Технологии организации, обработки и хранения статистических данных

ФИО преподавателя: Митина О.А.

e-mail: alogmi@yandex.ru

6

Лекция

Компоненты корпоративной информационной фабрики

Условия обучения

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

План

1. Информационная экосистема
2. Структура информационной фабрики
3. Репозиторий НСИ
4. Мастер-данные
5. Процессы ETL и ELT
6. Качество данных

Информационная экосистема компании

Корпоративная информационная фабрика (англ. Corporate Information Factory – CIF) - физическое воплощение информационной экосистемы.

Б. Инмон начало1980-х

Структура CIF

Уровень источников данных. Включает разнообразные источники первичных данных, таких, как OLTP и унаследованные системы, офисные документы, базы данных, файловые архивы, любые файлы, содержащие структурированные данные.

Уровень извлечения, преобразования и загрузки данных. Программно-аппаратный комплекс, реализующий извлечение данных из различных источников, преобразование к единому формату и загрузку в интегрированное хранилище.

Уровень хранения данных. Обеспечивает надежное, защищенное от несанкционированного доступа, хранение данных.

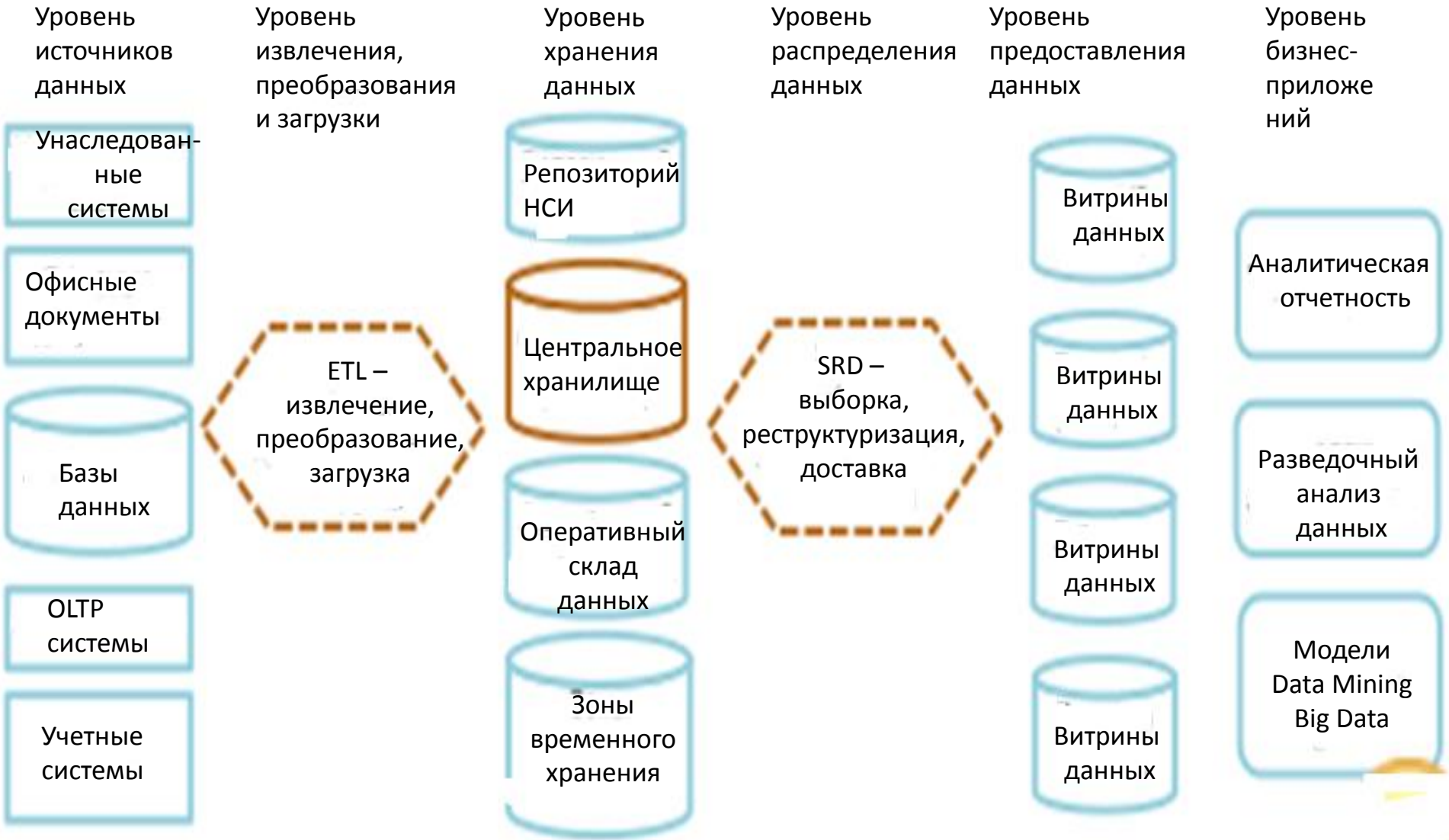
Уровень распределения данных. Выполняет предоставление данных из хранилища различным потребителям.

