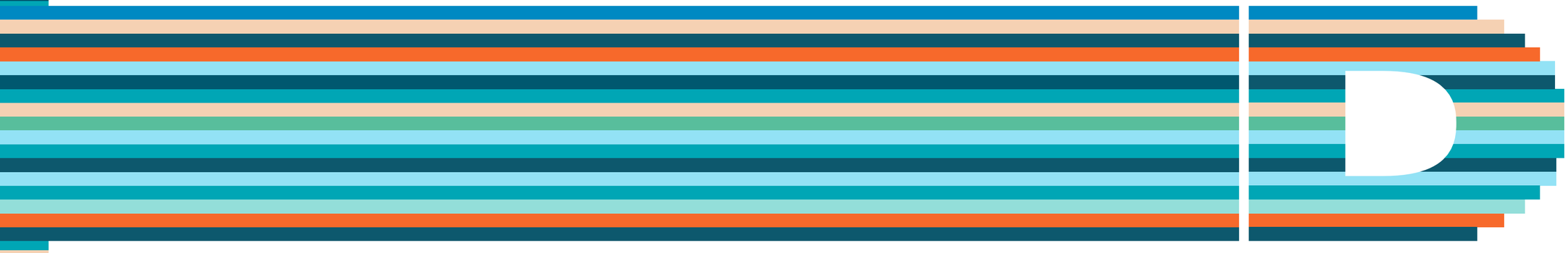
Лекция 4.

АРХИТЕКТУРА СИСТЕМ ХРАНЕНИЯ И ОБРАБОТКИ БОЛЬШИХ ДАННЫХ



ТЗОБД

К.т.н., доцент Зиновьева В.В.

Вопросы

1. **Классификация Больших данных**
2. **Алгоритм обработки Больших данных**
3. **Архитектура хранилищ данных**
4. **Системы хранения данных**

1. Классификация Big Data

- ✓ **официальные данные**. Информация, распространяемая государственными органами (заявления, пресс-релизы, прогноз погоды, сведения о планах муниципального развития).
- ✓ **операционные данные**. Это данные о клиентах, поставщиках, партнерах и сотрудниках, доступные в процессе онлайн-обработки транзакций и/или полученные из онлайн-базы данных аналитической обработки
- ✓ **коммерческие данные**. Коммерчески ценная информации.
- ✓ **информация из социальных сетей и сервисов**.
- ✓ **«темные» данные**. Информация, которая не хранится или не собирается организациями специально, а формируется случайно (попутно) в процессе ведения бизнеса или взаимодействия с сетевыми сервисами и остается в Интернет-архивах

Каждый из указанных выше источников данных обладает определенной ценностью, зависящей, от его достоверности и полноты

2. Алгоритм работы с Big Data

1. **Сбор данных.** Разнообразие способов сбора данных напрямую зависит от их источника. Также влияние оказывает и природа информации, подлежащей сбору и последующему анализу. Механизмы сбора можно классифицировать следующим образом:

- сбор структурированных данных (различные базы данных);
- сбор неструктурированных данных (данные GPS, аудио- и видеофайлы аналоговые источники информации и т.д.);
- сбор частично структурированных данных (данные журналов событий внутренних систем, сетевых служб, XML-данные и т.д.).

2. **Хранилище данных**. Все собранные данные распределяются на хранение и, в зависимости от типа данных, оказываются в распределенных/нераспределенных хранилищах или фиксируются в отдельных журналах записи событий.

Алгоритм работы с Big Data

3. **Преобразование данных**. Перед передачей данных на стадию обработки они должны быть преобразованы в понятный для программ формат с помощью инструментов импорта/экспорта.

4. **Обработка данных**. На данном этапе происходит объединение всех собранных данных. Обработка может проходить пакетами или режиме реального времени

3. **Анализ данных**. Инструменты, используемые на данном этапе, зависят от целей пользователя

4. **Вывод данных.** Результаты анализа должны быть представлены в формате, удобном для восприятия пользователем.

Архитектура хранилищ данных

Хранилище данных – это система, в которой собраны данные из различных источников внутри компании, которые используются для поддержки принятия управленческих решений.

Основная задача организации хранилищ данных – создание хорошо спроектированного централизованного банка данных.

Основное *преимущество* хранилища данных – это сокращение времени выполнения проекта.

Для описания стандартных процессов и инструментов для сопоставления, объединения и перемещения данных между базами используется термин *ETL (Extract, Transform, Load)* – извлечение, преобразование, загрузка.

Типичные операции, выполняемые в хранилище данных **OLAP -online analytical processing.**

Подходы к проектированию хранилищ

1. Подход «снизу вверх» Кимбалла.

Основывается на важности **витрин данных**, которые являются хранилищами данных, принадлежащих **конкретным направлениям бизнеса.**

