Технологии организации, обработки и хранения статистических данных

ФИО преподавателя: Митина О.А.

e-mail: alogmi@yandex.ru

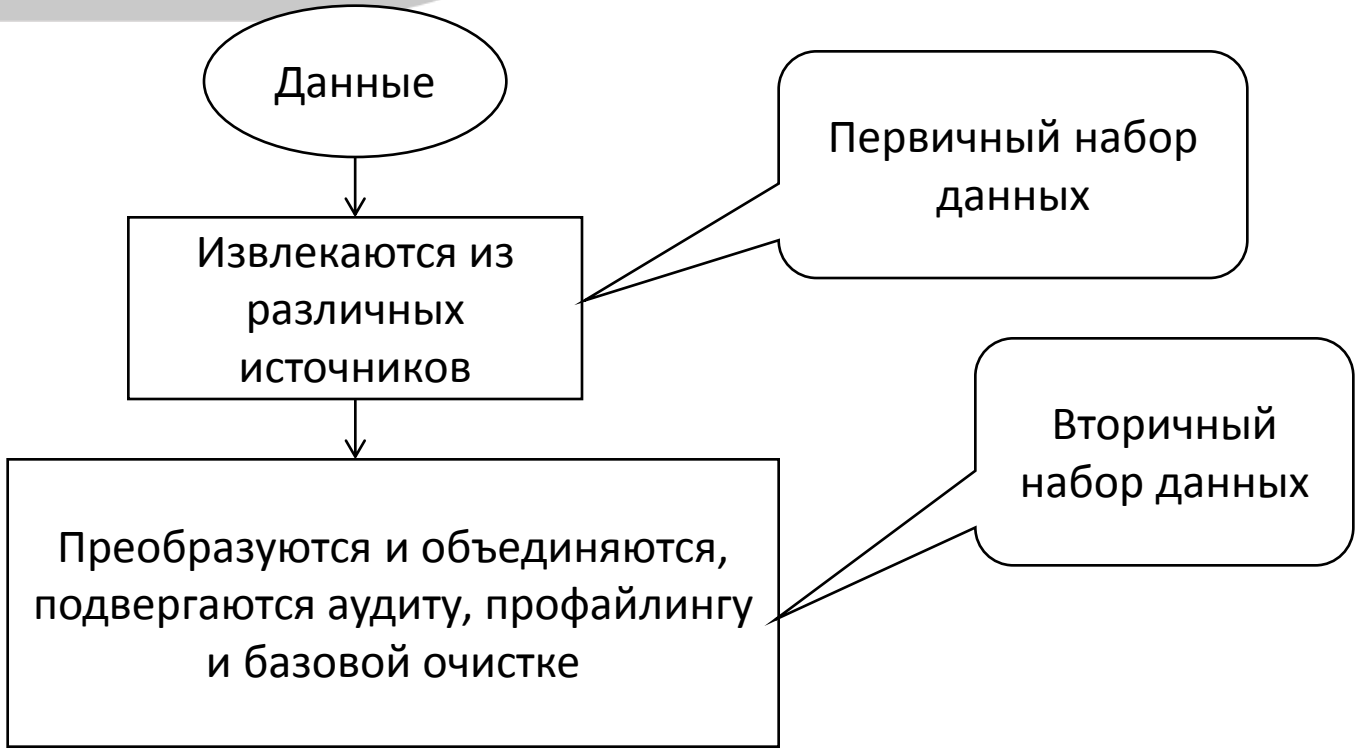
Лекция

Виды источников данных

Условия обучения

- По итогам изучения дисциплины проводится экзамен
- В течение семестра необходимо выполнить все практические работы

Виды источников данных



Поэтапное преобразование данных

Виды источников данных

- первичные;
- вторичные.

По характеру интегрируемых данных:

- фактографические;
- нормативно-справочные;
- метаданные.

Типы корпоративных данных

Типы корпоративных данных

Фактографические – это данные, отражающие факты, которые описывают процессы, объекты и явления предметной области (часто их называют историческими).

