# Технологии организации, обработки и хранения статистических данных

ФИО преподавателя: Митина О.А.

e-mail: alogmi@yandex.ru

7

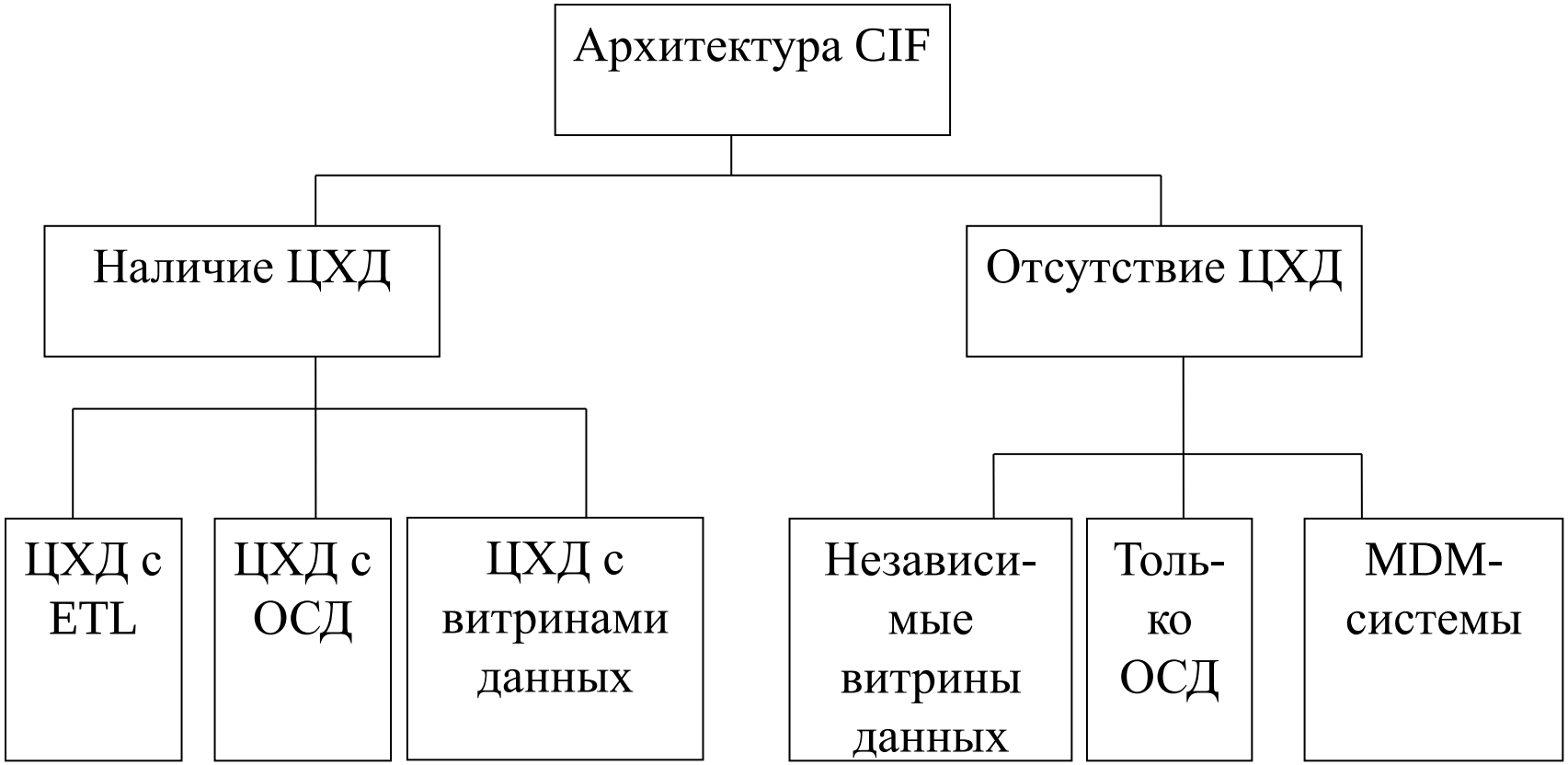
# Лекция

## Базовые архитектуры CIF. Преобразование данных

# Условия обучения

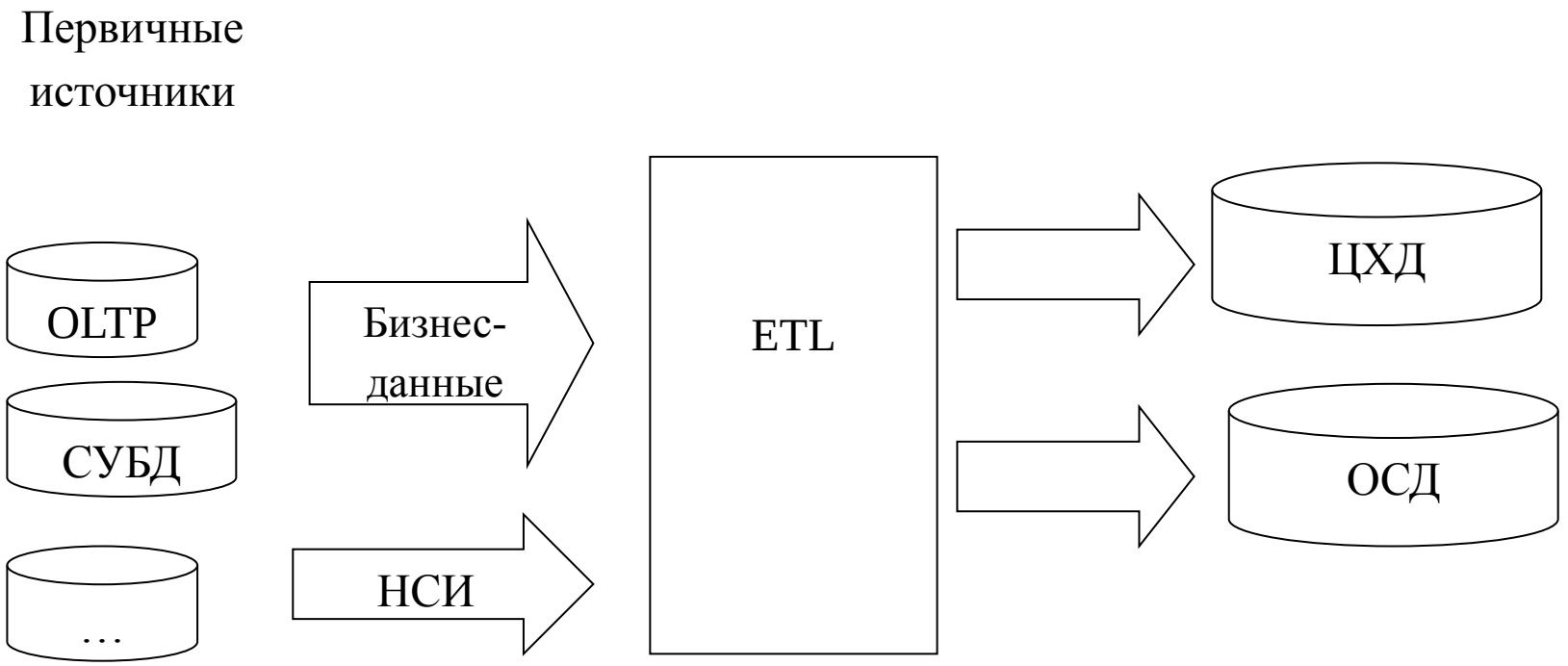
- По итогам изучения дисциплины проводится экзамен
- В течение семестра необходимо выполнить все практические работы