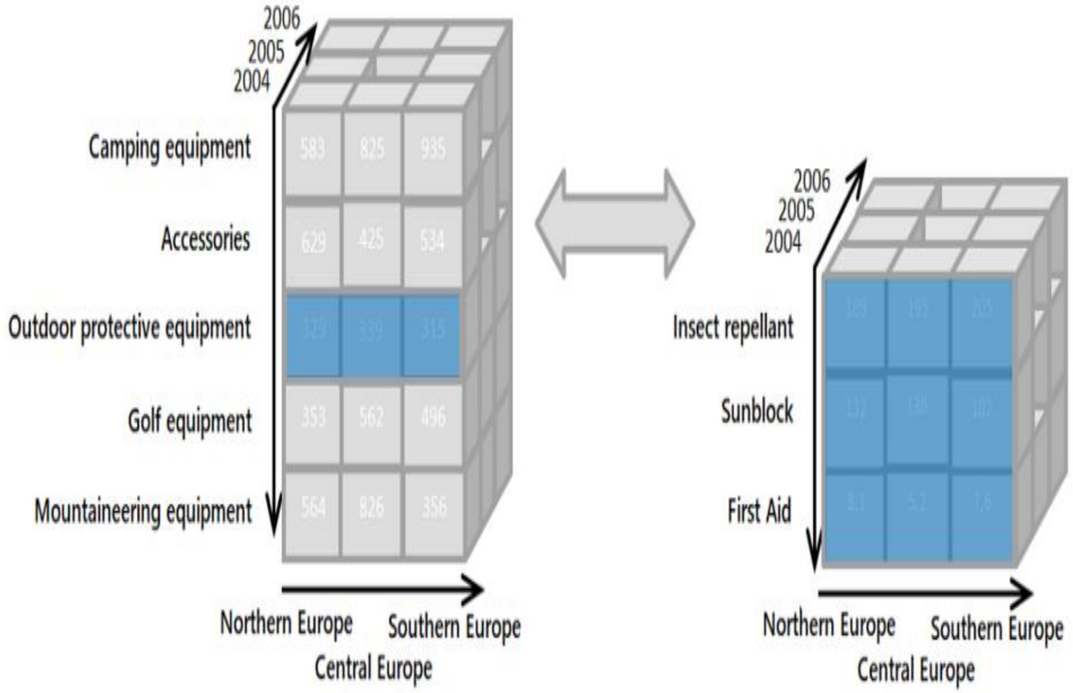
Хранилище данных по Кимбаллу – это сочетание различных витрин данных, которые облегчают отчетность и анализ данных.

Витрина данных (data mart) — это хранилище данных, предназначенное для **определенного круга пользователей в компании** (отдел маркетинга, производственный отдел и т.д.)

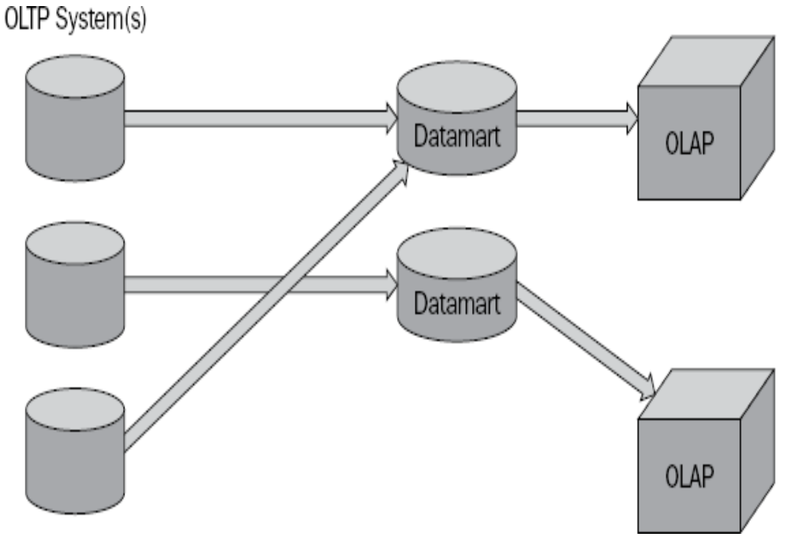
Витрины данных часто работают в реальном режиме времени.

Подход Кимбалла

```
SELECT SUM(price) as total_sales,
      region,
      store,
      product
FROM sales
GROUP BY CUBE(region, store, product);
```



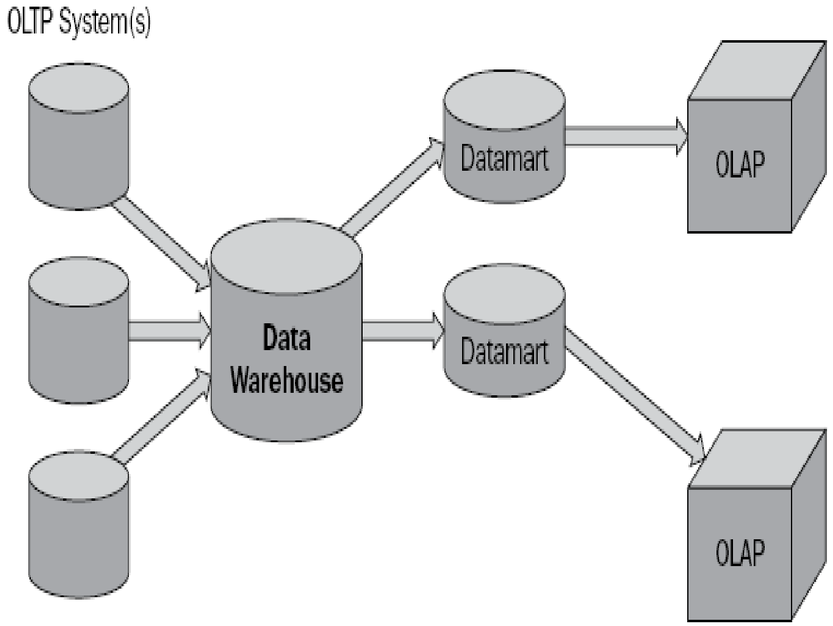
Используется *двухуровневая* архитектура. Она предполагает построение витрин данных без создания центрального хранилища, информация поступает из регистрирующих систем (*OLTP*- online Transaction Processing), транзакционная ***система*** — обработка транзакций в реальном времени) и ограничена конкретной предметной областью



