Технологии организации, обработки и хранения статистических данных

ФИО преподавателя: Митина О.А.

e-mail: alogmi@yandex.ru

Лекция

Интеграция данных

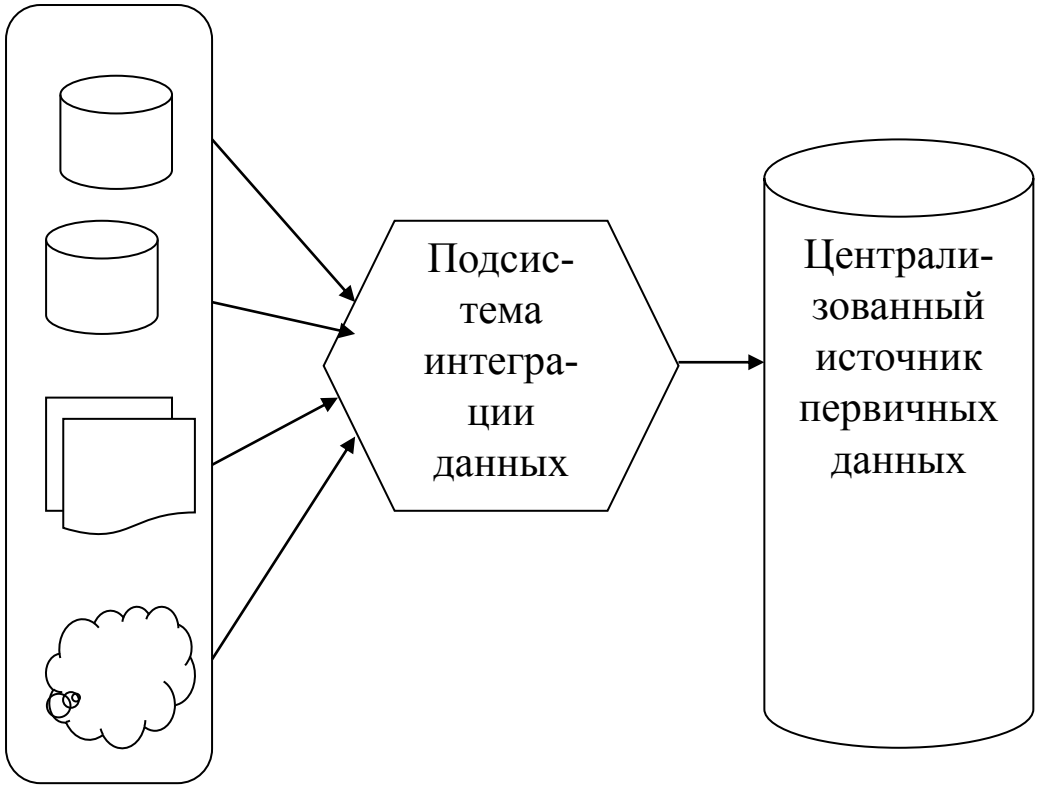
Условия обучения

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

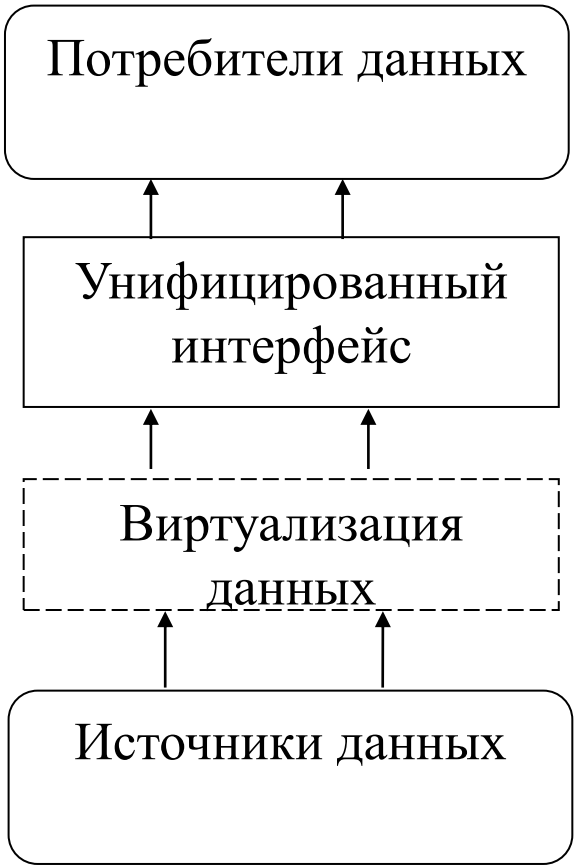
Интеграция данных

Интеграция данных (ИД) – это процесс объединения данных, находящихся в различных разнородных источниках, в единственном физическом источнике или обеспечение единого унифицированного интерфейса для некоторой совокупности источников (при этом физического объединения данных путем копирования информации не происходит).