Нормативно-справочные (НСИ) (англ.: *reference data)* включают различного рода словари (например, терминологические), справочники (адресов, телефонов), классификаторы (ОКПО, ОКАТО), нормативы, кодификаторы, рубрикаторы и т. д. ***Внешними*** называются такие нормативно-справочные данные, которые содержат информацию, не относящуюся к бизнес-процессам событиям и явлениям, происходящим внутри предприятия, реализующего аналитический проект. ***Внутренними*** называются нормативно-справочные данные, содержащие информацию, циркулирующую внутри компании.

Метаданные. Виды метаданных

Метаданные или «данные о данных» – разновидность данных, носящих служебный характер.

технические – обеспечивают функционирование баз данных, а также выполнение запросов к ним.

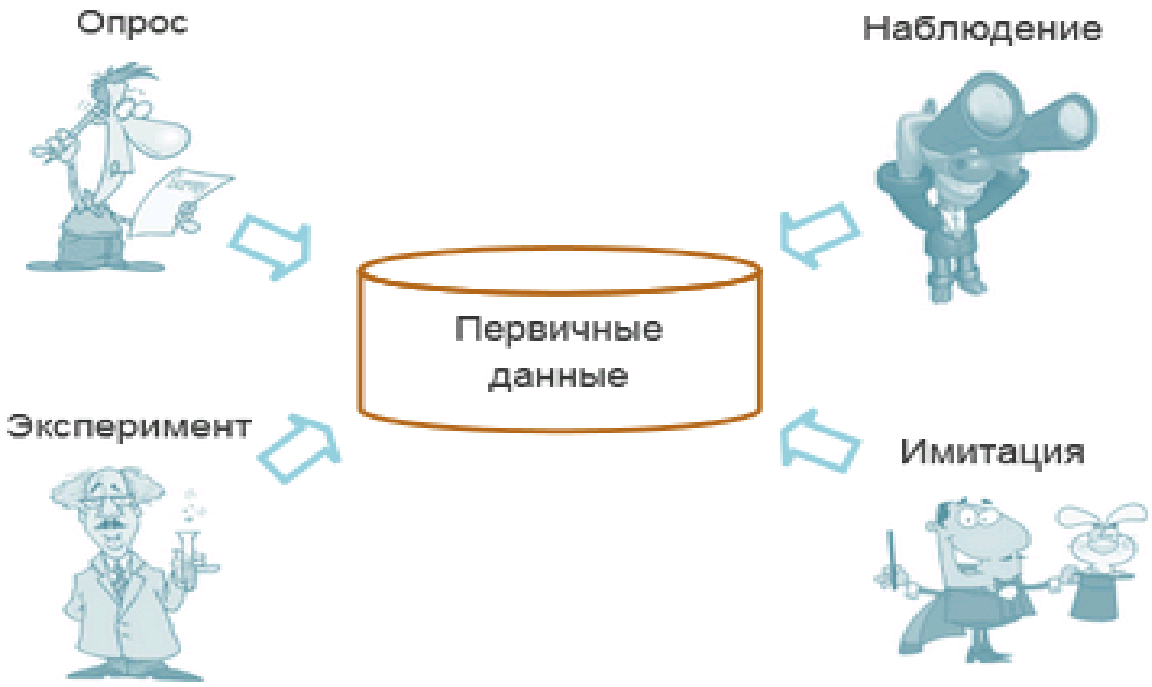
бизнес-метаданные – определяют сущности, хранящиеся в источниках данных, бизнес-термины и определения.

операционные метаданные – содержат информацию о процессе работы источника данных: происхождение загруженных и преобразованных данных, их статус (активные, архивированные или удаленные), статистику использования, сообщения об ошибках и т. д.

Первичные источники данных

Первичными называются источники, данные в которых являются результатом непосредственной регистрации и измерения характеристик бизнес-процессов, объектов и явлений.

Способы сбора первичных данных



Способы сбора первичных данных

Опрос – систематический сбор информации от респондентов посредством личных контактов с ними, по телефону, почте или через Интернет.

Способы сбора первичных данных

Наблюдение – метод, с помощью которого изучают и фиксируют реальное поведение бизнес-процессов, как текущее, так и ретроспективное. Наблюдение может быть открытым или скрытым.