Подходы к проектированию хранилищ
Нисходящий подход Инмона

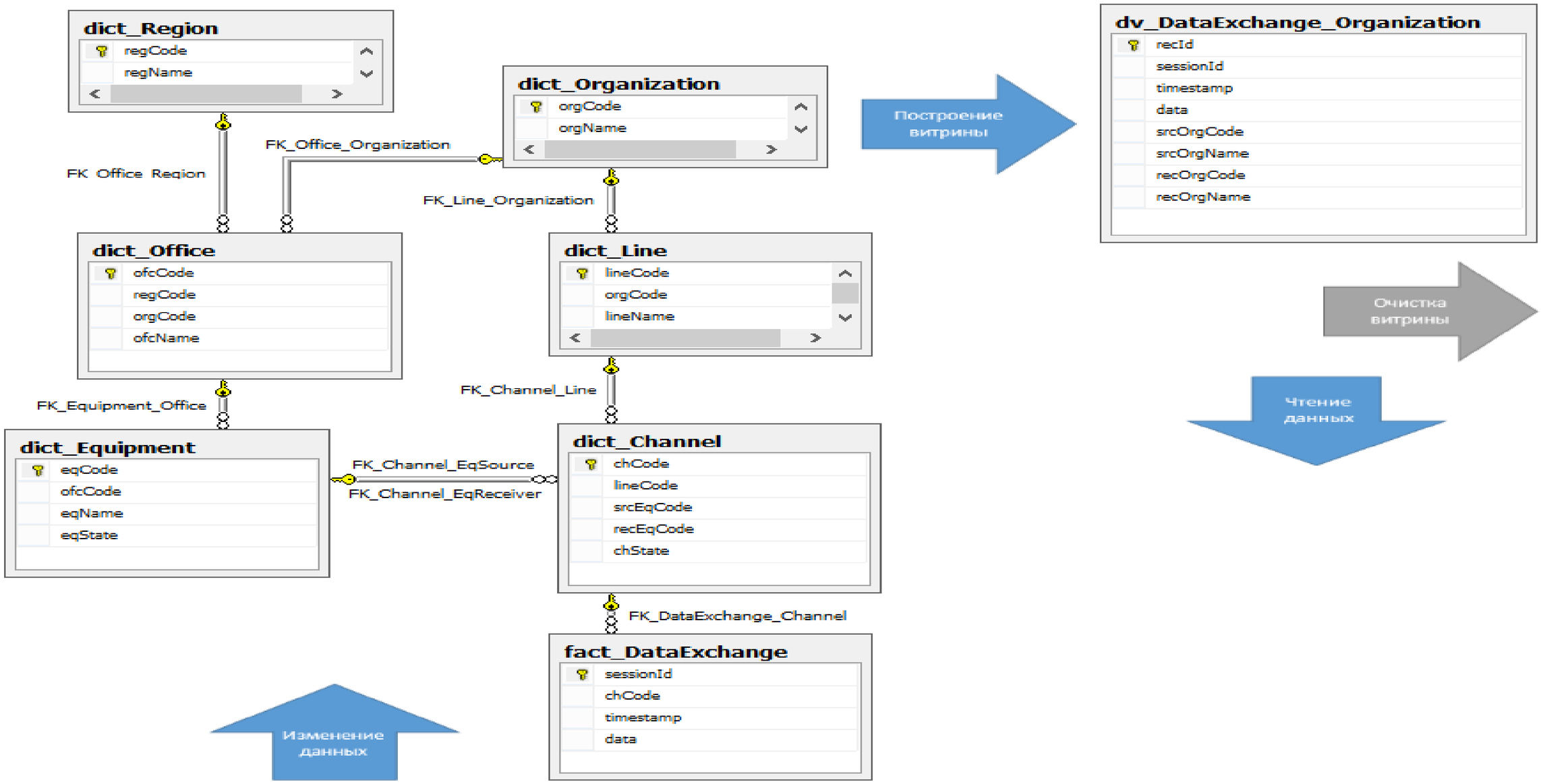
1. Он основывается на том, что хранилище данных является централизованным хранилищем всех корпоративных данных.

2. При таком подходе организация сначала создает нормализованную модель хранилища данных, а затем создаются витрины размерных данных на основе модели хранилища



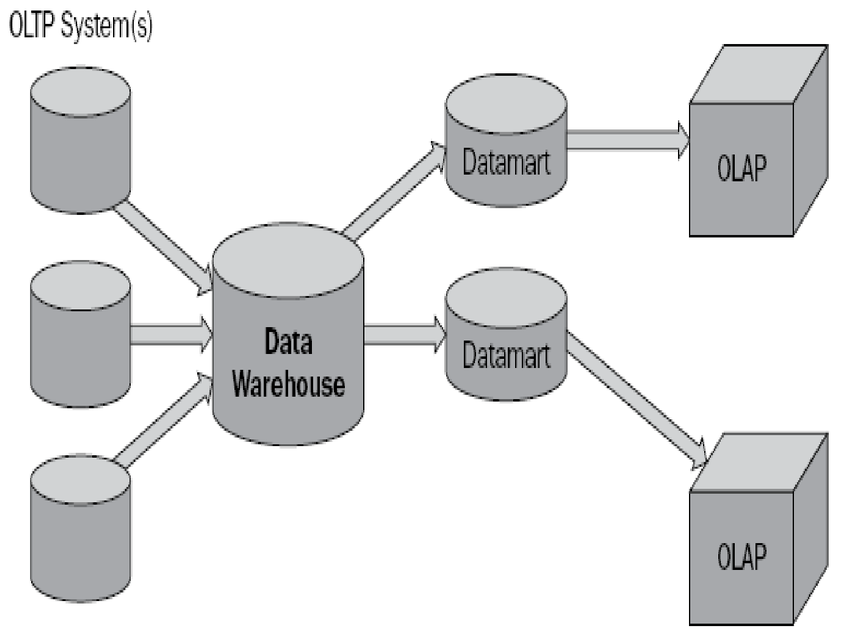
Подход Инмона – полноценное корпоративного хранилища данных (*Data Warehouse*) выполняется в *трехуровневой* архитектуре