Интеграция данных



Источники данных

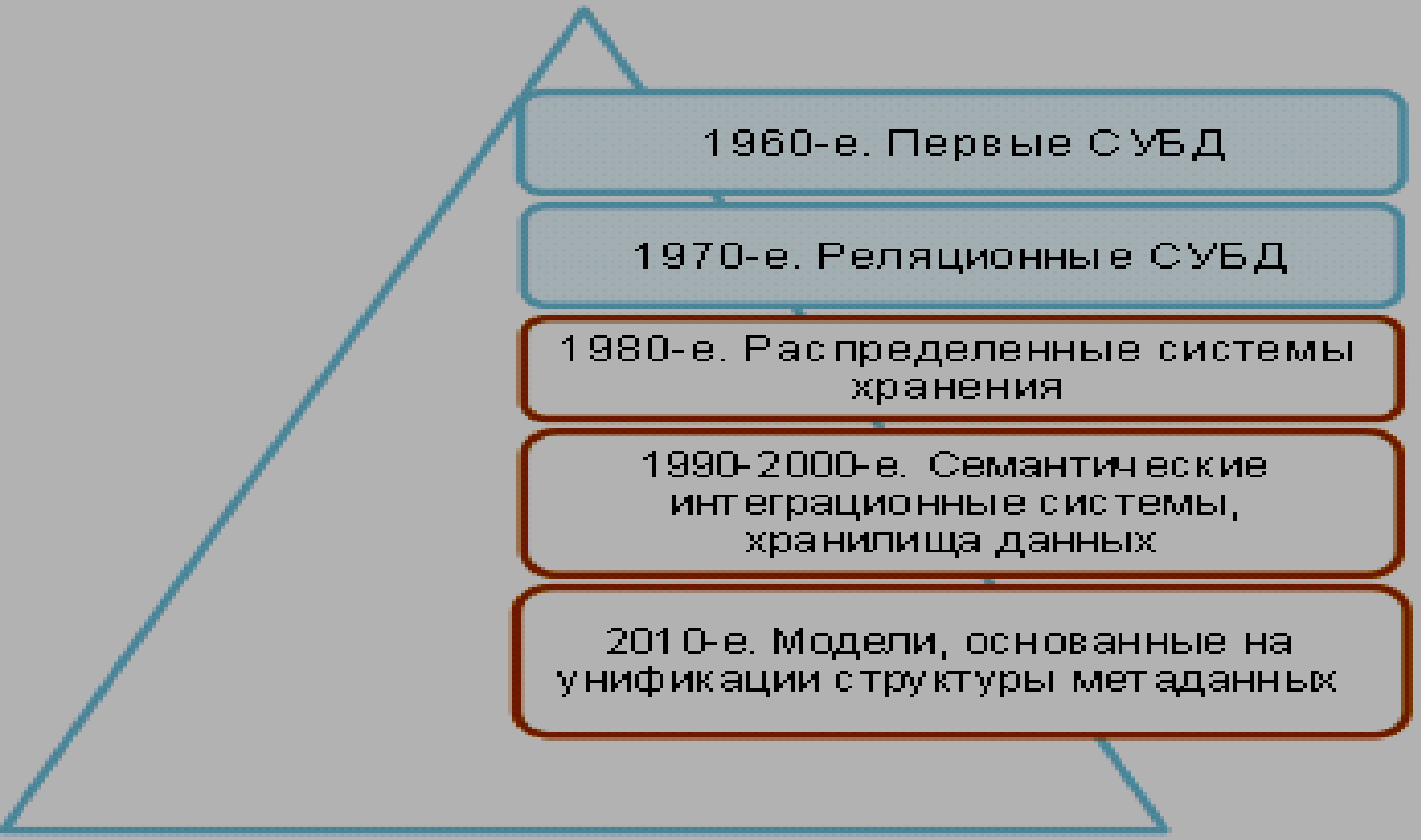


Источники данных

Источник данных