Структура CIF

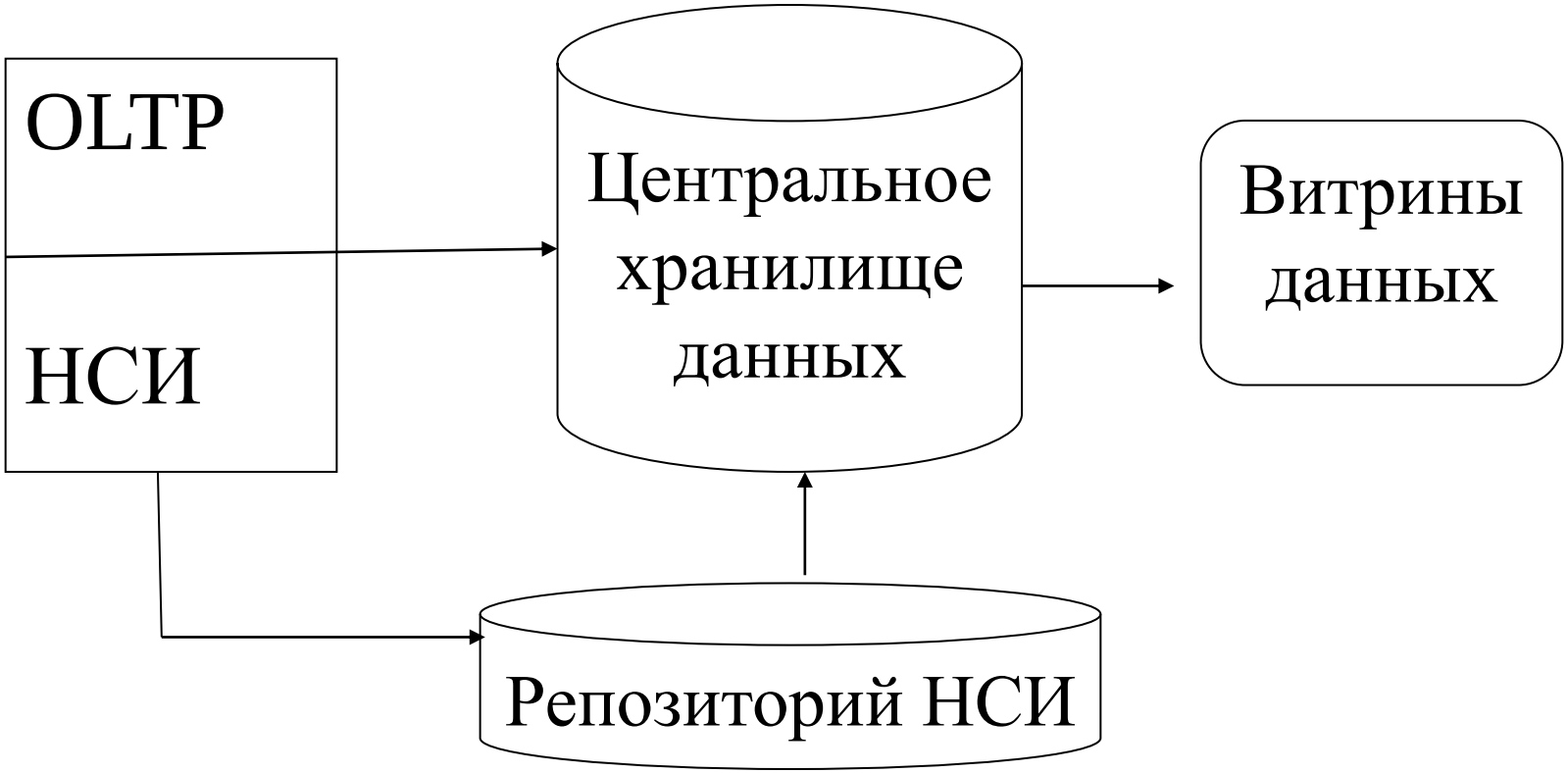
Уровень предоставления данных. Содержит источники данных для конечных пользователей, которым предоставляются данные.