Витрина данных из общей реляционной БД



Подходы к проектированию хранилищ
Нисходящий подход Инмона

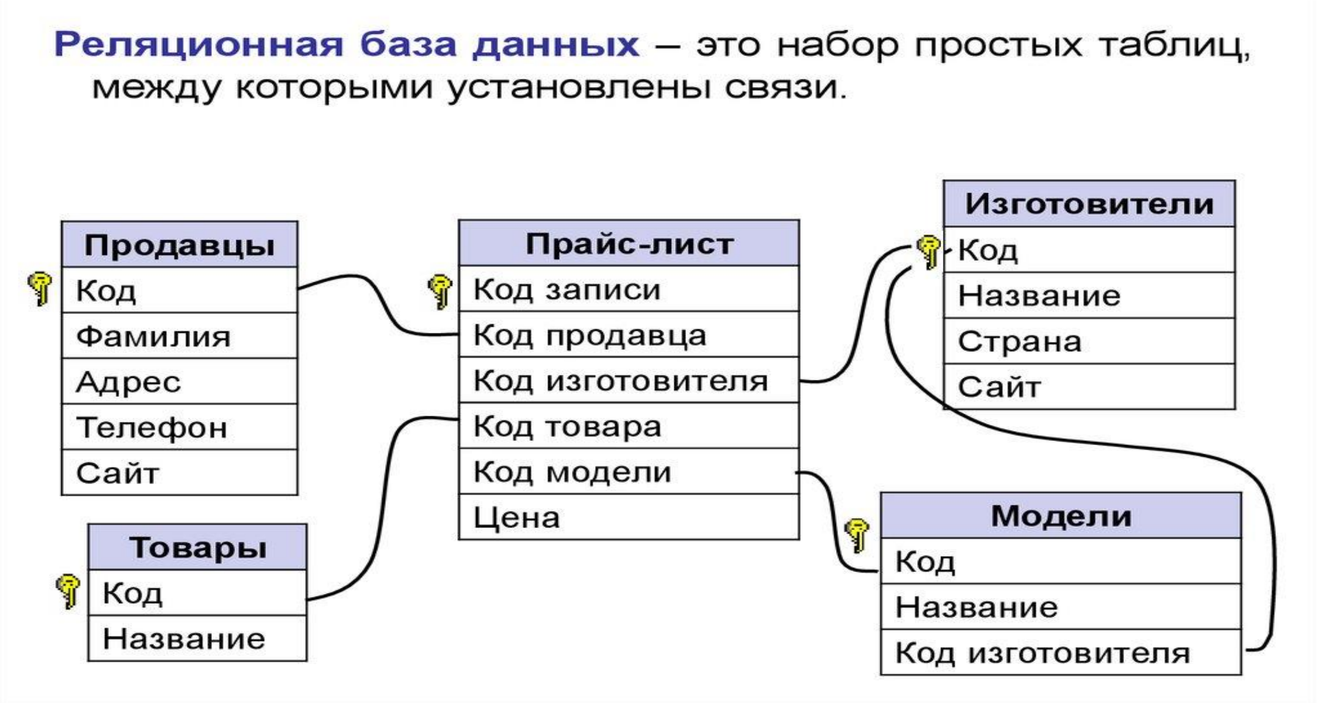
1. **На первом** (нижнем) уровне расположены разнообразные источники данных: внутренние регистрирующие системы, справочные системы, внешние источники + сервер БД, используемый для извлечения данных
2. **Второй (средний) уровень** содержит центральное хранилище, куда стекается информация от всех источников с первого уровня, и, возможно, оперативный склад данных + сервер OLAP, который преобразует данные в структуру, наиболее подходящую для анализа и сложных запросов
3. **Третий (верхний) уровень** представляет собой набор предметноориентированных витрин данных, источником информации для которых является центральное хранилище данных



Архитектуры хранилищ

Различают **два архитектурных направления** построения хранилищ: **нормализованные хранилища** данных и **хранилища с измерениями**.

В **нормализованных хранилищах данные** находятся в предметно ориентированных таблицах – принцип реляционных БД.

