Практика построения систем на основе технологии Хранилищ Данных (на примере системы SAS)

часть 1

Рассматривается методология и технология компании SAS Institute по созданию, управлению и эксплуатации Хранилищ Данных.

«Базы данных и знаний»

Лекция 7

План лекции

1. Методология создания, управления и эксплуатации Хранилищ Данных.
2. Архитектура ИС на основе технологии ХД.
3. Метаданные.
4. Хранение. Структура Репозитория ХД.

- В течение последнего десятилетия термин DataWarehouse или Хранилище Данных стал одной из самых популярных тем для обсуждения.
- Любая система, хранящая собираемые данные, претендует на название Хранилища Данных.
- Ведутся ожесточенные теоретические споры по поводу, что называть Хранилищем Данных, чем оно отличается от OLTP систем и какое определение более точно описывает его сущность.
- Мы попробуем подойти с другой стороны – практической.

- Одним из самых важных вопросов при построении любой информационной системы является четкое представление о том, что и как должно строиться, а говоря более строгим языком, – архитектура системы и методология ее построения.

- Для информационных систем в технологии Хранилищ Данных, главными особенностями которых являются большой масштаб проекта и постоянное развитие, наличие этих составляющих является обязательным.

- **Методология**

Любой проект, в конечном счете, должен принести прибыль.

По общему мнению, Хранилище Данных создается для информационной поддержки процесса принятия решений, осуществляет процесс доставки необходимой, актуальной и верной информации нужным людям в нужное время, с тем, чтобы они могли принимать обоснованные и своевременные решения.

- Для большой организации велико число лиц, принимающих решения и нуждающихся в такой информационной поддержке.

Реализация проекта, охватывающего деятельность всей организации, потребует длительного времени и больших затрат.

К тому же, конкретные требования к доставляемой информации меняются достаточно быстро, и возможно, что по окончании проекта его актуальность будет далека от ожидаемой.

Осознание этого факта приводит нас к двум выводам.

- Во-первых, создание ХД должно осуществляться короткими и быстро дающими ощутимые результаты этапами.

- Во-вторых, созданное Хранилище Данных не будет статической, раз и навсегда созданной системой.

Требования к доставляемой для принятия решений информации будут меняться в процессе его эксплуатации и развития, поэтому логическая и физическая структура ХД, должна позволять достаточно быстро и безболезненно осуществлять требуемые изменения.

- *Думать стратегически, начинать с решения самых насущных задач – вот лозунг, под которым идет создание и развитие Хранилища Данных.*