# Архитектура CIF




Архитектура CIF

# Централизованное ХД с ETL

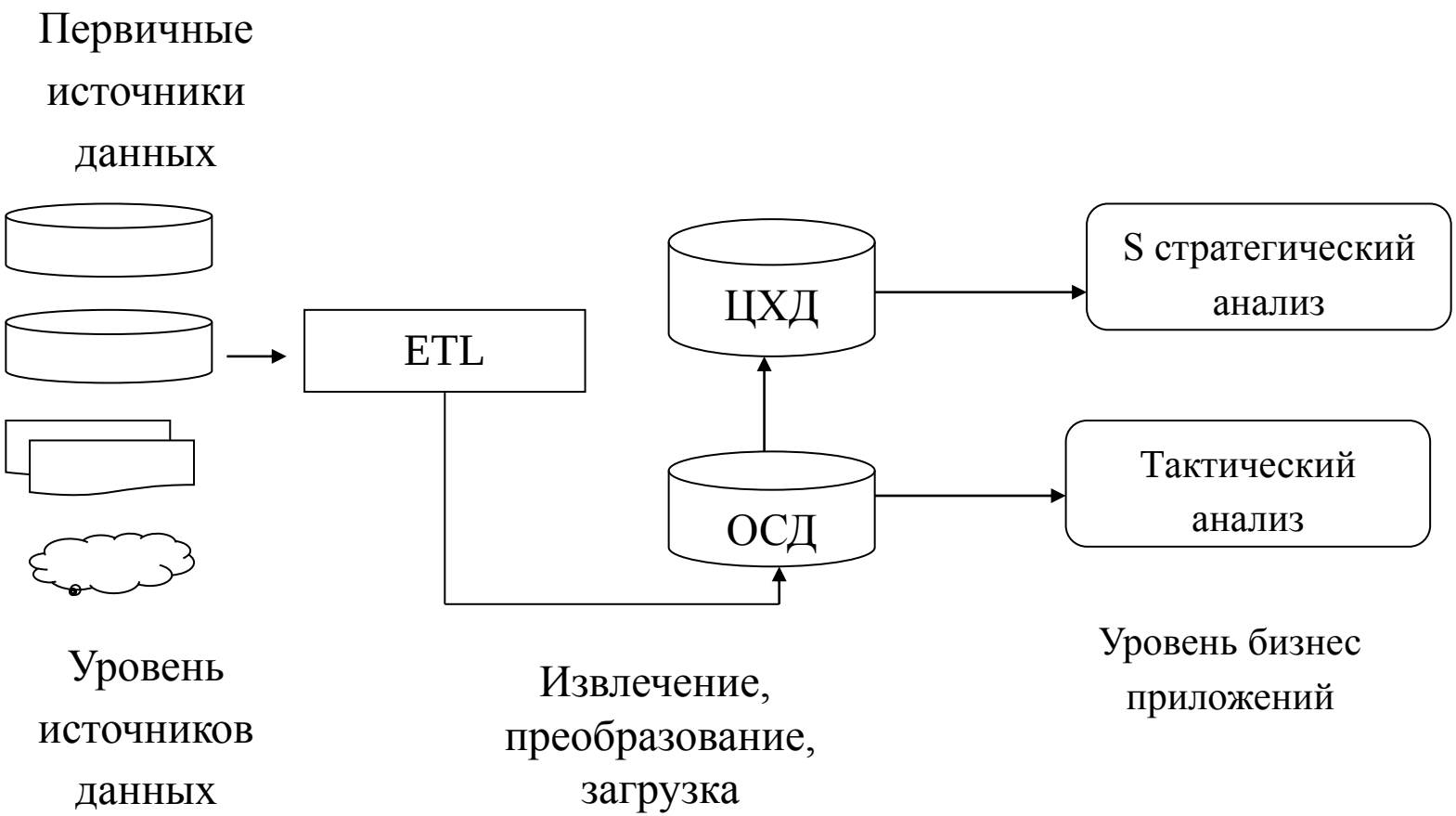



Централизованное ХД с ETL

# Централизованное ХД с ОСД (подходы)

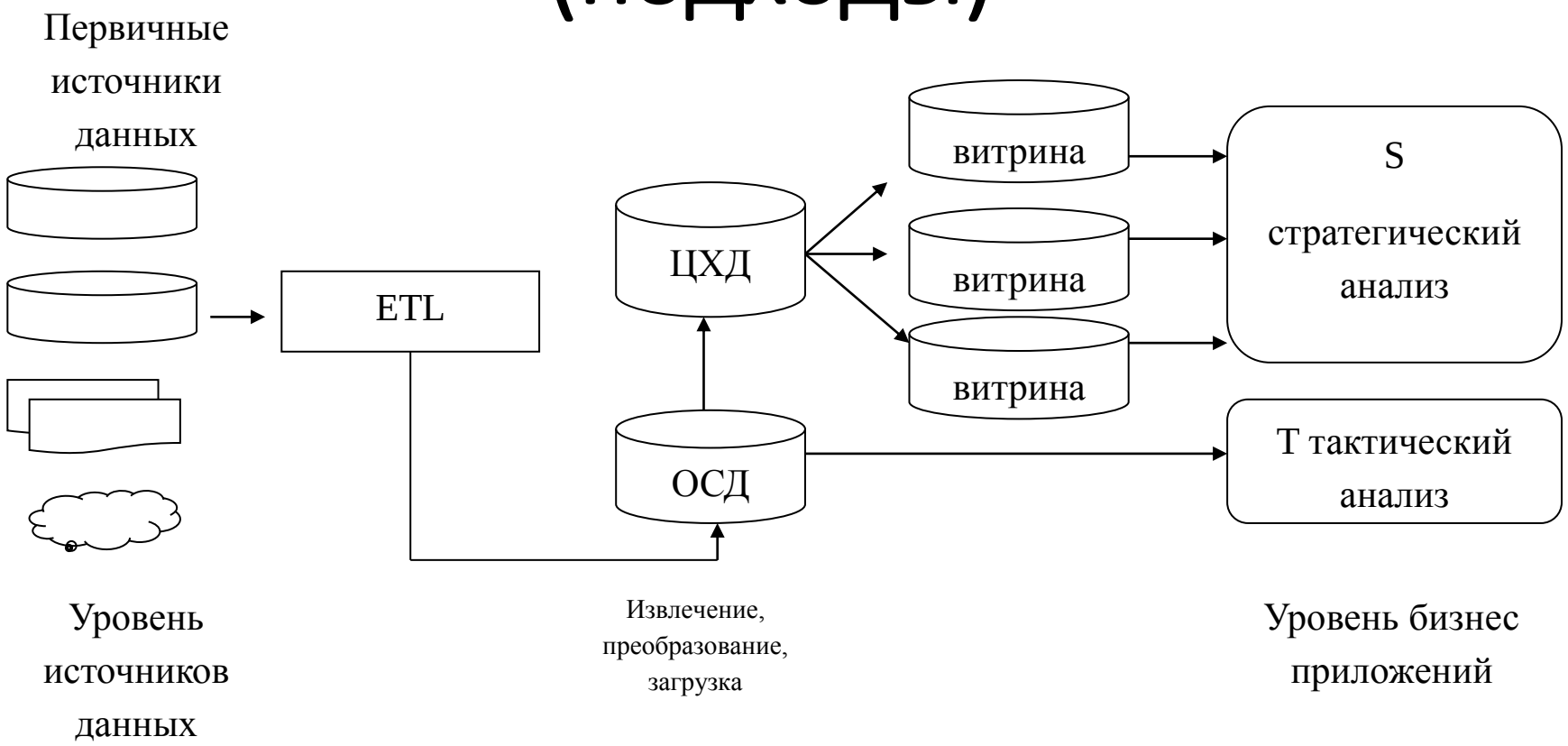
- последовательный;
- параллельный;
- независимый.