Уровень бизнес-приложений. Содержит приложения, реализующие различные виды анализа данных, формирующие отчетность и решающие задачи автоматизации управления бизнес-процессами компании.

Структура CIF



Репозиторий НСИ



Централизованное хранилище данных

Репозиторий НСИ

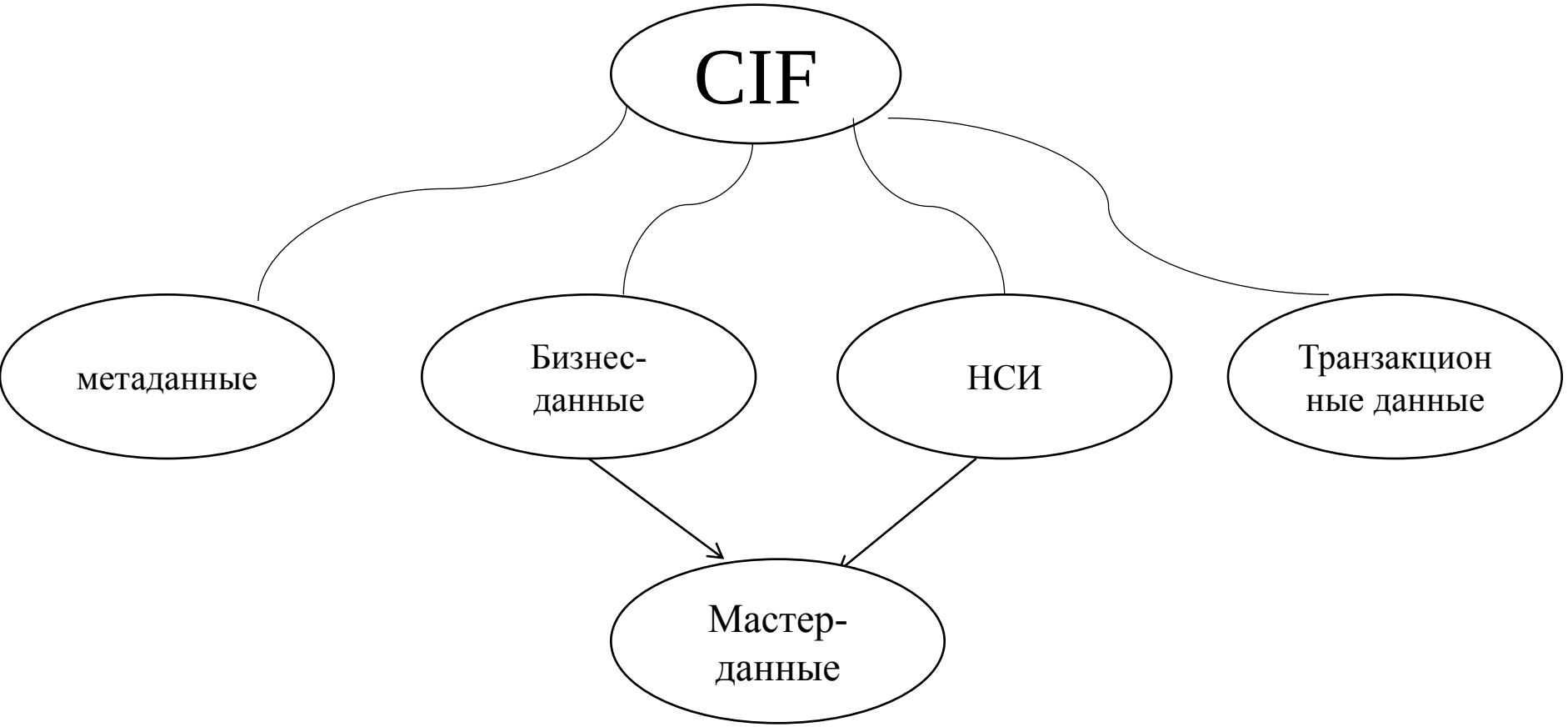
Преимущества:

- хранение НСИ в рамках единой модели, согласованной с остальными компонентами системы;
- ведение НСИ на основе корпоративных и отраслевых стандартов, классификации и кодирования;
- обеспечение единого регламента и технологической среды для доступа к НСИ, а также ведения экспертами классификаторов и справочников;

Репозиторий НСИ

- поддержание необходимого уровня безопасности НСИ и их синхронизации, исключение дублированной, ошибочной и противоречивой информации;
- возможность внедрения классификаторов и справочников НСИ в действующие управленческие, аналитические и другие системы, что позволяет сократить расходы на ведение НСИ;
- оперативность использования НСИ для формирования отчетов.

Мастер-данные (англ. Master Data)



Виды данных CIF

Виды мастер-данных

- Мастер-данные предприятия (англ. Enterprise Master Data) – отдельный источник основных бизнес-данных, используемых во всех системах, приложениях и бизнес процессах внутри предприятия
- Рыночные мастер-данные (англ. Market Master Data) – отдельный источник основных бизнес-данных внутри определенного сегмента рынка;
- Материальные мастер-данные (англ: Material Master Data) – данные о запасных частях, сырье и продуктах, используемых в системах планирования материальных ресурсов предприятия.

Оперативный склад (англ.: Operational Data Store, OSD, ОСД)

Оперативный склад данный – это элемент архитектуры корпоративной информационной фабрики, содержащий объектно-ориентированную, интегрированную и готовую к использованию информацию реального (или почти реального) масштаба времени, не являющийся первичным источником.

Оперативный склад (англ.: Operational Data Store, OSD, ОСД)

Преимуществами использования ОСД являются:

- обеспечение быстрого доступа к важным оперативным данным;
- при использовании ОСД компания всегда имеет представление о текущих параметрах бизнес-процессов и успешности текущей деятельности;