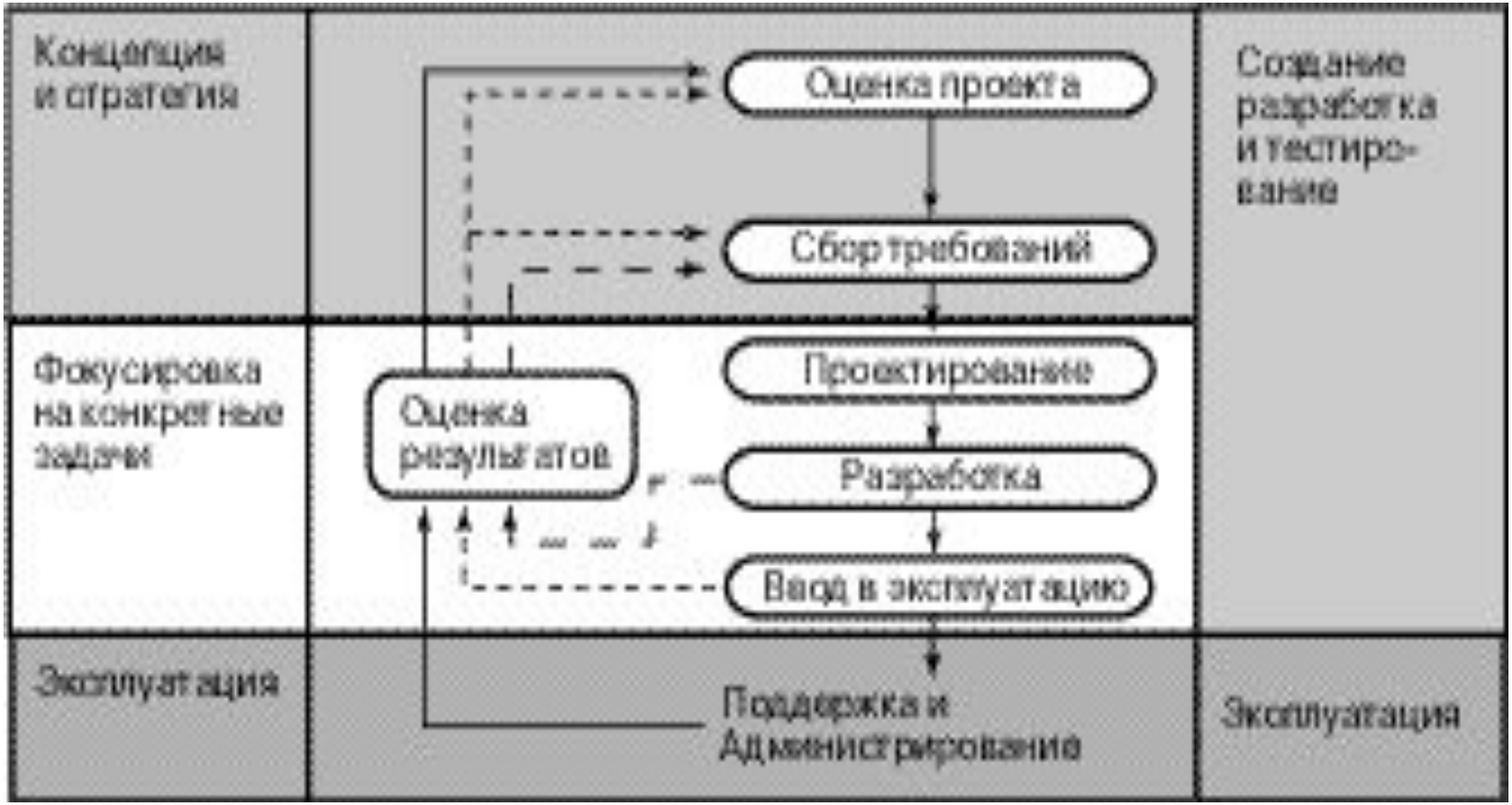
- *Практически это означает, что на первой стадии важно создать ясную картину будущего содержания Хранилища Данных в самых грубых чертах, затем выбрать наиболее актуальную задачу, и руководствуясь концепцией архитектуры Хранилища Данных работать над ней, не забывая об общей картине.*

- Этот подход включает в себя определение стратегических путей развития на самых ранних этапах построения Хранилища и поэтапную реализацию этих планов с учетом заданных приоритетов в виде краткосрочных (4-8 месяцев) проектов.

Такой подход позволяет получить ощутимые результаты в очень короткие сроки и при этом сохранить уверенность в том, что дальнейшее расширение создаваемого Хранилища не приведет к перестройке всей его структуры.

Основные стадии построения Хранилища данных

Каждый этап построения Хранилища реализуется как отдельный проект, который имеет основные стадии, представленные на рисунке



- **Оценка проекта**

1. Оценить условия для реализации проекта, которые помимо традиционных условий, таких как достаточный бюджет, наличие команды разработчиков и прочие, включают понимание концепции Хранилищ Данных, определенный уровень информационной инфраструктуры и другие.

2. Для эффективного контроля над ходом выполнения проекта важно с самого начало выбрать четкий и ясный критерий оценки проекта.

- *Оценка проекта*

3. Иметь четкое «видение» основных проблем организации в целом, а также определить основные источники данных, чтобы, получив общую, концептуальную модель деятельности организации, использовать ее как базис для стратегии построения Хранилища.

   Из всего разнообразия задач выбирается наиболее значимая задача, как цель первого (или очередного) этапа построения Хранилища Данных.

- **Сбор требований**

Определение круга пользователей, связанных с решаемой на данном этапе проблемой, и их опрос.

Параллельно осуществляется анализ имеющихся источников информации и способов доступа к ним.

- *Сбор требований*

В результате конкретизируются:

- структура данных в источниках;
- основные предметные области, связанные с задачей;
- требования пользователей.