# Централизованное ХД с ОСД (подходы)



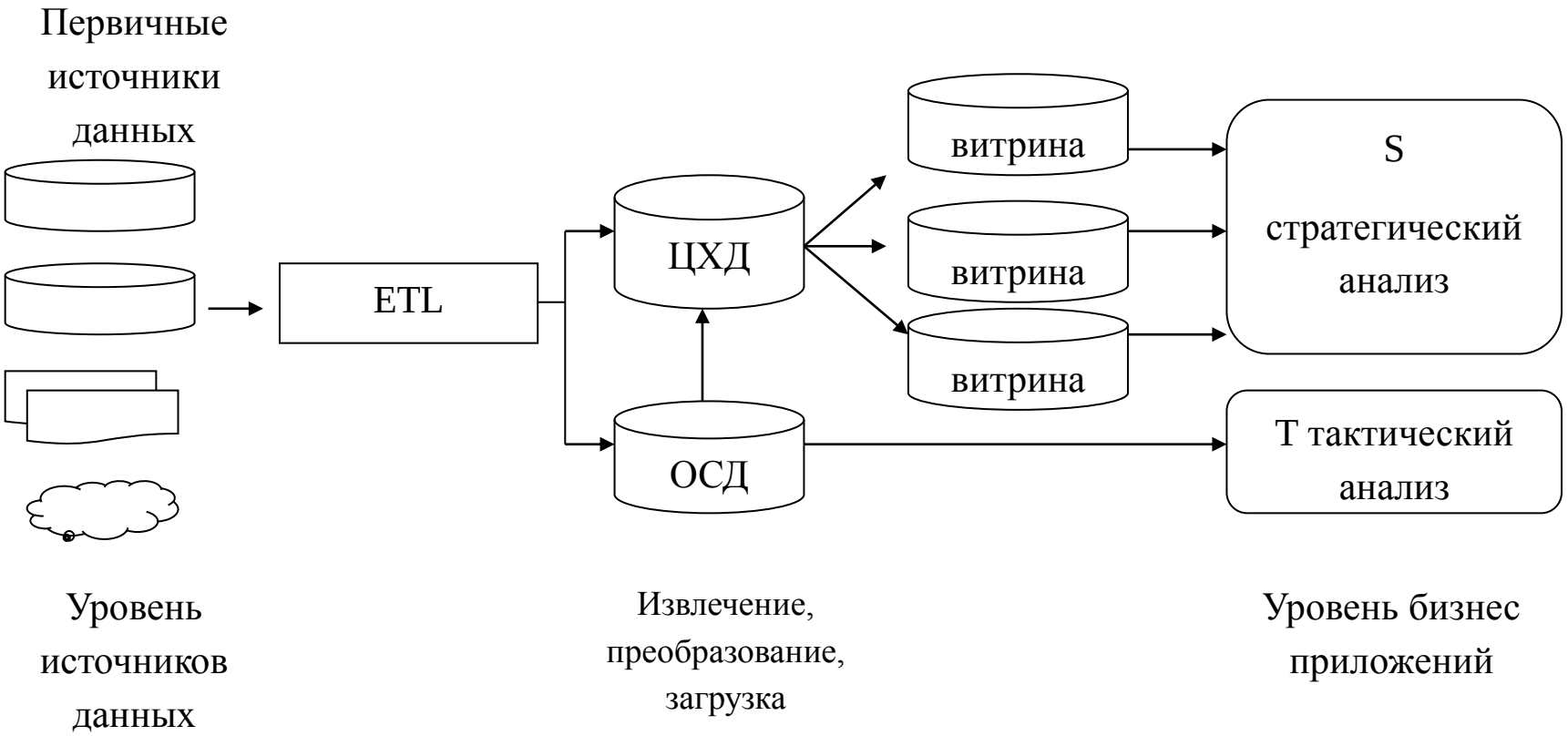

Архитектура последовательного соединения ОСД и ЦХД

# Централизованное ХД с ОСД (подходы)




Архитектура параллельного соединения ОСД и ЦХД

# Централизованное ХД с ОСД (подходы)




Архитектура независимого соединения ОСД и ЦХД

# Централизованное ХД с ОСД

Преимущества:

- данные с выхода ETL быстрее оказываются в ЦХД, что **улучшает** его синхронизацию с источниками данных,
- если в процессе тактического анализа в данные, расположенные в ОСД, вносятся нежелательные изменения, то они не попадут в ЦХД.

# Централизованное ХД с ОСД

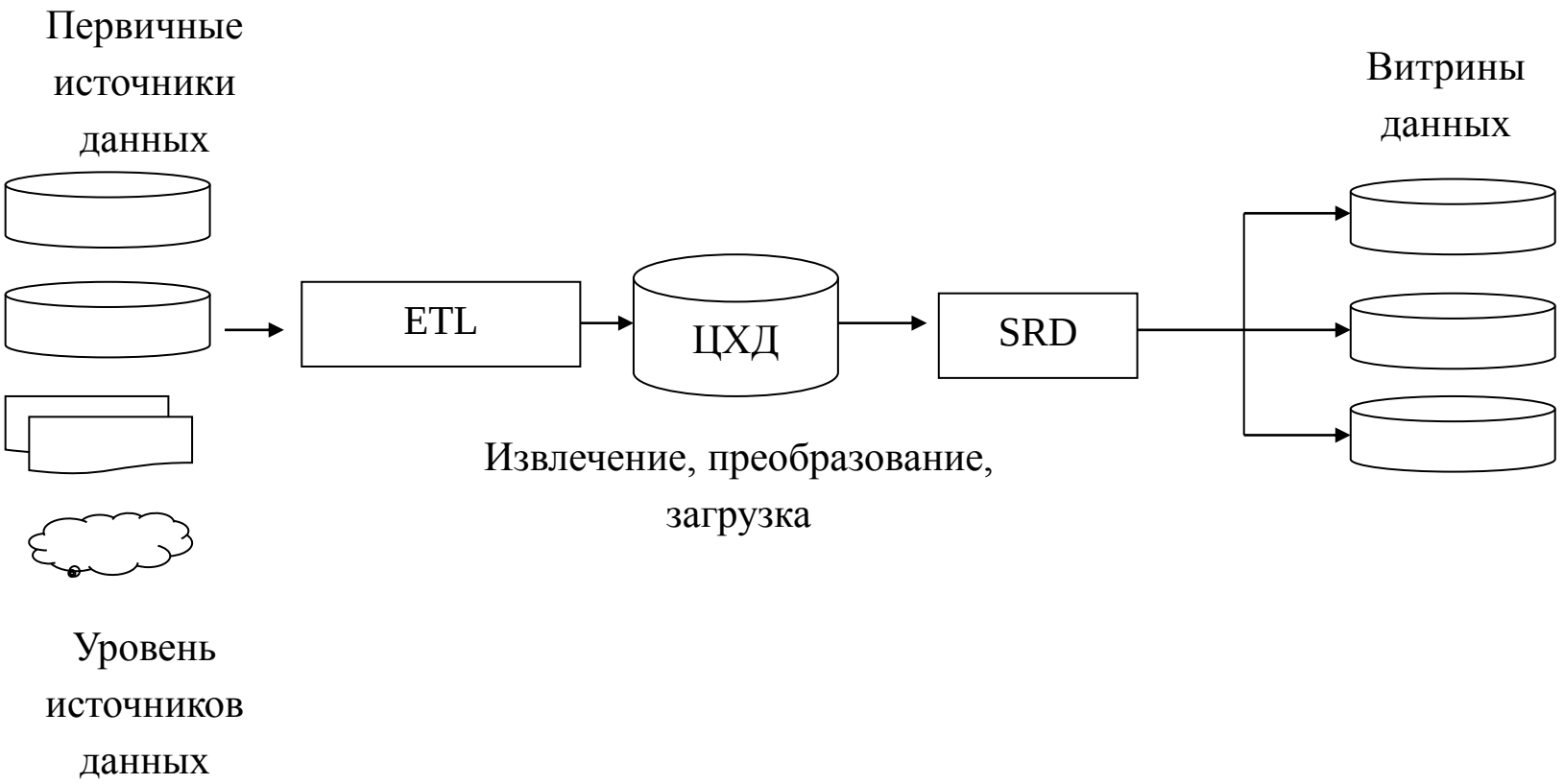
Недостатки:

- отсутствует дополнительный этап контроля и повышения качества данных, реализуемый в ОСД при параллельном соединении;
- регламент прохождения данных через ОСД должен соответствовать регламенту загрузки в ЦХД, что снижает время, доступное для работы с данными при их тактическом анализе.