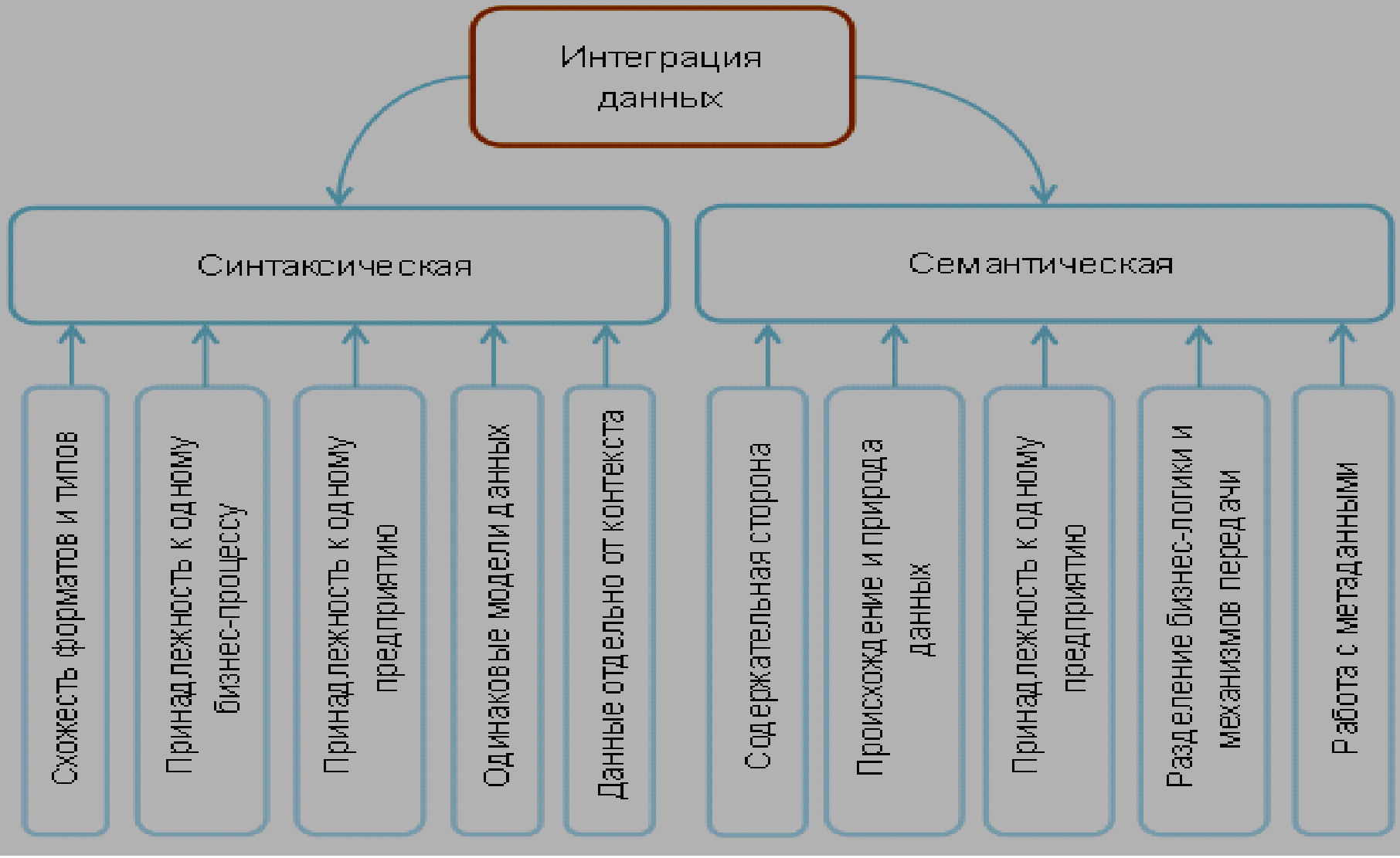
Источник данных - это ***объект, содержащий структурированные данные.***