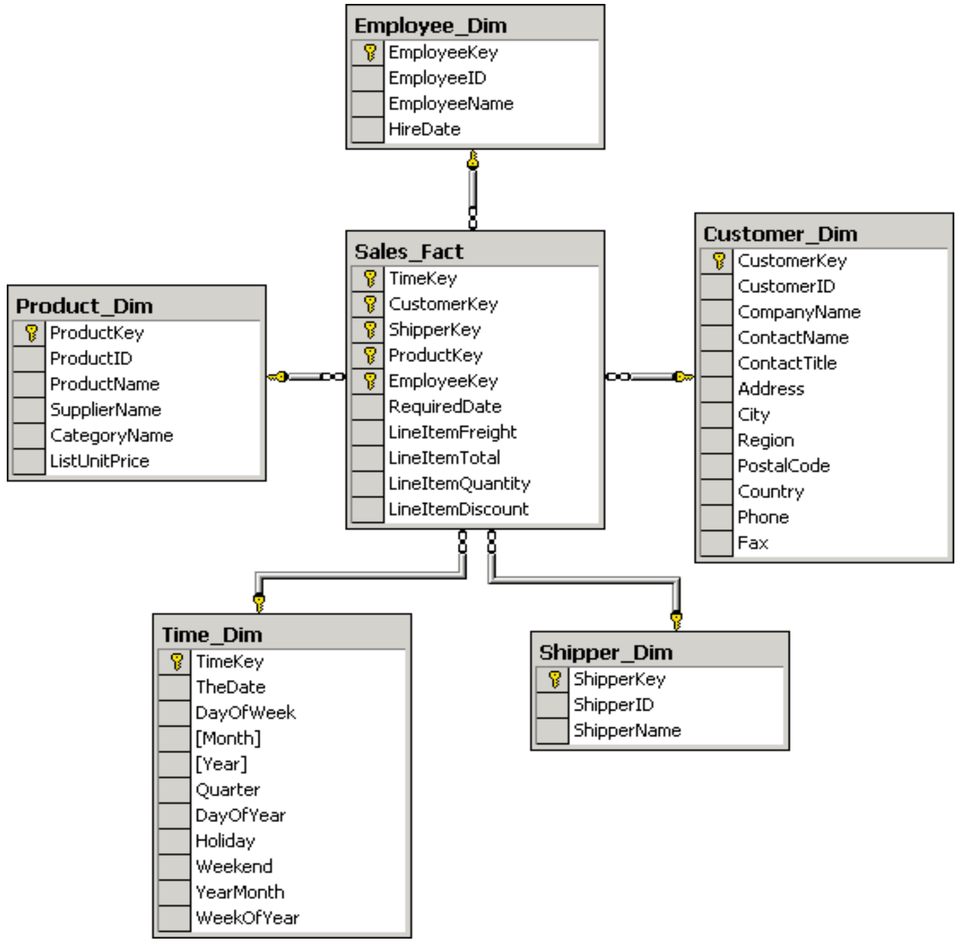


Архитектуры хранилищ. Звезда

Хранилища с измерениями используют разные типы схем хранения, такие как «звезда» и «снежинка».

Новый тип БД куб данных (ООП БД, noSQL БД)

Одно измерение куба может содержаться как в одной таблице так и в нескольких связанных таблицах, соответствующих различным уровням иерархии в измерении. Если каждое измерение содержится в одной таблице, такая схема хранилища данных носит название *«звезда» (star schema)*.



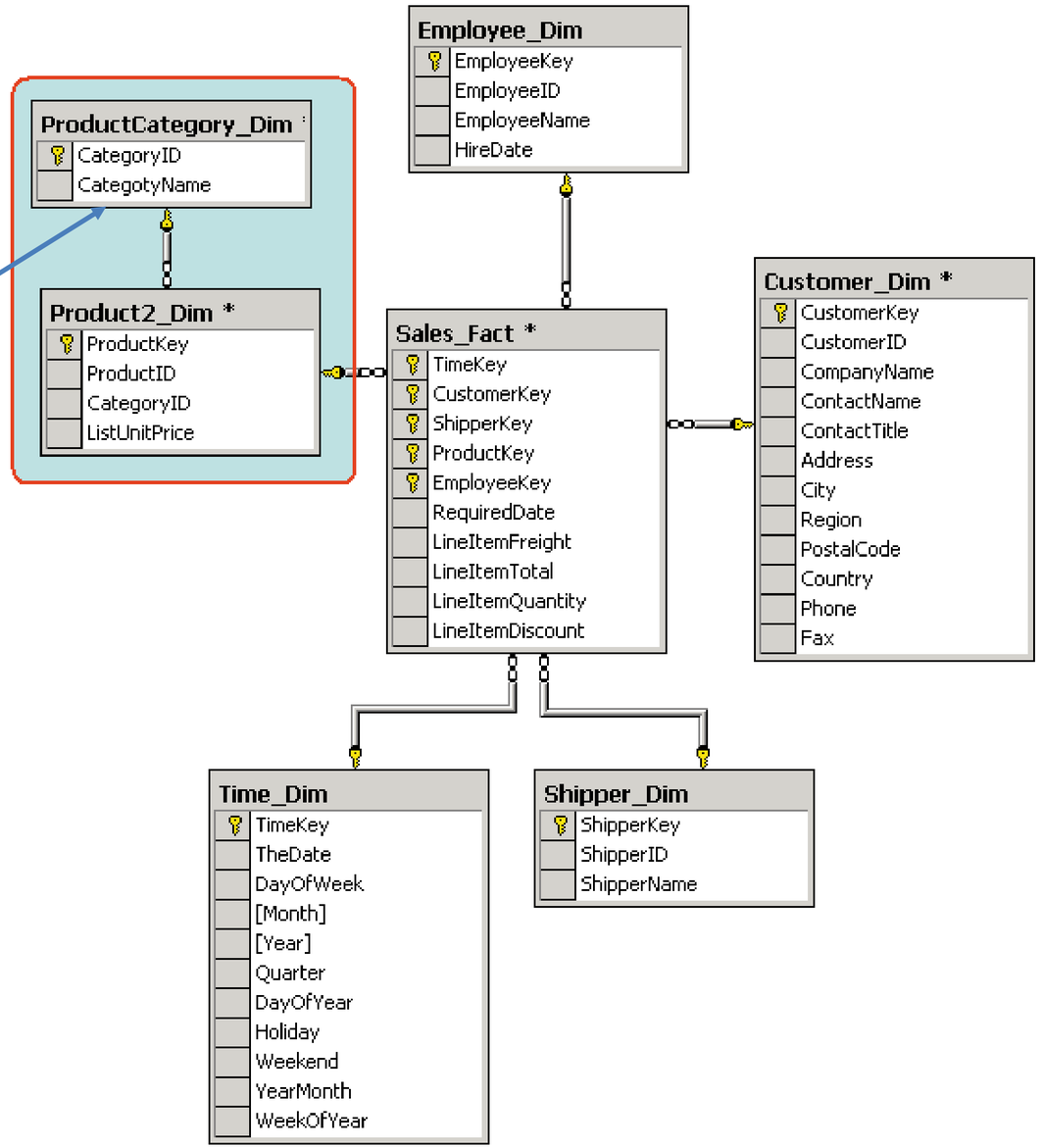
В центре схемы «звезда» находятся данные (таблица фактов), а измерения образуют ее лучи. Таблица фактов содержит агрегированные данные, которые будут использоваться для составления отчетов, а таблица измерений описывает хранимые данные.

Достаточно простая конструкция звездообразной схемы значительно упрощает написание сложных запросов.

Архитектуры хранилищ. Снежинка

Схема типа «снежинка» отличается тем, что использует **нормализованные данные**. Что означает эффективную организацию данных, каждая таблица содержала минимум избыточности.