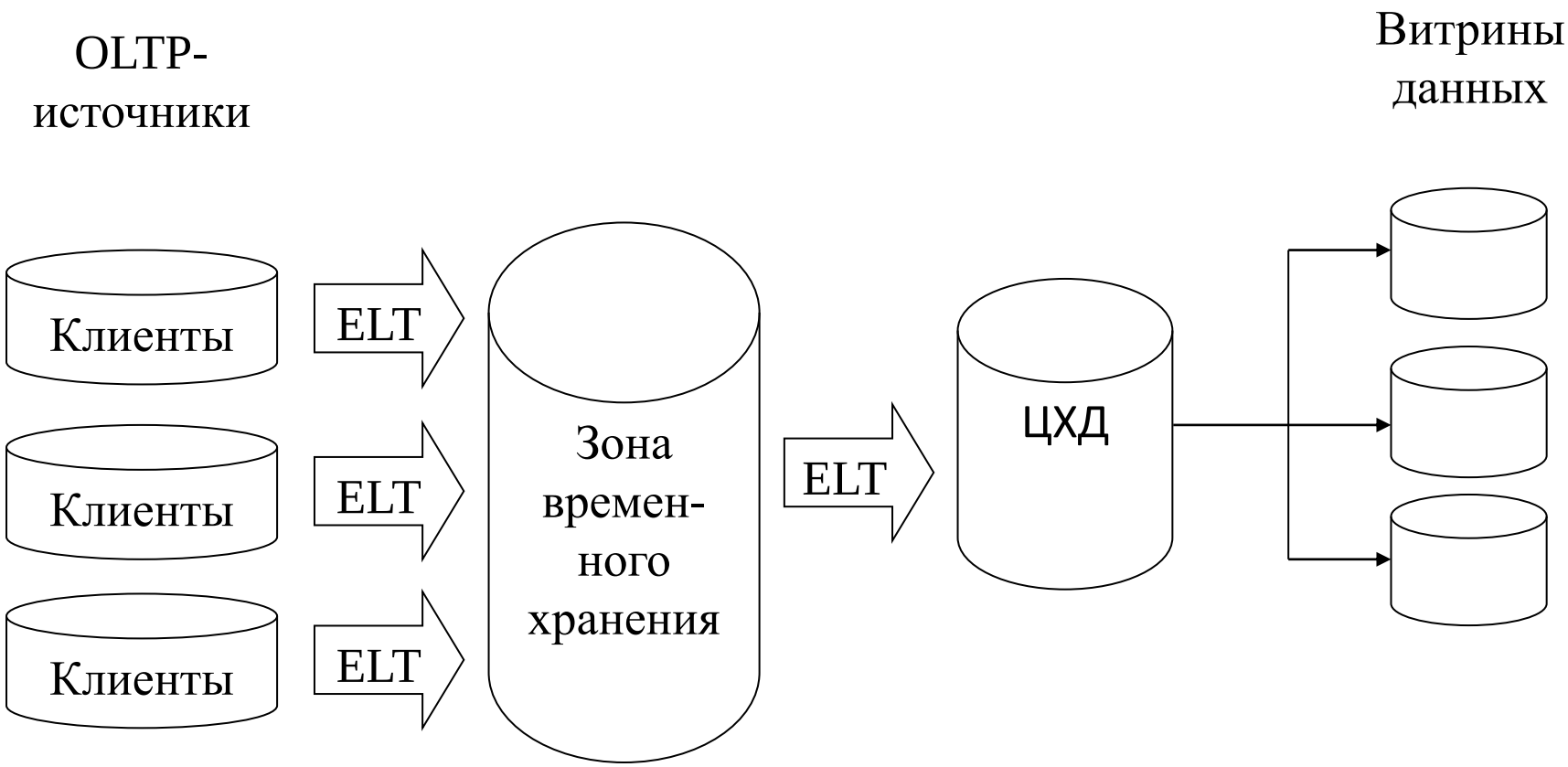
Оперативный склад (англ.: Operational Data Store, OSD, ОСД)

- повышается эффективность формирования оперативных отчетов и снижается нагрузка по запросам к первичным информационным источникам;
- при наличии центрального ХД в него ускоряются процессы загрузки, поскольку часть данных уже находится в ОСД;
- повышаются возможности администрирования.

Зоны временного хранения (англ: *Staging Area*)