Интеграция данных: краткая история проблемы



МИРЭА
Российский технологический университет

Подходы к интеграция данных



Уровни интеграции данных

•**Физический уровень.** Данные из различных источников преобразуются к единому формату и сохраняются в одном источнике.

•**Логический уровень.** Данные по-прежнему физически размещаются.

в своих источниках, доступ к ним реализуется на основе некоторой глобальной схемы, отражающей их требуемое совместное представление.

•**Семантический уровень.** Обеспечивает поддержку единого представления данных с учетом их семантических свойств в контексте единой онтологии предметной области.

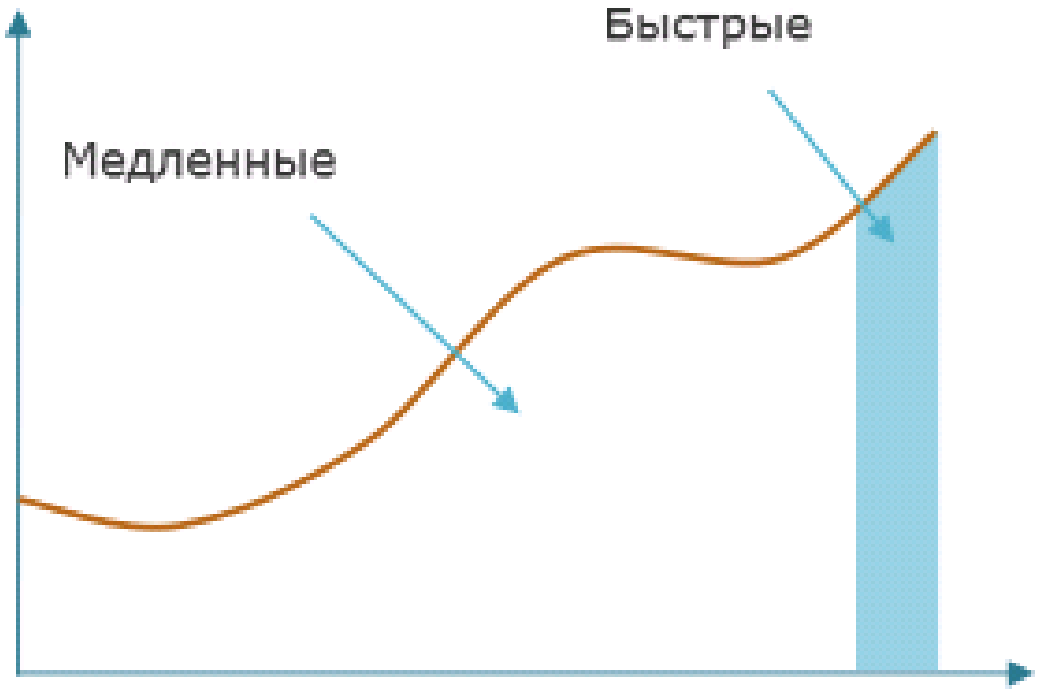
Способы интеграции данных

1. **Виртуальный** – реализуется с помощью механизма доступа, который при выполнении запроса пользователя формирует требуемое представление данных непосредственно из источников. Наиболее эффективен, если источники данных являются динамически обновляемыми.
2. **Материализованный** (актуальный) – формируется полное физическое представление данных, сосуществующее с источниками, на основе которых оно было получено.