# Централизованное ХД с витринами данных

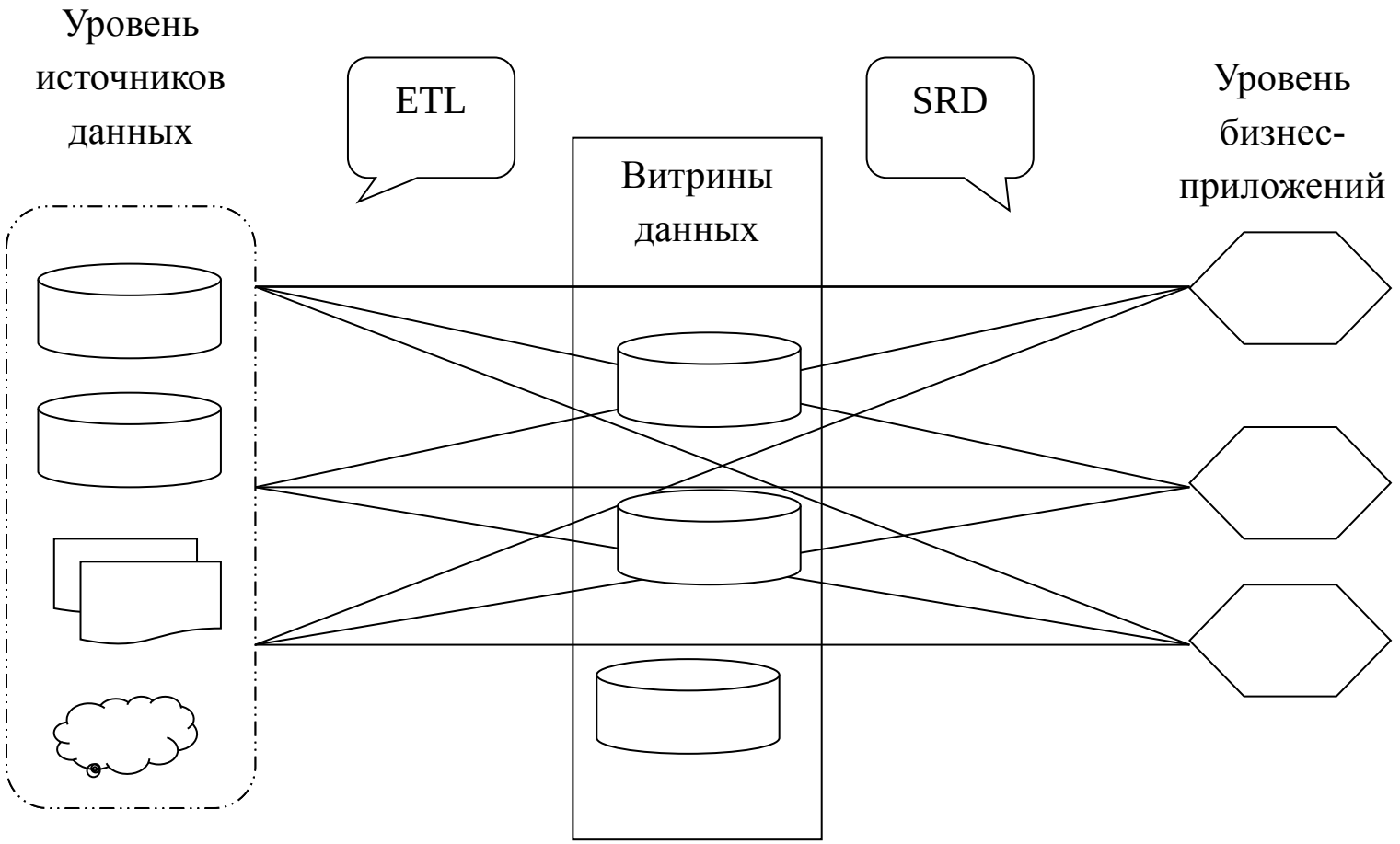
- обеспечение доступа к данным;
- содержание данных в виде, удобном для работы всеми подразделениями, по всем направлениям бизнеса;
- хранилище используется одновременно многими пользователями;
- искажение информации в ЦХД;
- недостаточная пропускная способность и ненадежность телекоммуникационных линий.

# Способы соединения витрин данных с ЦХД




Независимое соединение витрин данных с ЦХД

# Независимые витрины данных




Архитектура без централизованного ХД с независимыми витринами данных

# Независимые витрины данных

Недостатки:

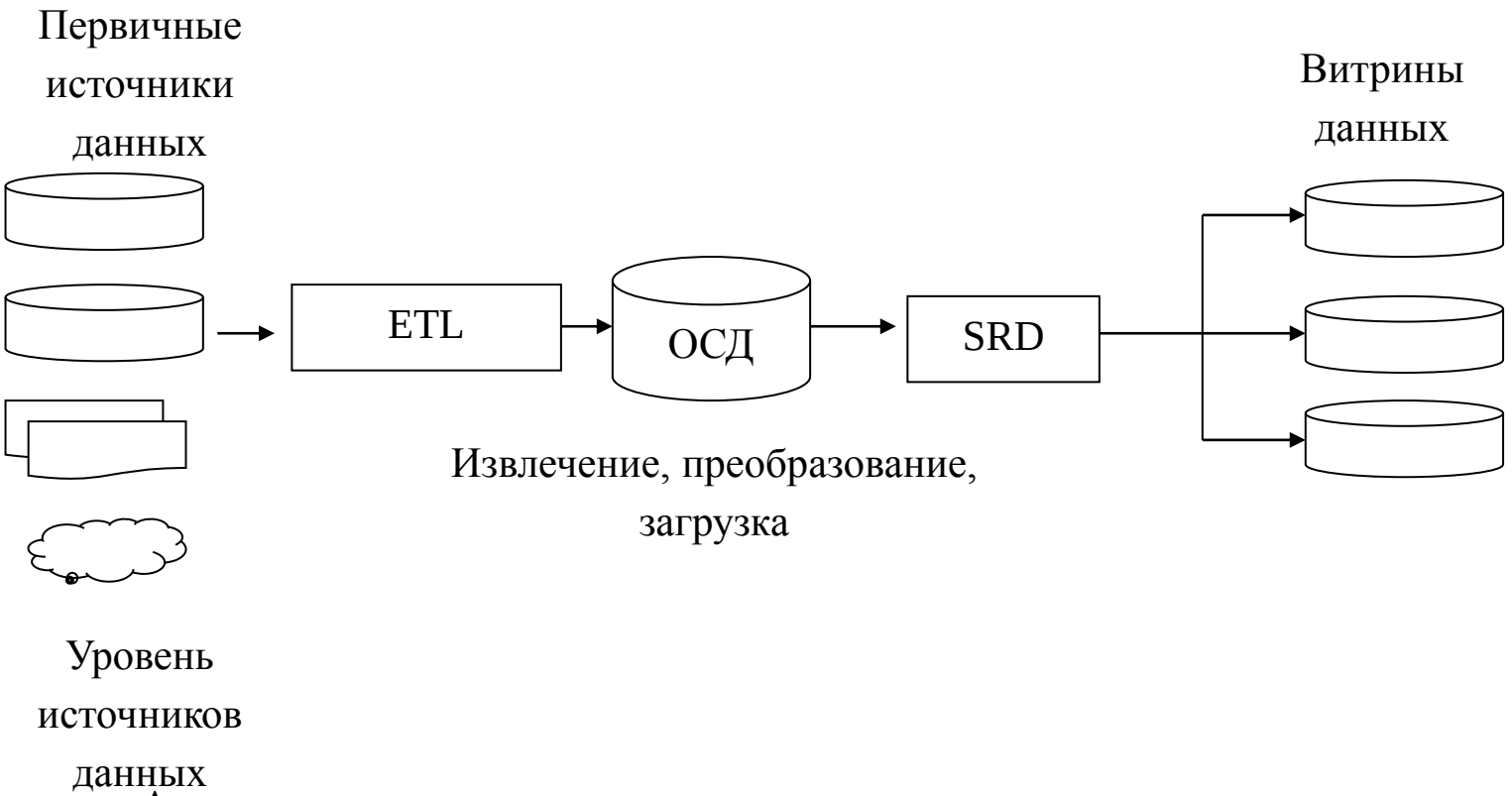
- избыточность данных;
- интеграция и трансформация данных выполняется для каждой витрины;
- плохая согласованность результатов анализа;
- ухудшение масштабируемости системы;
- нет общего взгляда на работу компании.

# Независимые витрины данных

Преимущества:

- низкая стоимость и простота реализации;
- возможность построения системы для отдельно взятого подразделения, если построение корпоративной системы для компании слишком дорого.