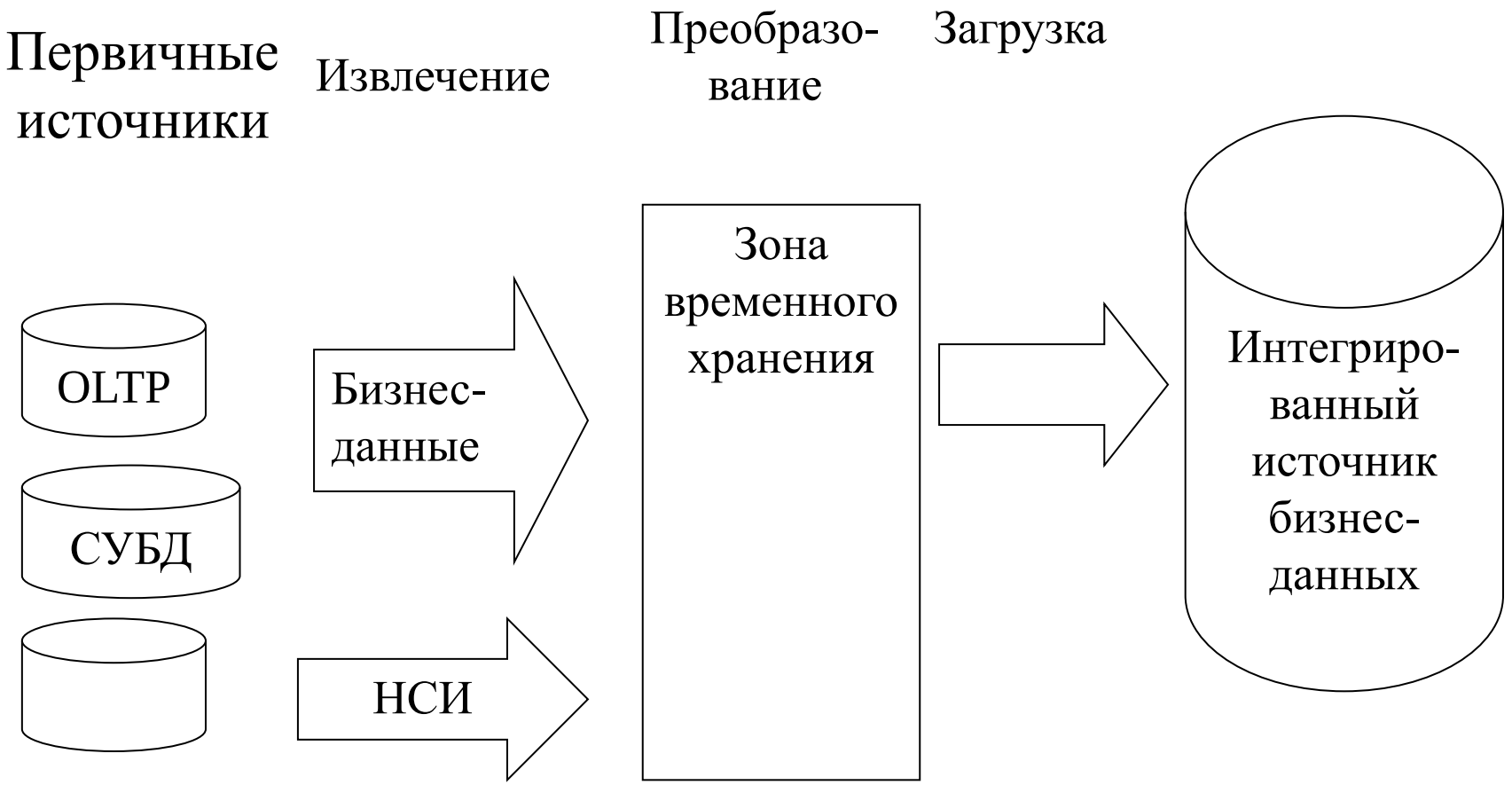
Зоной временного хранения называют буферную базу данных, необходимую для выполнения некоторых внутренних служебных технологических операций над данными, перемещаемыми из источников в ЦХД.

Зоны временного хранения (англ: *Staging Area*)