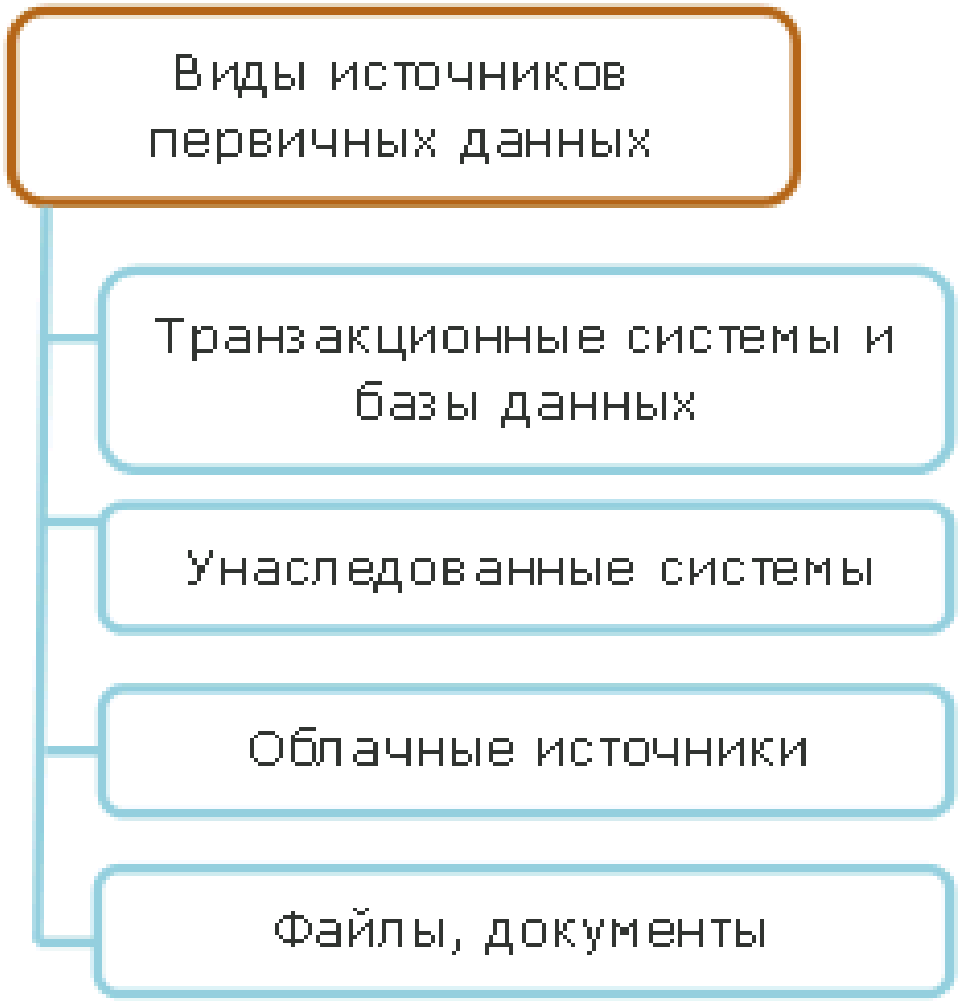
Способы сбора первичных данных

Эксперимент – имеет место тогда, когда один или несколько параметров бизнес-процесса изменяется, а один или несколько других – контролируется.