# Только оперативный склад данных




Архитектура с оперативным складом данных

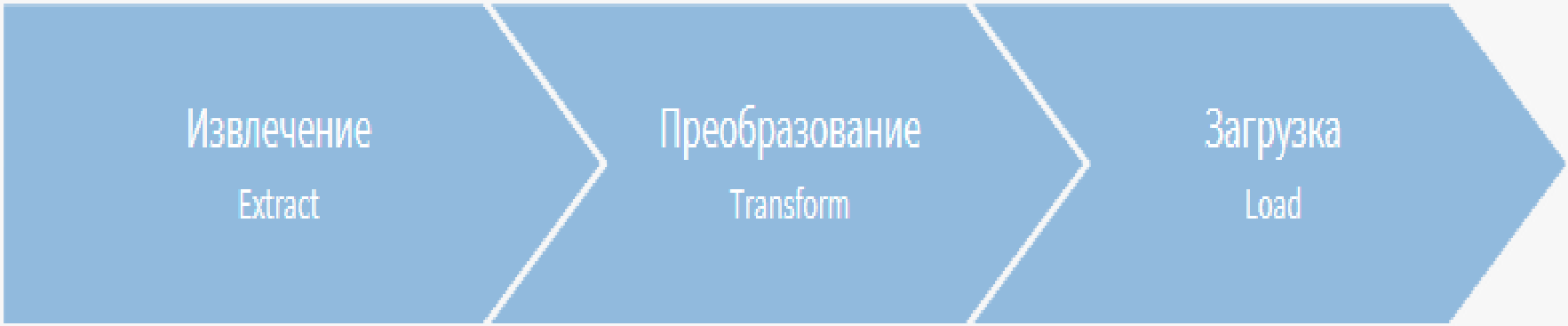
Лекция

Подготовка данных

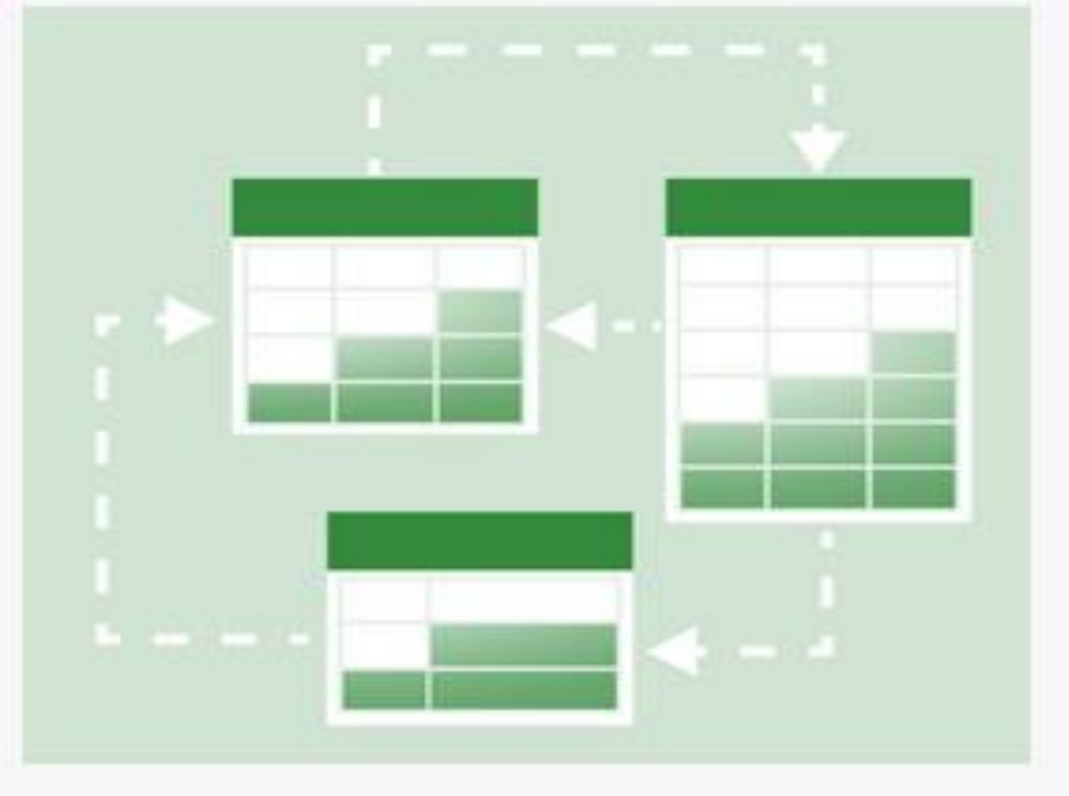
# БЛОК 1. ГРУППИРОВКА И ПРЕОБРАЗОВАНИЕ ДАТЫ

Информация | Запуск >

# ETL процесс


Извлечение
Extract
Преобразование
Transform
Загрузка
Load

# Преобразование данных как часть ETL-операций



# Преобразование данных как часть ETL-операций

в процессе переноса и загрузки данных в интегрированный источник или области временного хранения (ETL)

непосредственно при подготовке данных к анализу в бизнес-приложении (SRD)

# Преобразование данных как часть ETL-операций

приведение их в соответствие с моделью данных, используемой в хранилище

осуществление корректной консолидации данных и загрузка в хранилище

# Параметры полей

- Квантование
- Фильтр строк
- Сортировка
- Обогащение данных
- Табличная подстановка значений
- Группировка
- Вычисляемые значения
- Преобразование упорядоченных данных
- Транспонирование

# Группировка данных

| Дата | Количество |
|---|---|
| 01.03.2020 | 10000 |
| 02.03.2020 | 12000 |
| 01.04.2020 | 8000 |
| 03.04.2020 | 13000 |

| Дата | Количество |
|---|---|
| 01.03.2020 | 22000 |
| 01.04.2020 | 21000 |

# Группировка данных

# Группировка данных

Исходная таблица

| Дата | Цена | Количество | Сумма |
|---|---|---|---|
| 01.03.2016 | 583,33 | 85 | 16 500 |
| 02.03.2016 | 1 166,67 | 40 | 44 000 |
| 03.03.2016 | 800,00 | 78 | 53 800 |
| 04.03.2016 | 662,50 | 145 | 41 500 |
| 05.03.2016 | 562,50 | 125 | 46 500 |
| 06.03.2016 | 583,33 | 106 | 20 500 |

Группировка по полю «Дата»

| Клиент | Цена | Количество | Сумма |
|---|---|---|---|
| ЗАО «Монтажник» | 258,33 | 200 | 40 500 |
| ООО «Агрострой» | 766,67 | 98 | 37 800 |