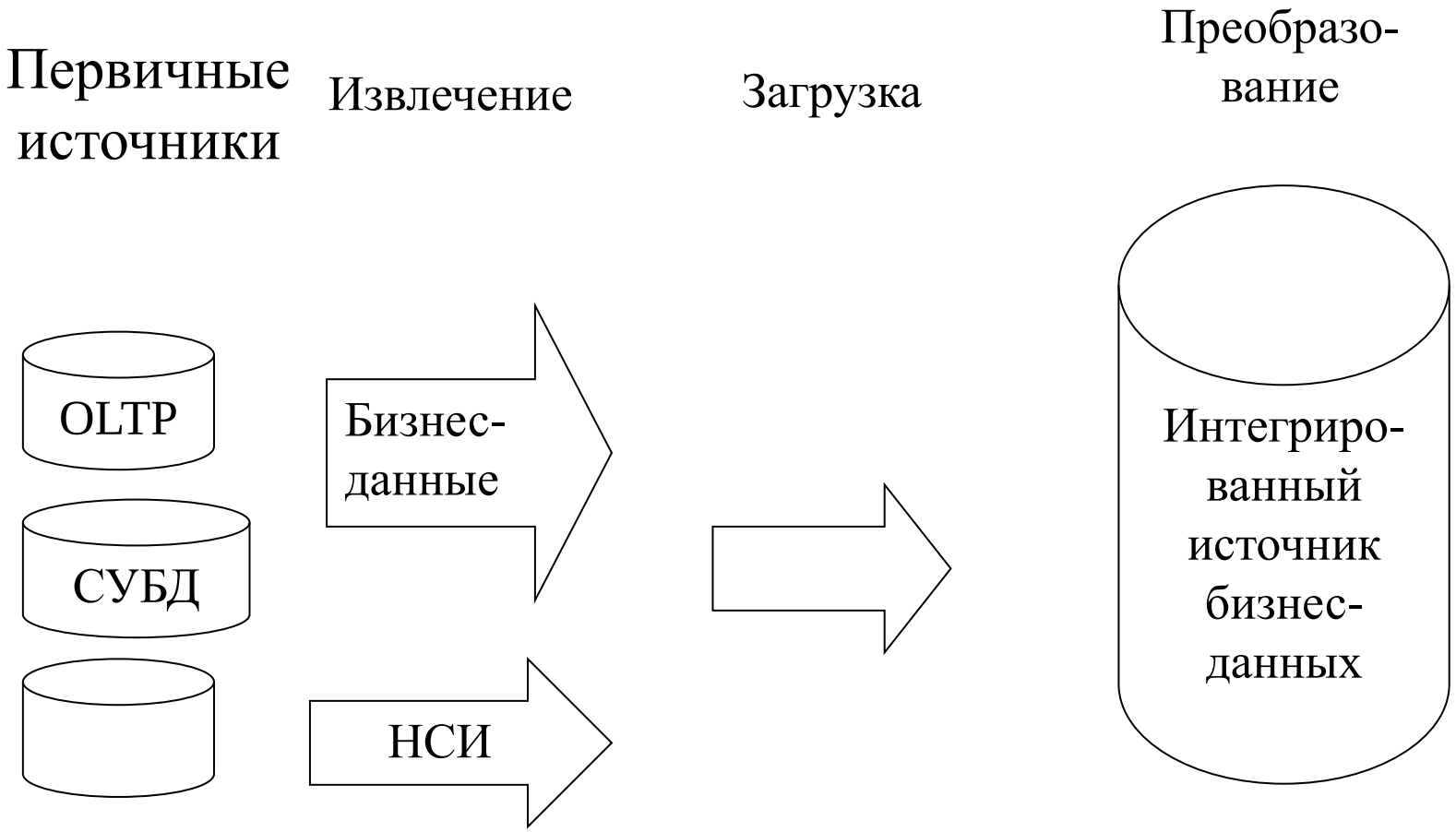
Выделение зон временного хранения в CIF

Процессы информативной корпоративной фабрики (ELT и ETL)



Вариант структуры процессов ETL в контуре CIF

Процессы информативной корпоративной фабрики (ELT и ETL)