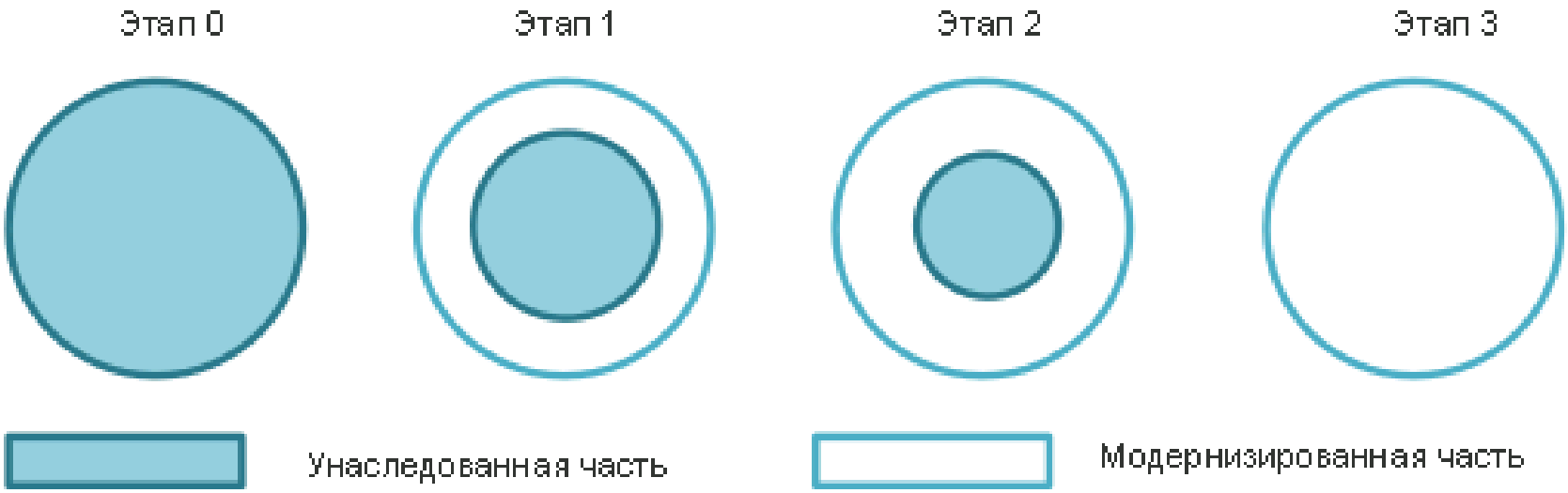
Способы сбора первичных данных

Имитация – метод, основанный на применении компьютерных моделей.

Способы сбора первичных данных



Унаследованные системы



Облачные источники данных

Облачные источники данных

К **преимуществам** относятся:

•отсутствие необходимости приобретать и содержать собственную аппаратную и программную инфраструктуру;

•оплата пользователем только того объема хранения, который реально занимают его данные, а не сервера целиком;

•процедуры по резервированию и сохранению целостности данных производятся провайдером облачного сервиса, освобождая клиента от этих задач.

Облачные источники данных

Недостатки облачного подхода:

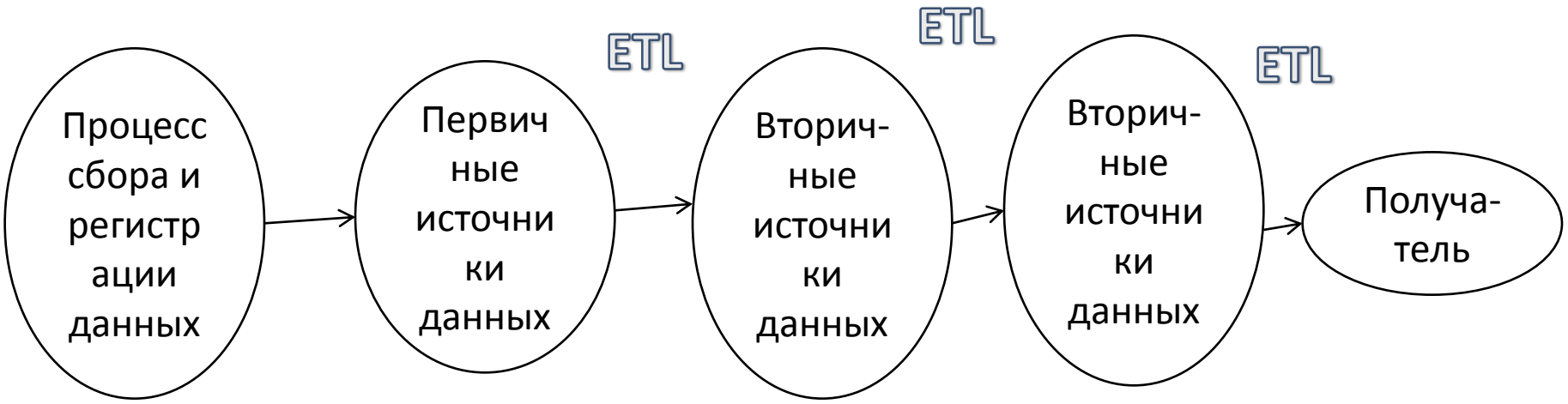
- возникают проблемы безопасности при пересылке данных;
- из-за необходимости пересылки данных и большой нагрузке на облачный сервис, как правило, скорость доступа к данным ниже, чем при использовании локальных источников;
- из-за технических проблем, провайдера облачного сервиса или нарушения работы каналов связи данные могут оказаться недоступными.

