратегия, концепция, стратегический план и корпоративный бизнес–проект
- Необходимы службы для непрерывной разработки и сопровождения КХД
- Нежелание разработчиков использовать визуальные средства разработки ETL-процедур

# Подход к управлению корпоративными данными Data Governance

Data Governance – политический процесс изменения поведения организации по распоряжению данными как стратегическим корпоративным активом



Внедрение Data Governance – фундаментальное изменение в методах и строгости определения, управления и использования данных как со стороны бизнеса, так и со стороны ИТ.

**Основные задачи Data Governance:**

- **Руководство принятием решений по управлению информацией**
- **Обеспечение согласованности определения и понятности информации**
- **Увеличение степени использования и доверия к данным как корпоративному активу**
- **Улучшение согласованности проектов в масштабе корпорации**

# Новые роли по управлению корпоративными данными

# Методические вопросы создания КХД

- Трудности адаптации моделей BDW в части концептуальной модели, многомерных моделей показателей и бизнес–глоссария
- Методология разработки КХД не адаптирована для его сопровождения
- Разрыв в создании и сопровождении метаданных КХД, метаданных форматов сбора и бизнес–аналитики
- Нарушение принципа неизменяемости КХД
- Нужна четкая политика формирования витрин
- Для нерегламентированных отчетов и OLAP нужно повышение качества данных
- Нужен переход от сбора отчетов к сбору показателей
- Не завершены методика создания КПБС и работы по интеграции КПБС с управлением метаданными

# Технические проблемы создания КХД

- Неповоротливая архитектура КХД
- В новой версии инструментов для BDW потеряны возможности генерации метаданных
- В составе BDW не поставляются модели витрин данных
- Не все модели BDW импортируются в общий репозиторий метаданных
- Версионность метаданных обеспечивается с помощью дополнительных средств Rational ClearCase
- Схема управления метаданными громоздкая, много ручного труда
- Нужны технология MDM для согласованного ведения НСИ и ILM для управления жизненным циклом информации
- Соотношение ETL и сбора данных

# Технология Master Data Management

- Собирает мастер–данные из отдельных приложений
- Создает центральный ресурс, независимый от приложений
- Упрощает текущие задачи интеграции и разработку новых приложений
- Гарантирует согласованность мастер–данных для транзакционных и аналитических систем
- Разрешает ключевые проблемы качества и согласованности данных проактивно до поступления данных в КХД

# Терпения и удачи всем, кто занимается интеграцией данных

**Спасибо за внимание!**

Валерий Иванович Артемьев

Центр информационных технологий Банка России

Тел.: +7(495) 753-96-25

e-mail: avi@cbr.ru