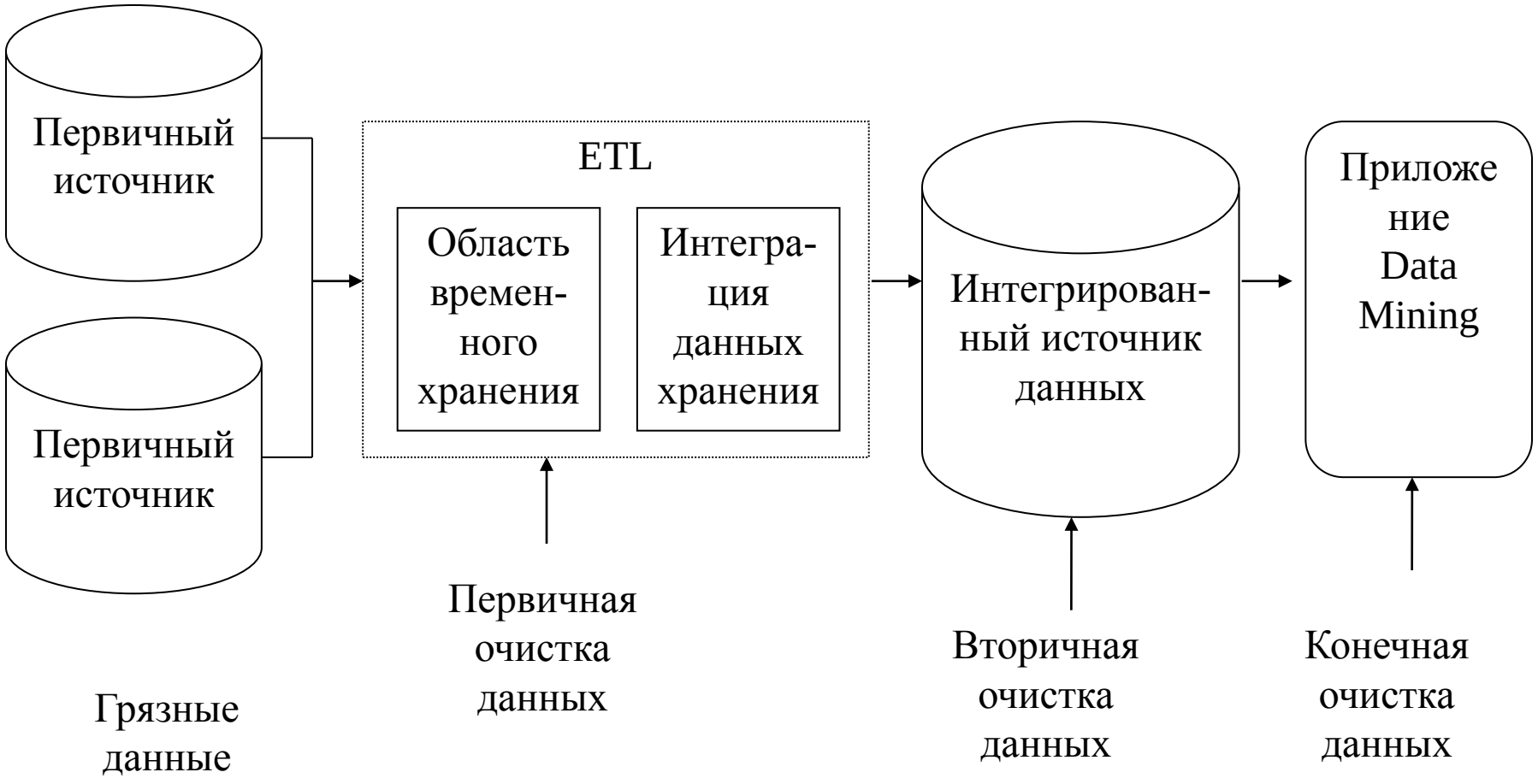
Вариант структуры процессов ELT в контуре CIF

Качество данных – Data Quality

- изначально низкое качество первичных источников данных ;
- отсутствие понимания причин снижения качества данных;
- неудачный выбор программных средств;
- попытка перенести мероприятия, направленные на повышение качества данных.

Обеспечение качества данных

Уровни отчистки данных



Путь перемещения данных в CIF

Очистка данных в консолидированных источниках

- очистка в ETL производится автоматически, без участия пользователя;
- не все проблемы в данных могут быть обнаружены и распознаны на этапе ETL;
- время на поиск и обнаружение проблем на этапе ETL является ограниченным;
- сам процесс интеграции может порождать проблемы в данных.

Очистка данных в бизнес-приложениях

ETL	SRD
Извлекает данные из различных внешних систем и источников	Извлекает данные из интегрированного источника (чаще всего – ЦХД)
На входе «сырые» данные, которые требуется преобразовать к единому формату и модели представления	На входе очищенные и хорошо структурированные данные, которые требуется преобразовать в формат, используемый приложением-потребителем
Загружает данные в интегрированный источник данных	Доставляет данные различным системам-потребителям, чаще всего витринам 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 Многомерная модель данных

5-6 Интеграция данных и бизнес-аналитика

7-8 Интеграция данных

9 Хранилища данных

10 Процессы информативной корпоративной фабрики

11 Базовые архитектуры корпоративной информационной фабрики

12 Технология OLAP и ее особенности

13 Понятие OLAP-куба. Операции над OLAP-кубами

14 Аналитические платформы. Инструменты бизнес-аналитики

15-16 Большие данные. Наука о данных

Спасибо за внимание!