Файлы, документы

Вторичные источники данных

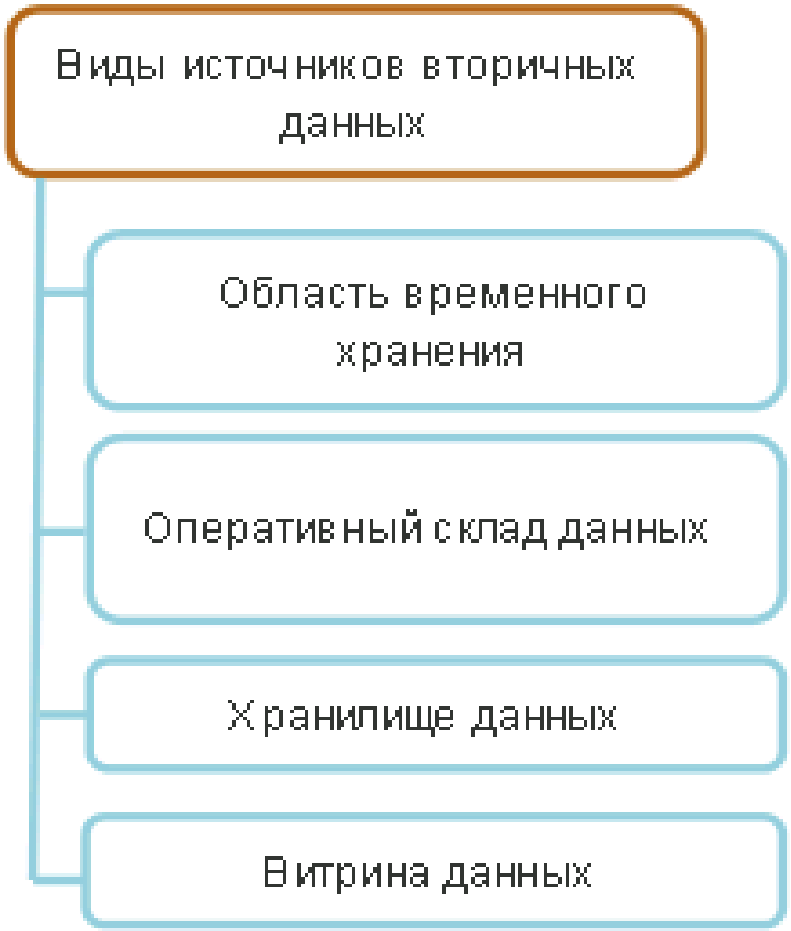
Вторичными являются источники, которые получают данные не в процессе их сбора и регистрации, а из первичных источников.

Вторичные источники данных



ETL (англ.: Extract, Transform, Load – извлечение, преобразование, загрузка) - программно-аппаратный комплекс

Вторичные источники данных



Вторичные источники данных

Область временного хранения (англ.: *Staging Area)* – используются для промежуточной обработки (очистки, трансформации) и синхронизации данных, поступающих из различных источников.

Вторичные источники данных

Оперативный склад данных (англ.: *Operational Data Stone)* – база данных, в которой хранятся оперативные данные – данные реального (или почти реального) времени, используемые для оперативного (тактического) анализа данных с целью поддержки принятия решений.

Синоним – *транспортная база данных.*

Вторичные источники данных

Хранилище данных (англ.: *Data Warehouse,* ХД) – предметно-ориентированный, интегрированный, неизменяемый, хронологический источник данных, специально разработанный для подготовки отчетов и анализа с целью поддержки принятия решений в организации.

Вторичные источники данных

Витрина данных (англ.: *Data Mart)* – массив тематической информации, ориентированный на пользователей одной рабочей группы или подразделения компании.