Задачи, решаемые в процессе интеграции данных

- Разработка архитектуры СИД.
- Разработка интегрирующей модели данных, являющейся основой единого пользовательского интерфейса СИД.
- Разработка методов представления моделей данных и построение отображений, поддерживаемых отдельными источниками данных.
- Интеграция метаданных, используемых в системе источников данных.
- Преодоление неоднородности источников данных.
- Разработка механизмов семантической интеграции источников данных.

Быстрые и медленные данные



Быстрые и медленные данные

Быстрые данные поступают непрерывно, сплошным потоком и являются сильно детализированными, поскольку отражают элементарные события в жизни бизнеса.

Быстрые данные отражают текущие тенденции в бизнесе, и позволяют принимать оперативные, тактические решения.

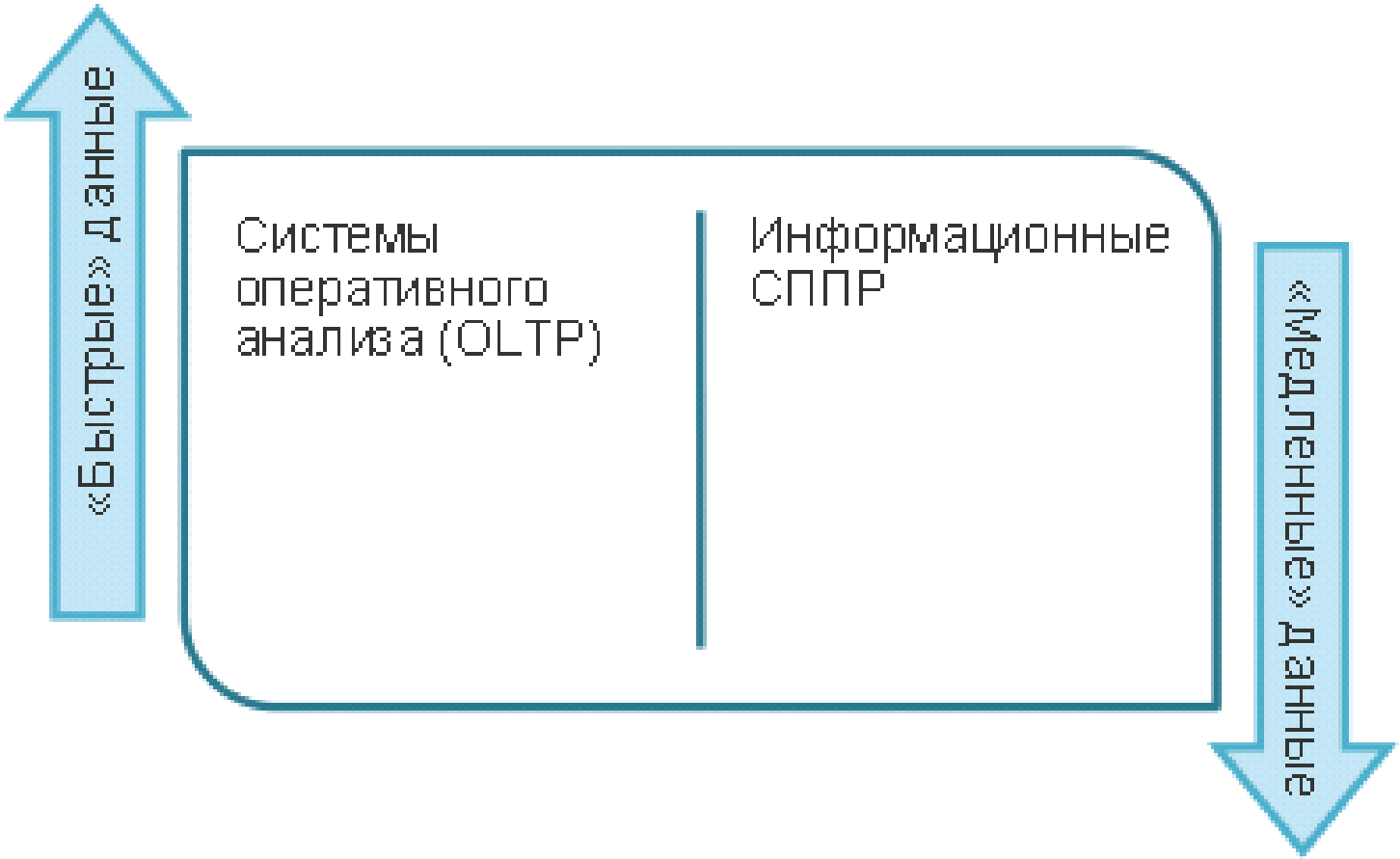