На этой же стадии вводится четко определенный и согласованный со всеми участниками проекта критерий завершенности проекта или его очередной итерации.

- **Проектирование**

На этой стадии, на основе информации с предыдущих стадий, производится анализ и проектирование структуры будущего проекта, включающие построение:

- моделей данных (логической и физической);
- модели процессов загрузки ;
- модели приложений.

- **Разработка**

- разработка процедур начальной загрузки ;
- проведение начальной загрузки;
- разработка процедур регулярной загрузки;
- разработка приложений;
- проведение тестовых испытаний.

- **Ввод в эксплуатацию**

- обучение пользователей;
- перевод проекта в стадию эксплуатации (определение администраторов, регламентов и т.д.) .

- **Оценка результатов**

В процессе всего этапа необходимо проводить регулярные оценки результатов ведения проекта с целью минимизировать последствия возникающих проблем на раннем этапе их возникновения.

В случае необходимости произвести откат на предыдущие стадии.

- *Оценка результатов*

Каждый из этапов проведения проекта заканчивается оценкой результатов и презентацией этих результатов всем членам рабочей группы, включая и будущих пользователей системы.

Участие конечных пользователей на всех этапах проектирования, а не только на этапе сбора требований, является ключевым моментом.

Это дает возможность пользователю почувствовать себя самого творцом данного проекта и существенно облегчает процесс внедрения и способствует общему успеху процесса.

Архитектура

Ясное и четкое представление об архитектуре будущей системы должно быть у всех участников проекта: от архитекторов и разработчиков, до опрашиваемых представителей конечных пользователей.

Архитектура

Основными элементами доставки требуемой информации являются:

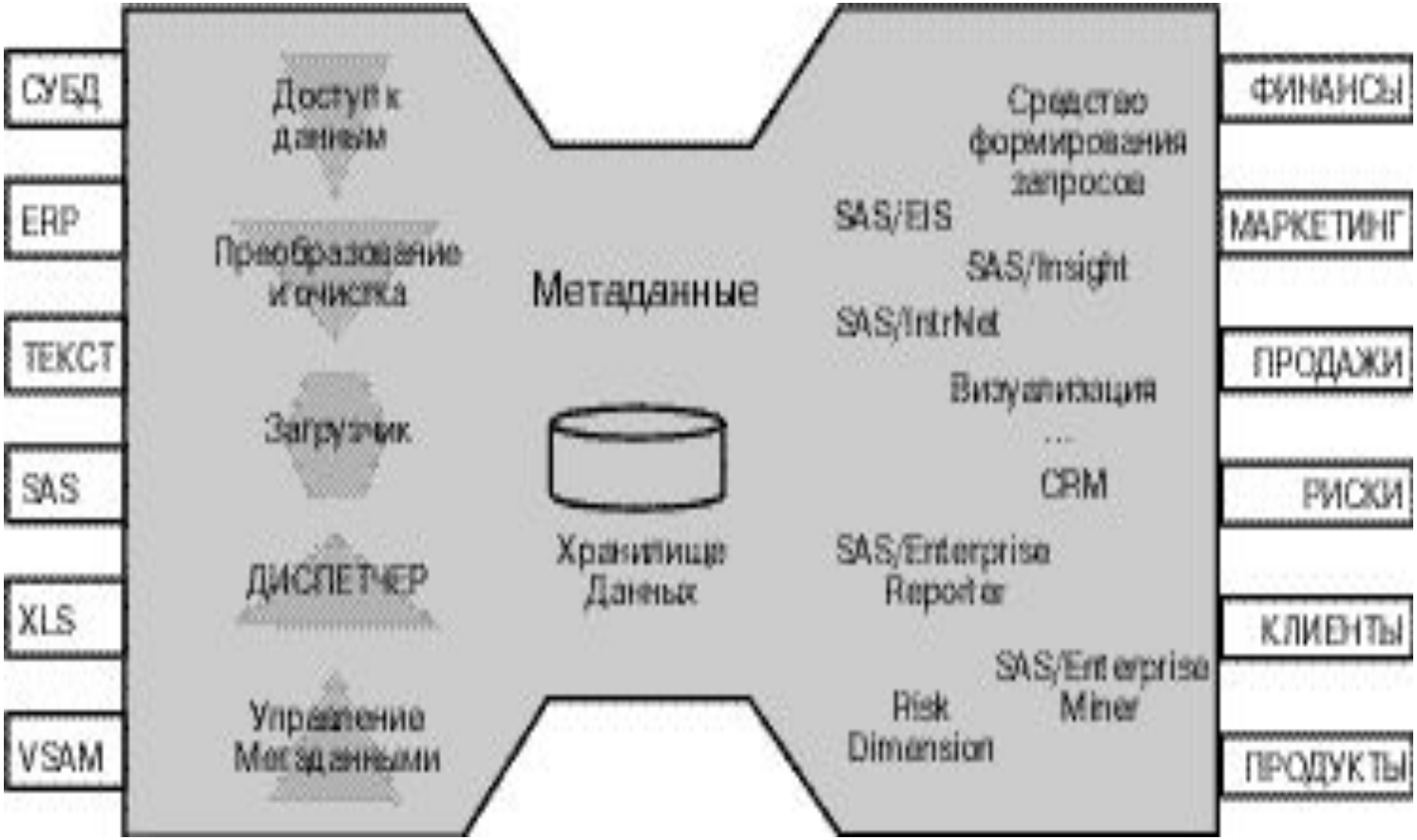
- загрузка данных из источников, в число которых входят различные СУБД, ERP (Enterprise Resource Planning) системы, а также электронные таблицы, текстовые файлы, ленты новостей и другие;
- хранение данных в специально спроектированных структурах, отражающих их предметную специфику и обеспечивающих эффективный доступ;
- использование информации через многочисленные типы приложений в различных разрезах и терминах хорошо понятных конечному пользователю.