Таким образом, таблицы измерений разветвляются на отдельные таблицы измерений.



Основной недостаток – сложность запросов, необходимых для доступа к данным: каждый запрос должен пройти несколько соединений таблиц, чтобы получить соответствующие данные.

Способы загрузки данных в хранилища

ETL (*Extract, Transform, Load****)*** – сначала извлекают данные из пула источников данных. Данные хранятся во временной промежуточной БД. Затем выполняются операции преобразования, чтобы структурировать и преобразовать данные в подходящую форму для целевой системы хранилища данных. Затем структурированные данные загружаются в хранилище и после этого становятся готовы к анализу.

ELT (*Extract, Load, Transform*) – данные сразу же загружаются после извлечения из исходных пулов данных. Промежуточная БД отсутствует, что означает, что данные немедленно загружаются в единый централизованный репозиторий. Данные преобразуются в системе хранилища данных для их использования с инструментами бизнес-аналитики и аналитики

4. Системы хранения данных

1. ***Традиционный подход основан на использовании системы***

SAN (Storage Area Network) для структурированных данных.
Первичные данные хранятся в виде блоков в дата-центре. Функции блочного хранения используются на низких уровнях в виде блоков фиксированного размера, которые легко индексируются и находятся в системе хранения. Подходи для небольших БД.

2. **Горизонтально-масштабируемые** *(Scale-out)* **файловые системы**, такие как ***HDFS*** (*Hadoop Distributed File System*). НО поддержка таких систем трудоемка+ механизм репликации увеличивает объем вдвое.

3. **Облачные хранилища.** *HyperText Transfer Protocol* (*HTTP*) – протокол передачи данных, предназначенный для передачи гипертекстовых документов содержbn ссылки, позволяющие организовать переход к другим документам. **НО** многие облачные провайдеры берут плату не только за объем хранимых данных, но и за трафик и число обращений к хранилищу. Итог цена может быть сравнима с HDFS.

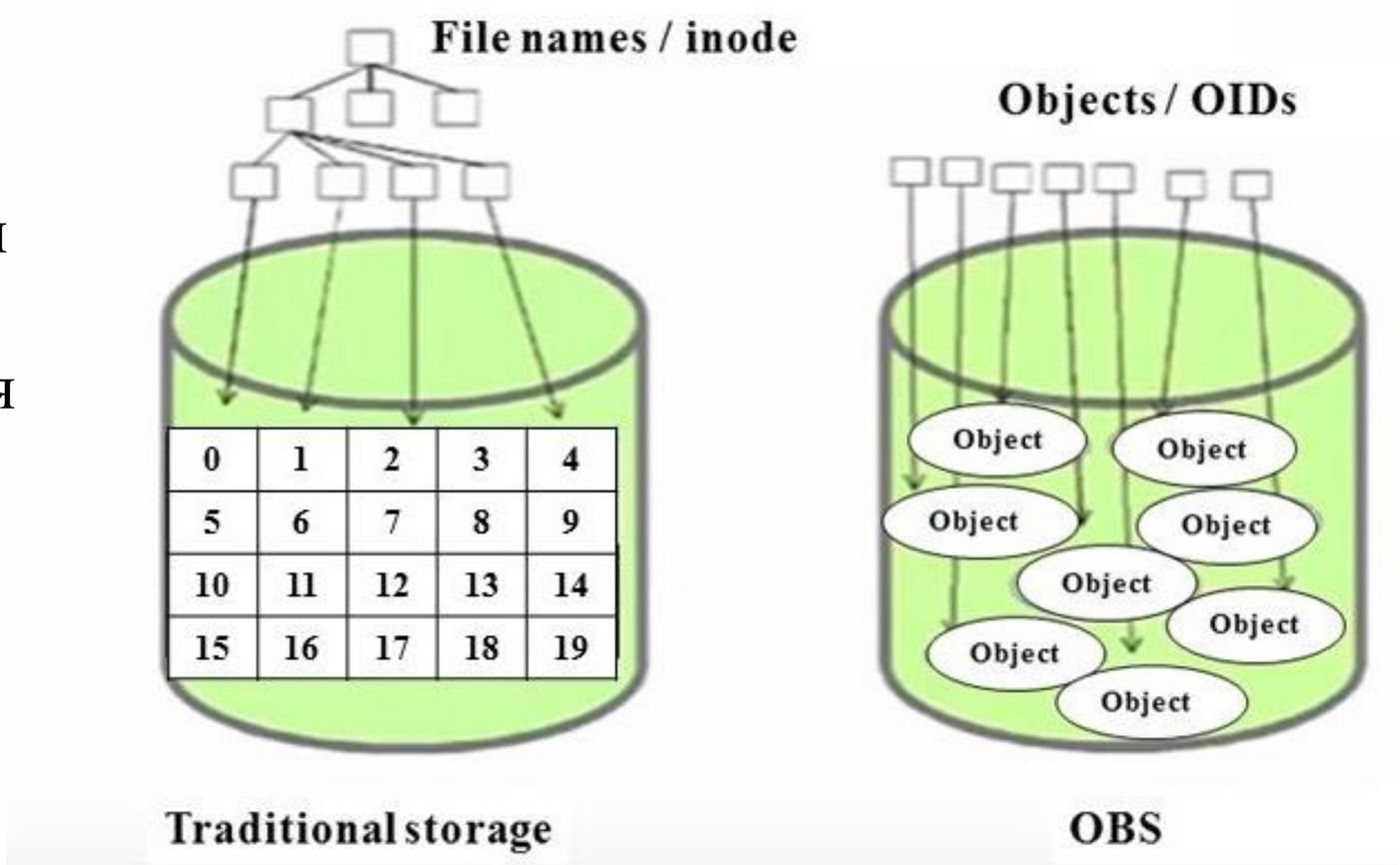
Особенности объектных систем хранения

1. данные хранятся как объекты, неструктурированы и могут включать в себя самые разные форматы и размеры файлов
2. в объектных системах хранения используется простой список объектов, хранящихся в «пакетах» (*buckets*)
3. вместо имен - идентификаторы;
4. объекты хранятся вместе с определенными пользователем метаданными, что облегчает поиск объектов