Быстрые и медленные данные

Медленными называют данные, которые не являются медленно меняющимися или перемещающимися, а отражают долгосрочные зависимости и закономерности бизнес-процессов, что позволяет использовать их для решения задач стратегического анализа и поддержки принятия решений.

Быстрые и медленные данные

Медленными называют данные, которые не являются медленно меняющимися или перемещающимися, а отражают долгосрочные зависимости и закономерности бизнес-процессов, что позволяет использовать их для решения задач стратегического анализа и поддержки принятия решений.

Быстрые и медленные данные



Системы оперативного анализа

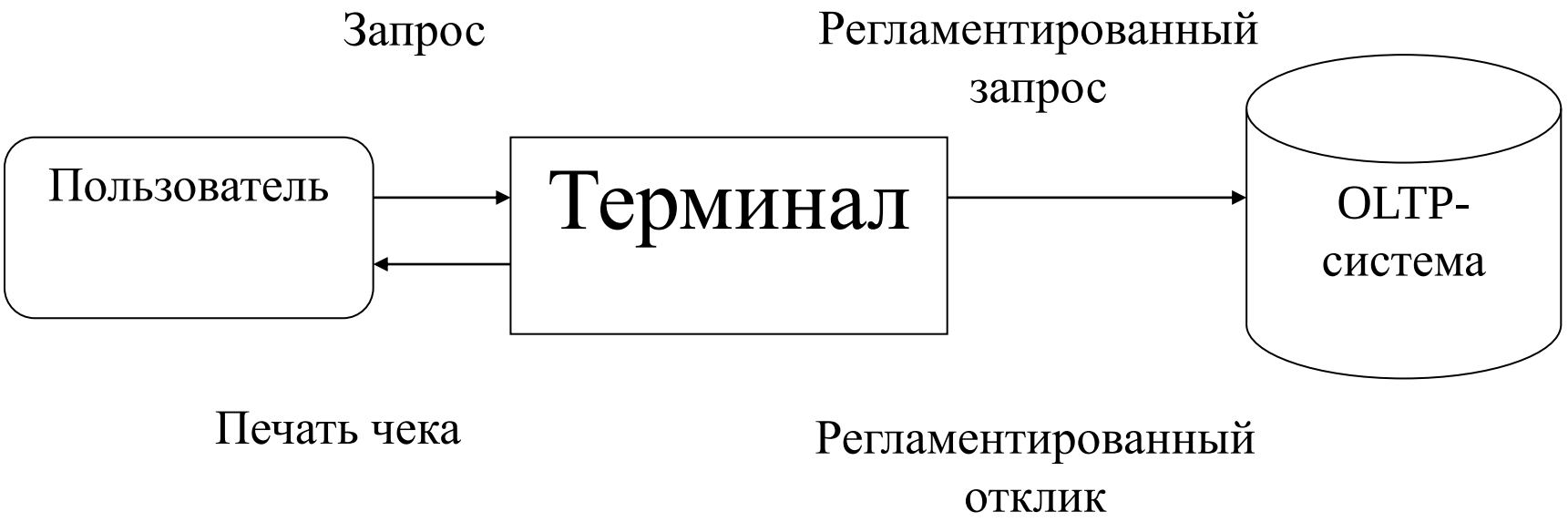
Системы оперативной обработки информации получили название OLTP (англ.: *On-Line Transaction Processing* – оперативная, то есть в режиме реального времени, обработка транзакций).