- Все элементы системы базируются на краеугольном камне Хранилища Данных – Метаданных.

В Метаданных содержится вся жизненно важная информация о Хранилище, а именно:

- логическая и физическая структуры Хранилища;
- процессы загрузки и их регламент;
- приложения и возможные способы представления информации .

Общая схема информационной системы в технологии Хранилищ Данных



Метаданные

Метаданные (или данные о данных) являются ключевым элементом в Хранилище Данных.

Именно благодаря использованию Метаданных Хранилище становится гибким и удобным средством доставки информации для поддержки принятия решений.

Метаданные

Они содержат:

- полное описание логической и физической структур данных,
- всех процессов загрузки данных,
- специализированных приложений для анализа и представления данных в определенных областях,
- дополнительную информацию обо всех элементах Хранилища, помогающую легко ориентироваться в его сложной структуре.

Метаданными явно или скрыто пользуются все группы пользователей Хранилища, начиная от наименее подготовленных конечных пользователей, приложения для которых управляются Метаданными, до Администратора данных и разработчиков.

По функциональным требованиям эти средства можно разделить на две основные группы:

- средства просмотра и поиска,
- средства создания и поддержки.

Пример

Система SAS предлагает средство **MetaSpace Explorer** для осуществления просмотра и поиска Метаданных Хранилища для конечных пользователей и имеет следующую функциональность:

- навигация по объектам Метаданных Хранилища в разрезах: предметная область, тип, владелец и др.;
- навигация по распределенному Хранилищу в разрезе серверов;
- поиск объектов Хранилища Данных по заданному критерию;
- отображение метаданных, связанных с объектами

В качестве рабочего инструмента Администратора Хранилища Данных и разработчиков создан продукт **SAS/Warehouse Administrator**, позволяющий осуществлять следующие операции:

- определять объекты Хранилища, их атрибуты и взаимосвязи;
- задавать доступ к внешним источникам;
- описывать процедуры загрузки в Хранилище;
- задавать физическую модель Хранилища;
- выполнять процедуры загрузки;
- регламентировать выполнение процедур загрузки;
- генерировать метаданные автоматически;
- представлять структуру Хранилища в удобной графической форме.

Продукт **SAS/Warehouse Administrator** состоит из следующих основных компонент:

- Warehouse Explorer – проводник по Хранилищу;
- Process Editor – редактор процессов;
- Scheduler – диспетчер процессов.