Системы хранения данных

4. Объектная система хранения (*object storage*). Используются:
HTTP – протокол передачи данных, что и в облаке.
Application Programing Interface (*API*) – интерфейс прикладного программирования

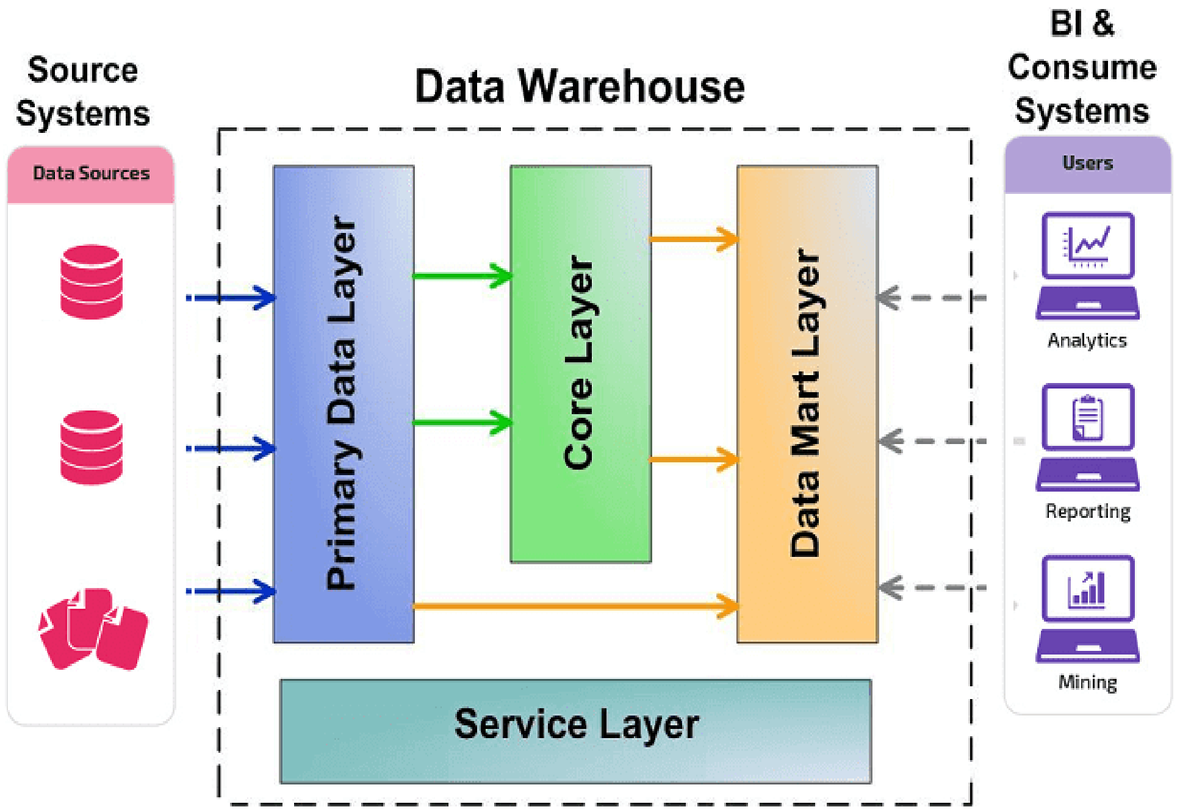
1. На верхнем уровне размещаются средства управления.
2. Файлы различного вида хранятся как «объекты», которые имеют атрибуты- метаданные
3. Файлами управляет локальная файловых система



Пример архитектуры корпоративной системы хранилища – DWH

DWH – предметно-ориентированные БД для консолидированной подготовки отчетов, интегрированного бизнес-анализа и оптимального принятия управленческих решений на основе полной информационной картины. Архитектура *DWH* – многоуровневая, слоеная

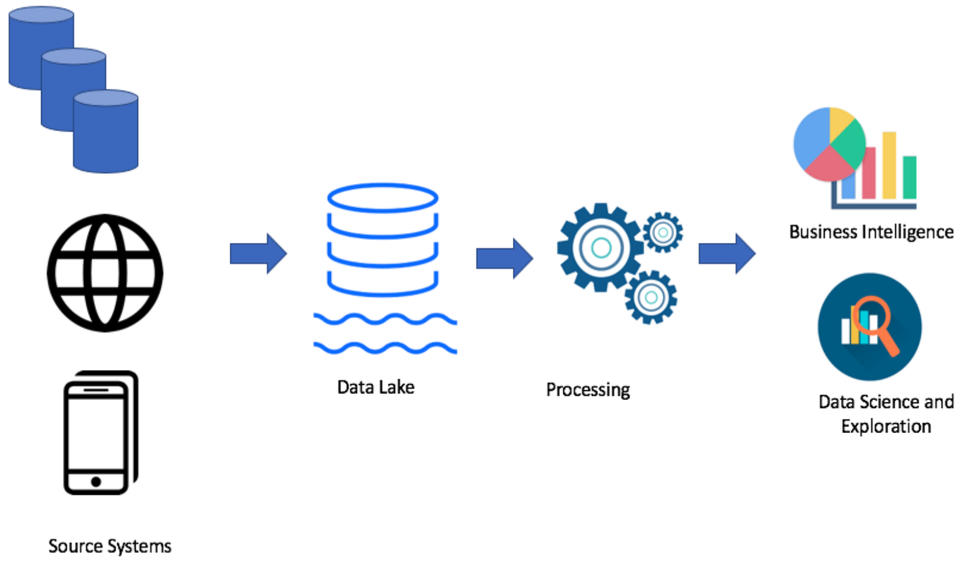
1. **операционный слой первичных данных** (*Primary Data Layer*, или стейджинг), на котором выполняется загрузка информации
2. **ядро хранилища** (*Core Data Layer*) – центральный компонент, который выполняет консолидацию данных из разных источников
3. **аналитические витрины** (*Data Mart Layer*), где данные преобразуются в структуры, удобные для анализа и использования
4. **сервисный слой** (*Service Layer*) обеспечивает управление всеми вышеописанными уровнями.