| ООО «Полигон» | 1030,00 | 81 | 73 500 |
| ООО «Тандем» | 1200,00 | 40 | 45 000 |
| ООО «Шплинт» | 433,33 | 160 | 26 000 |

Группировка по полю «Клиент»

| Товар | Цена | Количество | Сумма |
|---|---|---|---|
| Блоки | 900 | 60 | 54 000 |
| Керамзит | 100 | 310 | 31 000 |
| Кирпич | 1500 | 31 | 46 500 |
| Плиты | 1100 | 68 | 74 800 |
| Цемент | 150 | 110 | 16 500 |

Группировка по полю «Товар»

| Дата | Клиент | Товар | Цена | Количество | Сумма |
|---|---|---|---|---|---|
| 01.03.2016 | ООО «Полигон» | Цемент | 150 | 20 | 3 000 |
| 01.03.2016 | ЗАО «Монтажник» | Керамзит | 100 | 60 | 6 000 |
| 01.03.2016 | ООО «Тандем» | Кирпич | 1 500 | 5 | 7 500 |
| 02.03.2016 | ООО «Шплинт» | Плиты | 1 100 | 10 | 11 000 |
| 02.03.2016 | ЗАО «Монтажник» | Блоки | 900 | 20 | 18 000 |
| 02.03.2016 | ООО «Полигон» | Кирпич | 1 500 | 10 | 15 000 |
| 03.03.2016 | ООО «Агрострой» | Плиты | 1 100 | 8 | 8 800 |
| 03.03.2016 | ЗАО «Монтажник» | Керамзит | 100 | 30 | 3 000 |
| 03.03.2016 | ООО «Тандем» | Блоки | 900 | 10 | 9 000 |
| 03.03.2016 | ООО «Полигон» | Плиты | 1 100 | 30 | 33 000 |
| 04.03.2016 | ООО «Шплинт» | Керамзит | 100 | 100 | 10 000 |
| 04.03.2016 | ООО «Тандем» | Кирпич | 1 500 | 10 | 15 000 |
| 04.03.2016 | ЗАО «Монтажник» | Цемент | 150 | 20 | 3 000 |
| 04.03.2016 | ООО «Полигон» | Блоки | 900 | 15 | 13 500 |
| 05.03.2016 | ООО «Агрострой» | Плиты | 1 100 | 20 | 22 000 |
| 05.03.2016 | ООО «Шплинт» | Керамзит | 100 | 50 | 5 000 |
| 05.03.2016 | ЗАО «Монтажник» | Цемент | 150 | 40 | 6 000 |
| 05.03.2016 | ООО «Тандем» | Блоки | 900 | 15 | 13 500 |
| 06.03.2016 | ООО «Полигон» | Кирпич | 1 500 | 6 | 9 000 |
| 06.03.2016 | ООО «Агрострой» | Керамзит | 100 | 70 | 7 000 |
| 06.03.2016 | ЗАО «Монтажник» | Цемент | 150 | 30 | 4 500 |