Warehouse Explorer – проводник по Хранилищу, позволяет:

- создавать,
- редактировать,
- удалять, группировать,
- просматривать основные объекты Хранилища Данных.

В их состав входят:

- описания внешних источников данных,
- предметные области и их логические структуры,
- таблицы детальных данных,
- таблицы агрегированных данных и MDDB,
- витрины данных,
- информационные витрины и приложения,
- а также вспомогательные объекты.

Также это средство позволяет задавать основные атрибуты Хранилища: рабочие библиотеки, сервера данных и доступ к ним, лиц, ответственных за эксплуатацию и разработку элементов Хранилища.

Process Editor – редактор процессов позволяет :

- задавать процедуры загрузки для всех элементов Хранилища, от процедур выборки данных из внешних источников до обновления информации в Информационных Витринах и Витринах Данных,
- атрибуты выполнения этих процедур,
- задавать связи между объектами Хранилища,
- графически представлять процесс загрузки любого объекта Хранилища в виде дерева процессов.

Scheduler – диспетчер процессов позволяет регламентировать выполнение процессов загрузки Хранилища Данных.

Включает поддержку распределенных Хранилищ Данных.

Продукт **SAS/Warehouse Administrator** имеет следующие возможности для расширения функциональности и настройки под конкретный проект:

- **MetaData API** – программный интерфейс доступа к Метаданным Хранилища;
- **Scheduler API** – программный интерфейс управления диспетчером загрузки;
- **Средства расширения функциональности** – стандартный интерфейс вызова, дополнительных средств работы с Хранилищем Данных.

Продукт **SAS/Warehouse Administrator** имеет возможности для расширения функциональности и настройки.

Ряд весьма полезных дополнительных средств расширения функциональности имеется в Общем Фонде SAS средств, который пополняется разработками из реальных проектов по всему миру и включает такие полезные средства:

- средство автоматического создания документации Хранилища Данных,
- импорт описания структур данных из разнообразных CASE средств и другие.

- **Загрузка**

Для принятия обоснованных решений необходимо, чтобы доставляемая информация была актуальной и непротиворечивой.

Поэтому организация процесса регулярной загрузки данных в Хранилище является важной задачей.

Выделим основные этапы этого процесса:

- доступ к источникам данных;
- доставка на сервер репозитория Хранилища;
- преобразование данных (унификация, изменение структуры и т.д.);
- проверка на корректность и непротиворечивость, очистка ;
- очистка ;
- агрегирование.

- **Доступ к источникам**

Компания SAS Institute предоставляет возможности для доступа ко всем данным в организации независимо от места и формата хранения.

Доступ осуществляется посредством продукта SAS/ACCESS:

- позволяет работать напрямую с данными из таких известных СУБД как: Oracle, Sybase, Informix, Rdb, DB2, ADABAS, SAP R/2 и R/3, Ingres и другие,
- обеспечивает доступ к стандартным интерфейсам ODBC и OLE DB и файлам со стандартными форматами: VSAM, XLS, DIF, DBF, WKn и др.

Всего система SAS имеет прямой доступ к более чем 40 различным форматам данных на 15 различных платформах.

- **Доставка на сервер**

При реализации Хранилища Данных для организации, автоматизированные системы которой работают на нескольких серверах и даже на разных платформах, процесс транспортировки данных может создавать определенные проблемы.

Система SAS функционирует на многочисленных платформах и имеет хорошо развитые средства межплатформенного общения:

- SAS/CONNECT и SAS/SHARE между такими платформами как: MVS, VM/CMS, различные UNIX, OpenVMS VAX & AXP, OS/2, Window, Windows NT, Macintosh и другие.

- **Преобразование данных**

- Физическая структура Хранилища Данных часто сильно отличается от структуры источников.

- Основной причиной обычно является требования к эффективному исполнению запросов и прогнозируемое время отклика.

- Кроме изменения структуры, при интеграции данных из разных источников необходимо унифицировать форматы представления.