Под *транзакцией* в данном случае понимают некоторый набор логически связанных операций над базой данных, который рассматривается как единое, завершенное, с точки зрения бизнес-логики, действие над некоторой информацией, связанное с выполнением определенной бизнес-функции.

Системы оперативного анализа

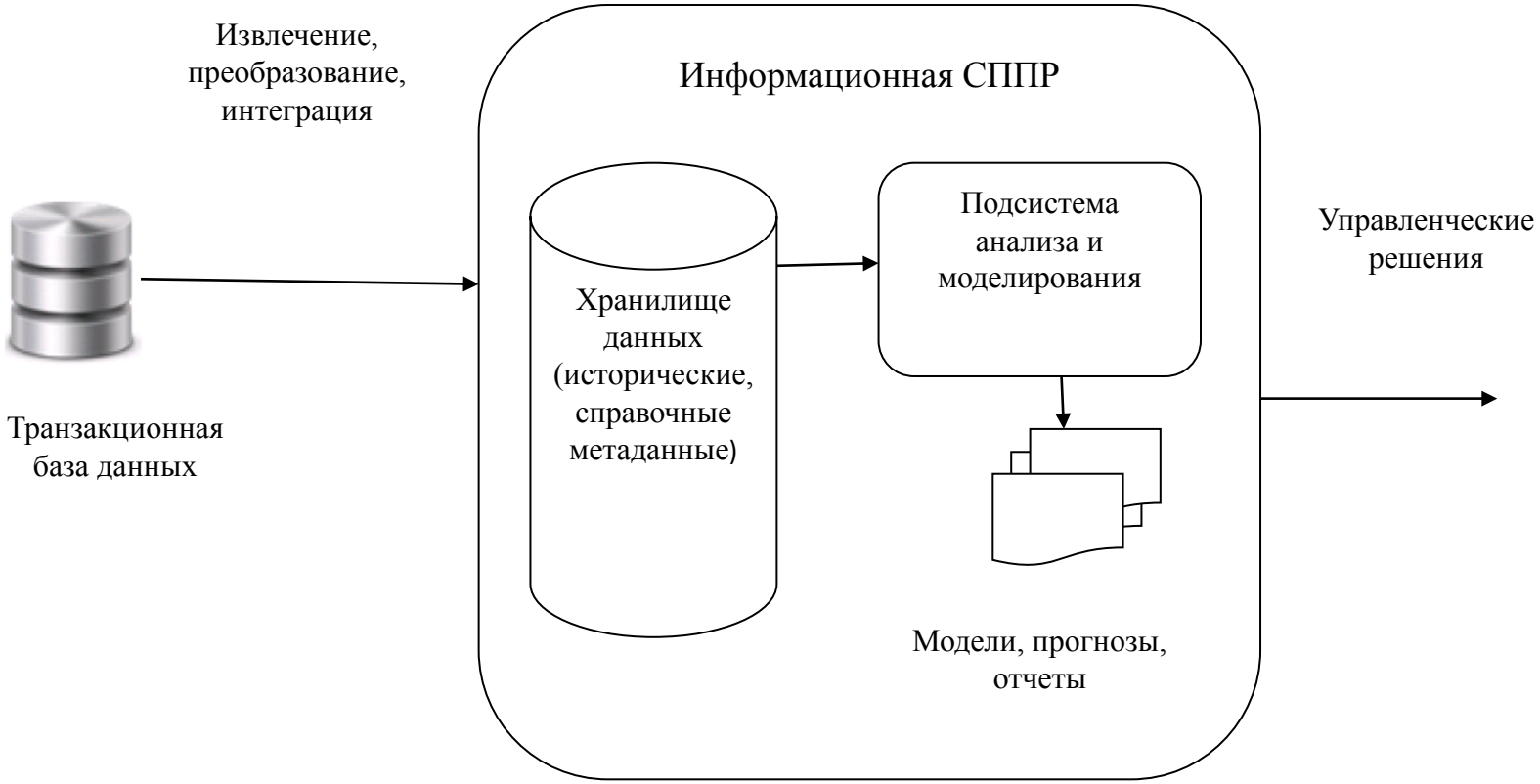
Обобщенная схема движения данных в транзакционной системе