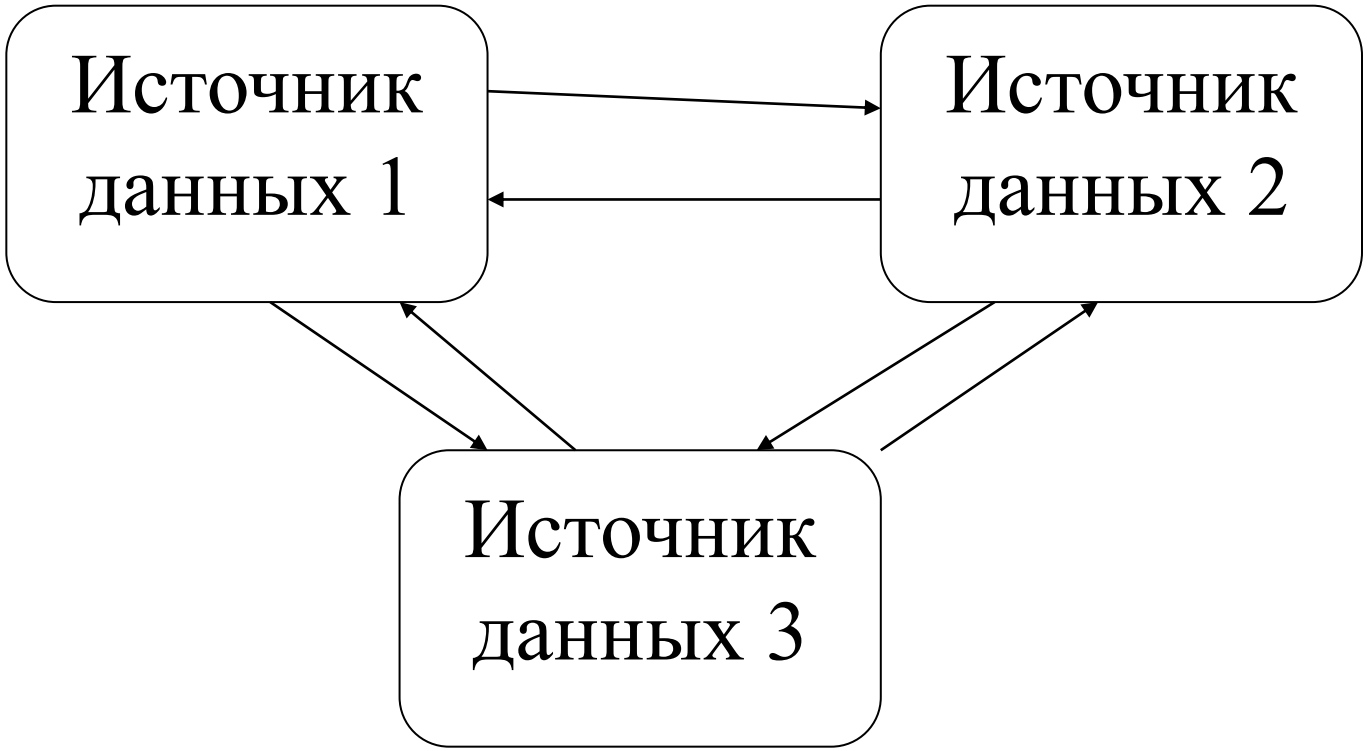
Вторичные источники данных

корпоративное ХД, содержащее все данные организации. В этом случае хранилище является центром информационной инфраструктуры организации, и на него замыкаются основные потоки данных, циркулирующие внутри нее.