- *Преобразование данных*

В системе SAS разработчику доступны следующие средства обработки и преобразования данных:

- SQL – стандартный язык обработки реляционных данных;
- Data Step – эффективный 4GL язык обработки данных, разработка SAS Institute;
- SAS/IML – язык для работы с матрицами, в виде математической нотации;
- Различные специализированные процедуры обработки данных на основе эффективных алгоритмов сортировки (SORT), преобразования временных рядов (EXPAND), шкалирования (RANK) и пр.

- **Проверка на корректность и очистка**

- Одним из самых важных свойств Хранилища Данных является достоверность доставляемой информации.
- Ричард Хекатрон, один из пионеров концепции, назвал Хранилище Данных как «единый образ истины» для всей организации.
- Поэтому, проверка на непротиворечивость и корректность загружаемых данных, а также очистка и снятие противоречий является ключевым элементом Хранилища Данных.

- *Проверка на корректность и очистка*

Кроме простых и достаточно тривиальных процедур, легко реализуемых с помощью стандартных языков обработки данных, возникает необходимость определять данные, выпадающие из общего набора.

Система SAS включает в себя процедуры:

- ANOVA для анализа дисперсий,
- REG , NREG и LOGISTIC для использования моделей линейной, нелинейной и логистической регрессии,
- MODEL для более сложных моделей,
- а также анализ на основе нейронных сетей.

Эти процедуры входят в состав следующих продуктов: SAS/Base Software, SAS/STAT, SAS/OR и SAS/Enterprise Miner.

- **Агрегирование**

Для обеспечения эффективности отрабатываемых запросов и обеспечения удовлетворительного времени отклика на них, часто используемые агрегированные показатели рассчитываются заранее и включаются в состав Хранилища.

Для агрегирования данных, хранящихся в многомерных базах данных, используется процедура MDDB.

В зависимости от типа производимых вычислений система SAS предоставляет целый спектр процедур агрегирования, входящих в продукты SAS/Base Software, SAS/ETS и SAS/MDDB Server.

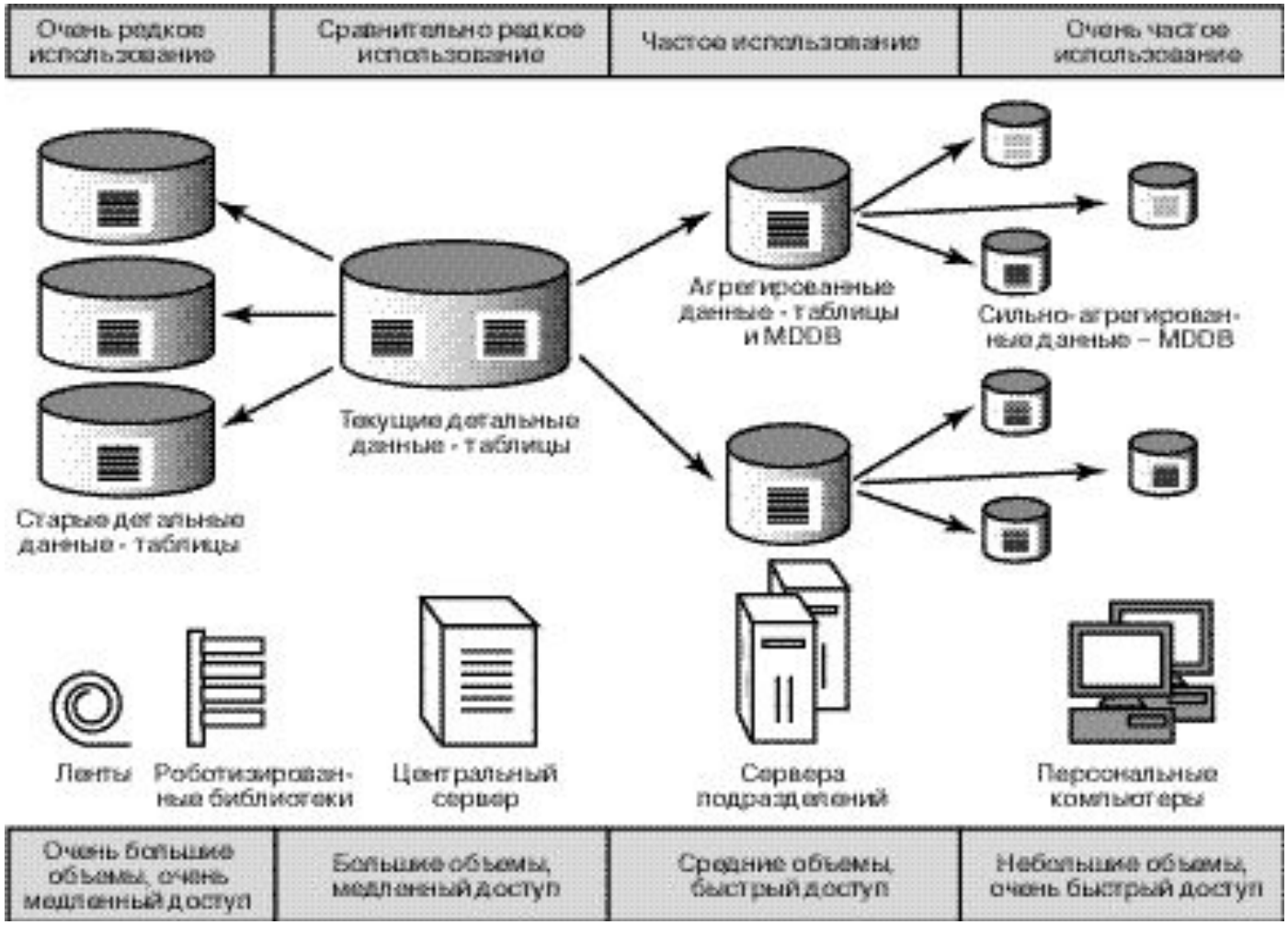
•Хранение.
Структура Репозитория Хранилища