Системы оперативного анализа



Пример транзакции OLTP-системы

Системы поддержки принятия решений



Системы поддержки принятия решений. Способы организации данных

- для выполнения нерегламентированных запросов необходима обработка массивов данных из множества разнородных источников;
- для выполнения запросов, связанных с анализом тенденций, прогнозированием протяженных во времени процессов, необходимы исторические данные, накопленные за достаточно длительный период, что не обеспечивается обычными OLTP-системами;
- транзакционные данные в OLTP-системах являются максимально детализированными, что не оптимально с точки зрения анализа. При аналитической обработке предпочтение отдается данным с некоторым уровнем их обобщения.

Системы поддержки принятия решений. Способы организации данных

интеграцию – извлечение и объединение данных из множества разнородных источников в централизованную систему хранения;
преобразование – приведение данных к наиболее удобному для анализа виду (агрегирование, кодирование и так далее);

профайлинг («англ. profile» – профиль)

аудит это понятие, обозначающее совокупность психологических методов и методик оценки и прогнозирования поведения человека на основе анализа наиболее информативных частных признаков, характеристик
внешности, невербального и вербального поведения;

Системы поддержки принятия решений. Способы организации данных

аудит – это независимая проверка деятельности отдельно взятой организации с целью изучения достоверности финансовой отчётности компании. Данной процедуре подвергаются также все процессы, проходящие внутри фирмы, производимые продукты, а также реализуемые проекты.

очистка – восстановление нарушения полноты и целостности данных, исключение из них пропусков, дубликатов, противоречий и других факторов, мешающих их корректному анализу.

Разница между OLTP-системами и информационными СППР

	Цели использования данных	Уровень обобщения (детализации) данных	Требования к качеству данных	Формат хранения данных	Время хранения данных
OLTP	Формирование отчетности, простые алгоритмы обработки	Максимально детализированы	"Сырые" данные с ошибками, пропусками и т.д.	Могут храниться в различных форматах	В пределах отчетного периода (как правило, 1-2 года)
СППР	Аналитическая обработка с целью поиска скрытых закономерностей, построения прогнозов и т.д.	Различные уровни детализации (обобщения)	Данные, прошедшие профайлинг, аудит, очистку	Хранятся и обрабатываются в едином формате	Годы, десятиления

Разница между OLTP-системами и информационными СППР

	Изменение данных	Периодичность обновления	Доступ к данным	Характер выполняемых запросов	Время выполнения запросов
OLTP	Данные могут добавляться, изменяться, удаляться	Часто, но в небольших объемах	Обеспечивается доступ ко всем текущим (оперативным) данным	Стандартные (регулярные), настроенные заранее	Несколько секунд
СППР	Допускается только добавление новых данных; ранее добавленные данные изменяться не должны	Редко, но в больших объемах (в соответствии с регламентом)	Обеспечивается доступ к историческим данным с соблюдением их хронологии	Нерегламентированные, формируемые аналитиком "на лету"	До нескольких минут

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!