Вторичные источники данных

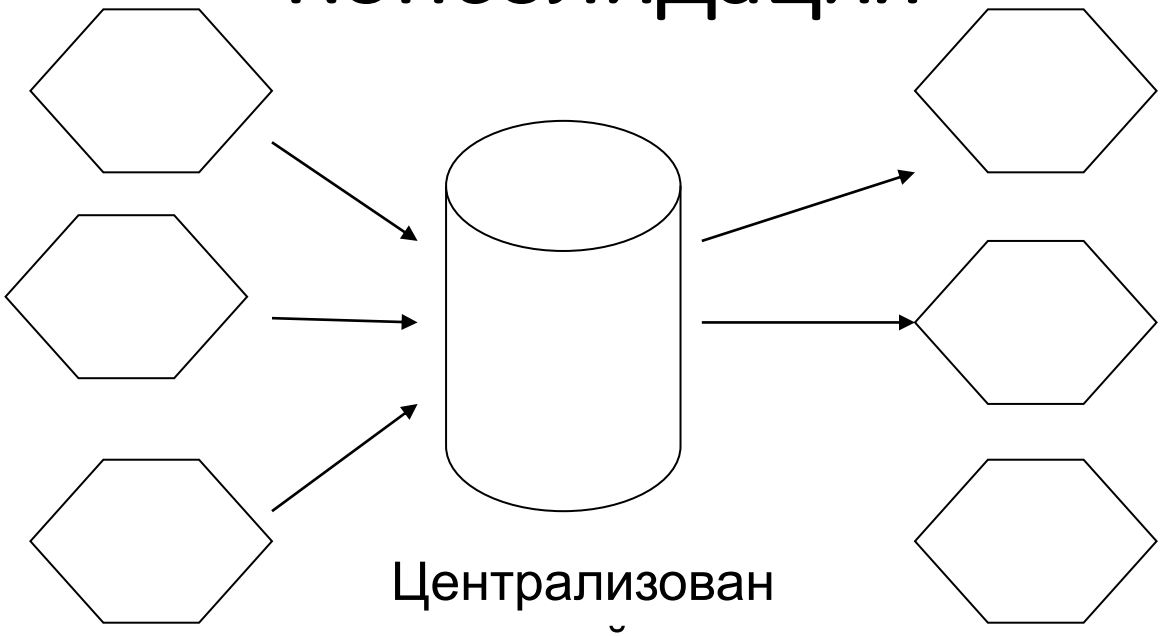
центральное ХД – также содержит все данные организации и замыкает потоки данных, но при этом является доступным всем пользователям внутри компании.

Методы и интеграция данных



Метод «точка-точка»

Консолидация



Консолидация данных

Консолидация

Недостатки подхода:

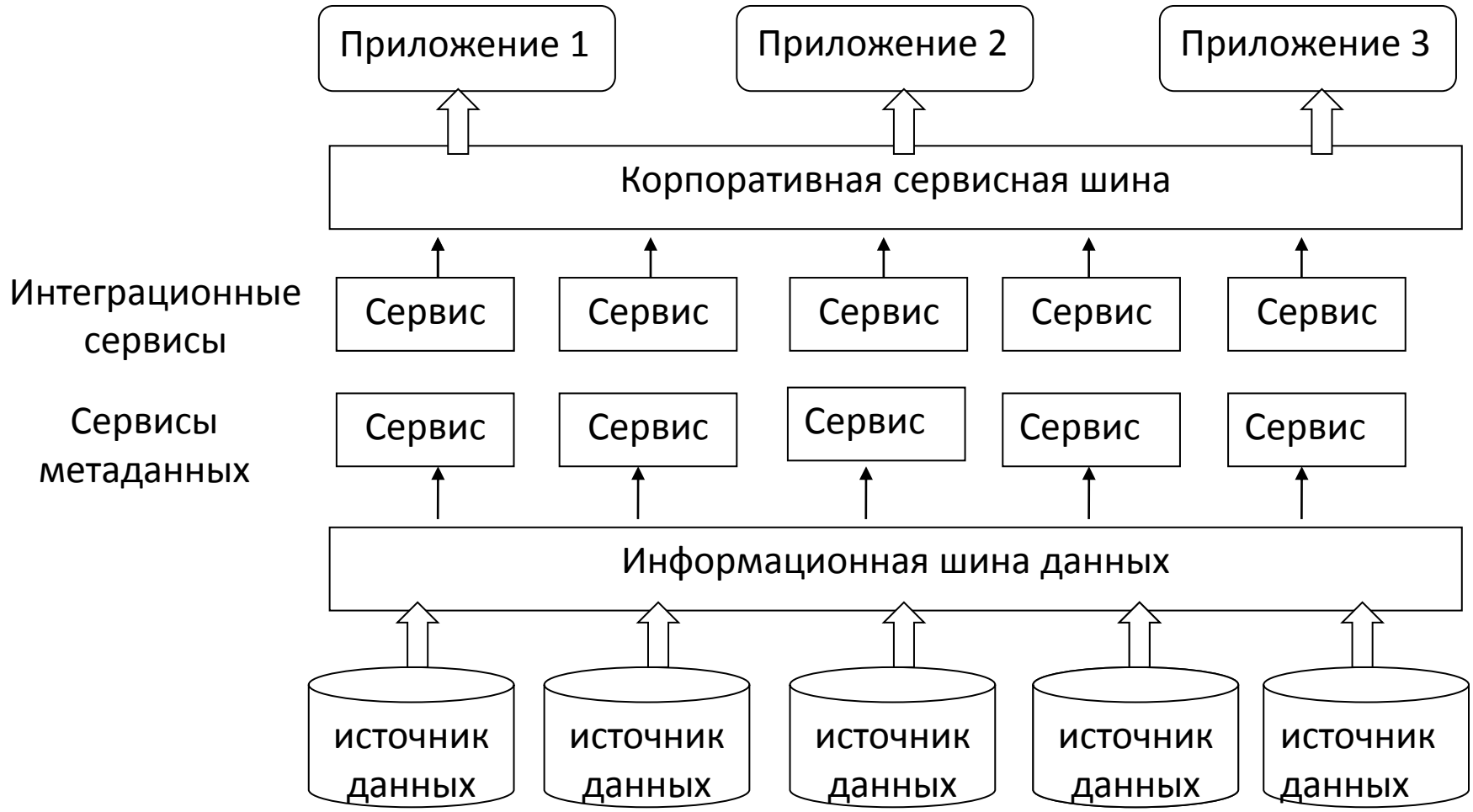
- совместное существование источников и консолидированного ХД удваивает требования к ресурсам дисковой памяти;
- состояния консолидированного ХД практически невозможно синхронизировать с текущим состоянием источников, поскольку данные в ХД всегда будут появляться с задержкой;
- интеграция новых источников данных проблематична, поскольку для нее потребуется изменять весь процесс ETL и структуру метаданных ХД.