Общая структура репозитория Хранилища Данных является в своем роде отражением главной цели его построения, а именно, максимально полно и быстро удовлетворить потребности пользователей в той или иной информации.

В зависимости от потребностей пользователей в информации можно выделить следующие основные типы:

- персональная информация;
- информация по бизнес темам;
- текущие детальные данные;
- старые детальные данные

Структура репозитория распределенного хранилища данных



- **Персональная информация** – это информация, используемая пользователями со строго определенными обязанностями и информационными потребностями.

Обычно требует большой предварительной обработки или другими словами имеет высокий уровень агрегации (под агрегацией мы будем понимать не только суммирование, но и другие преобразования данных производимых с помощью как аддитивных так и не аддитивных статистик).

Чаще всего хранятся в многомерных базах данных.

- **Информация по бизнес темам** – информация, относящаяся к определенной тематике, такой, например, как финансовая деятельность организации.

Для организаций имеющих, близкие функциональные и организационные структуры ее можно определить как информация для подразделения (например, для финансовой службы).

Имеет более широкий спектр, как в предметных областях, так и во времени, но в то же время напрямую используется реже, чем персонализированная информация.

Обычно хранятся в смешанных структурах: многомерные базы данных и реляционные таблицы.

- **Текущие детальные данные** – самая подробная информация доступная в Хранилище Данных.

Обычными пользователями используется весьма редко, только в случае необходимости сильного уточнения информации.

Обычно является полем деятельности аналитиков по поиску знаний (или поиску скрытых зависимостей в больших объемах информации).

Обычно хранится в реляционных структурах.

- **Старые детальные данные**, – по сути, это тот же самый низкий уровень агрегирования, что и у текущих детальных данных.

Выделяется в особой тип по следующей причине:

- с одной стороны, детальные данные часто требуют больших ресурсов для хранения,
- а с другой детальные данные с возрастом, например, несколько лет необходимы в очень редких случаях.

Решением в данном случае является использование более дешевых и емких способов хранения, например, ленты или роботизированные библиотеки.

Для системы SAS доступ к таким данным не отличается от доступа к таблицам, хранящимся на диске.

Единственное отличие – медленная скорость чтения, что делает реализацию такой схемы особенно легкой и предпочтительной.

Вопросы по лекции

1. Цель создания Хранилища Данных.
2. Назовите основные стадии построения Хранилища Данных.
3. На каком этапе выбираются критерии оценки проекта.
4. Что является результатом сбора требований?
5. Что такое Метаданные? Какая информация хранится в Метаданных?

- Спасибо за внимание