Озеро данных

Озеро данных- файловое хранилище всех типов сырых данных, которые доступны для анализа кем-угодно в организации

1. Собирайте и храните все что угодно — озеро данных содержит все данные, как сырые необработанные данные за любой период времени, так и очищенные данные.
2. Глубокий анализ — озеро данных позволяет пользователям исследовать и анализировать данные.
3. Гибкий доступ — озеро данных обеспечивает гибкий доступ для различных данных и различных сценариев.
4. НО требуются навыки для извлечения данных и контроль за вносимыми данными



Проблемы в использовании озера данных

Чтобы озеро не стало «болотом», нужно наладить в и процесс управления данными (*Data governance*).

Главная составляющая этого процесса – определение достоверности и качества данных еще до загрузки в озеро

К 2018 году 90% внедренных озер данных будут бесполезны потому что они будут переполнены информацией, собранной неизвестно с какой целью. (Gartner, Strategic Planning Assumption, Gartner BI Summit, 2015).

Данные в озере могут быть неконсистентны и не иметь метаданных, поэтому реально только очень опытные аналитики, хорошо знающие контекст, смогут сливать и согласовывать данные из разных источников.

Угрозы безопасности хранилищ

Внешние угрозы:

✓Национальные государства.
✓Террористы.
✓Хакеры, кибермошенники, организованные преступные группировки.
✓Конкуренты, занимающиеся промышленным шпионажем.

Внутренние угрозы:

✓Инсайдеры с нечистоплотными помыслами.
✓Плохо обученный или безответственный персонал.
✓Недовольные сотрудники.

Другие угрозы:

✓Пожары, наводнения и другие катастрофы природного характера.
✓Перебои с электроэнергией.

Защита хранилищ

Стандарт ISO/IEC 27040 в области безопасности информационных хранилищ предусматривает использование физических, технических и административных мер для защиты систем хранения и инфраструктуры вместе с хранимой информацией.

Физический уровень

Меры безопасности на физическом уровне предусматривают защиту инфраструктуры и данных от физического неправомерного доступа и могут включать:

- ✓Найм персонала для мониторинга дата центров и хранилищ.
- ✓Closed Circuit Television - Система телевидения замкнутого контура с сохранением видео.
- ✓Использование систем доступа на базе биометрии / смарт-карт и турникетов с защитой от проникновения нескольких лиц одновременно и обратного хода, которые разрешают проходить только одному человеку после аутентификации.
- ✓Мониторинг внутреннего пространства при помощи датчиков температуры и дыма.
- ✓Использование альтернативных источников питания (например, запасного генератора).

Защита хранилищ

Технический уровень

Аутентификация пользователей и контроль доступа:

- Изменение всех стандартных учетных записей.
- Избегание совместного использования учетных записей, отследить которые сложно или невозможно.
- Назначение ровно таких прав, которые нужны для выполнения роли.
- Изменение или снятие прав при увольнении или смены роли пользователя.

Анализ трафика: при помощи приложений для поведенческого анализа (user and entity behavior analytics; UEBA), которые все чаще становятся частью решений SIEM (security information and event management).

- **Защита интерфейсов управления.**
- **Надежное шифрование**
- **Защита конечных узлов: настольных компьютеров,**
- **Специальные меры для защиты баз данных**

Защита хранилищ

Административный уровень

Административные меры сводятся к трем «П»:

Политика, Планирование, Процедуры

Рекомендации

Документирование хранения наиболее важных и критичных для бизнеса категорий информации и требований к защите.

Внедрение мер сохранности и защиты данных в целом.

Внедрение мер по удалению информации и утилизации носителей.

АУДИТ- Проверка на соответствие, что все элементы инфраструктуры хранения соответствуют политикам безопасности.

Источники

Источники

1. Big Data = Большие данные : учеб. пособие / И. Б. Тесленко [и др.] ; Владим. гос. ун-т им. А. Г. и Н. Г. Столетовых. – Влади- мир : Изд-во ВлГУ, 2021. – 123 с.

2. Радченко И.А, Николаев И.Н. Технологии и инфраструктура Big Data. – СПб: Университет ИТМО, 2018. – 52 с.

Библиографический список

Источники:

http://ipcmagazine.ru/legal-issues/big-data-technology-principles-and-architecture

https://habr.com/ru/company/ibm/blog/555086/

https://habr.com/ru/company/vk/blog/703508/

https://habr.com/ru/company/jugru/blog/568638/93 Хранилище данных [Электронный ресурс]. URL: https://www.in- tuit.ru/studies/courses/599/455/lecture/10155 (дата обращения: 07.07.2020).

Где хранить корпоративные данные: краткий ликбез по Data Warehouse [Электронный ресурс]. – URL: https://www.bigdataschool.ru/bigdata/lsa-data- warehouse-architecture.html

BIG DATA 2017: Где хранить Большие Данные [Электронный ресурс]. URL: https://www.computerworld.ru/articles/BIG-DATA-2017-Gde-hranit-Bolshie- Dannye (дата обращения: 15.06.2020).

99 Где хранить корпоративные данные: краткий ликбез по Data Warehouse.

https://habr.com/ru/post/485180/