Консолидация

Преимущества консолидации:

- физическое наличие ХД повышает устойчивость системы интеграции данных к сбоям и нарушениям в работе оборудования;
- при использовании ХД больше возможностей для поддержания *целостности, непротиворечивости* и *качества данных*.

Сервисный подход (SOA – Service Oriented Architecture)



Сервисный подход к интеграции информационных систем

Сервисный подход к интеграции данных

Преимущества:

- один и тот же сервис может использоваться для различных бизнес-процессов;
- оперативность изменения – изменения в одном сервисе распространяются на все бизнес-процессы, в которых этот сервис использовался;
- масштабируемость – благодаря возможности распределения вычислительной нагрузки сервисно-ориентированные системы более устойчивы к пиковым нагрузкам.

Сервисный подход к интеграции данных

Недостатки:

- высокая сложность интеграции с внешними приложениями, не предоставляющими сервисов для доступа к данным;
- для эффективного информационного взаимодействия все ресурсы, входящие в состав схемы, должны быть хорошо структурированными;
- некоторые реализации SOA накладывают существенные ограничения на объемы передаваемых данных.

Список литературы

- Тюрин Ю.Н. Анализ данных на компьютере / Ю.Н. Тюрин, А.А. Макаров. – М.: МЦНМО, 2016. – 368 с.
- Мхитарян В.С. Анализ данных: учебник для академического бакалавриата / под ред. В.С. Мхитаряна. – М.: Изд. Юрайт, 2017 – 490 с.
- Хрусталёв Е.М. Агрегация данных в OLAP-кубах. http :// www . olap . ru /

Темы дисциплины

1 Анализ данных. Основные понятия и определения

2 Бизнес-аналитика. Основные понятия и определения

3 Методология CRISP-DM

4 Интеграция данных

5 Виды источников данных

6 Компоненты корпоративной информационной фабрики

7 Структура CIF

8 Базовые архитектуры CIF

9 Базовые архитектуры CIF

10 Хранилища данных

11 Витрины данных

12 Визуализация

13 Технологии бизнес-аналитики

14 Прикладные задачи бизнес-аналитики

15 OLAP-технологии

16 Аналитические платформы

Спасибо за внимание!