# Группировка по полю «Дата»

| Дата | Цена | Количество | Сумма |
|---|---|---|---|
| 01.03.2016 | 583,33 | 85 | 16 500 |
| 02.03.2016 | 1 166,67 | 40 | 44 000 |
| 03.03.2016 | 800,00 | 78 | 53 800 |
| 04.03.2016 | 662,50 | 145 | 41 500 |
| 05.03.2016 | 562,50 | 125 | 46 500 |
| 06.03.2016 | 583,33 | 106 | 20 500 |

# Группировка по полю «Клиент»

| Клиент | Цена | Количество | Сумма |
|---|---|---|---|
| ЗАО «Монтажник» | 258,33 | 200 | 40 500 |
| ООО «Агрострой» | 766,67 | 98 | 37 800 |
| ООО «Полигон» | 1030,00 | 81 | 73 500 |
| ООО «Тандем» | 1200,00 | 40 | 45 000 |
| ООО «Шплинт» | 433,33 | 160 | 26 000 |

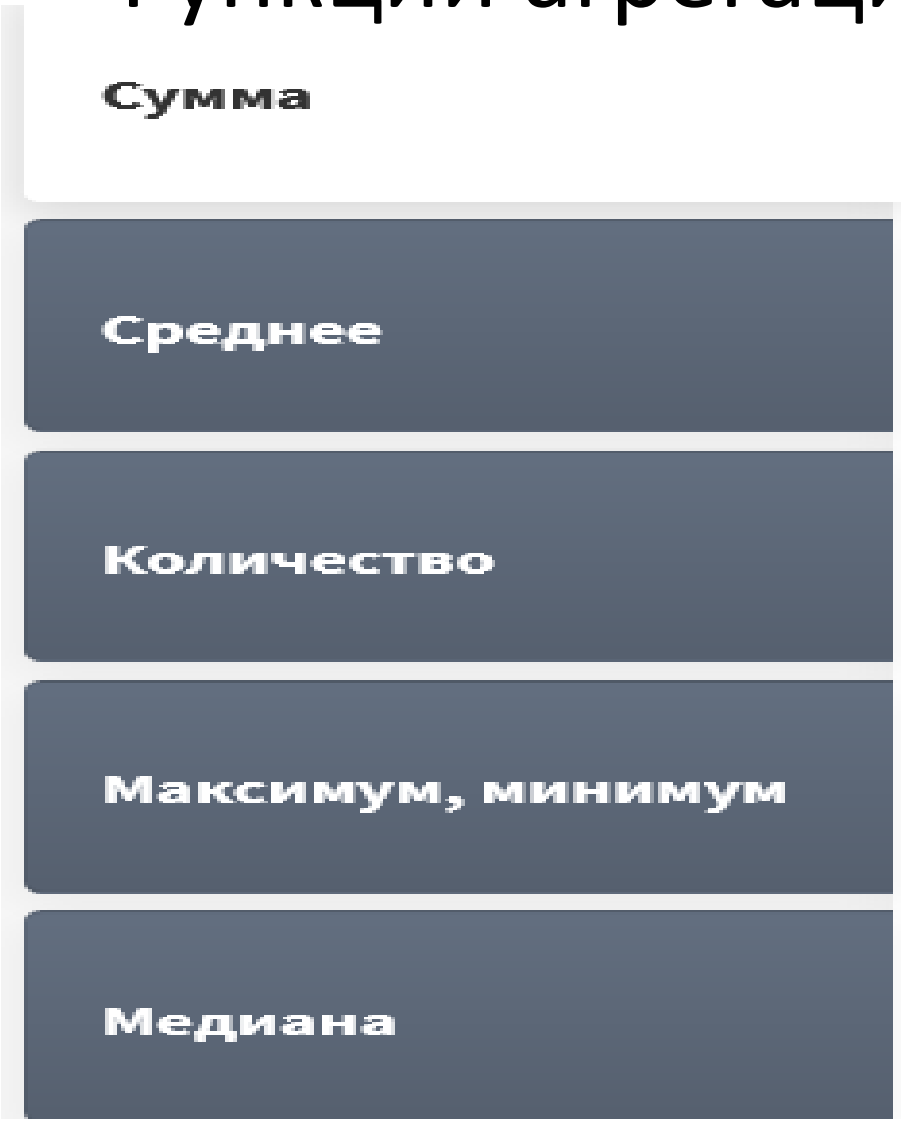
# Группировка по полю «Товар»

| Товар | Цена | Количество | Сумма |
|---|---|---|---|
| Блоки | 900 | 60 | 54 000 |
| Керамзит | 100 | 310 | 31 000 |
| Кирпич | 1500 | 31 | 46 500 |
| Плиты | 1100 | 68 | 74 800 |
| Цемент | 150 | 110 | 16 500 |

# Группировка - пример

| Распределение сумм продаж по датам | Группировка по клиенту | Группировка по товару |
| --- | --- | --- |
| ▪ анализ динамики продаж<br>▪ выявление тенденций и т.д. | ▪ оптимизация работы с клиентами, например, предоставление скидок наиболее активным и т.д. | ▪ определение наиболее и наименее продаваемых товаров<br>▪ оценка вклада конкретного товара в общий объем продаж и т.д. |

# Функции агрегации



# Группировка строковых значений

# Операции  с датой и временем

| Дата | Значение 1 | Значение 2 | Значение N |
|---|---|---|---|
| 01.01.2020 | 10 | 25,6 | … |
| 01.02.2020 | 20 | 245,25 | … |
| … | … | … | … |
| 01.12.2020 | 560 | 54,34 | … |

# Операции  с датой и временем

| Дата | Дата (Неделя) | Дата (Месяц) | Дата (Квартал) | Дата (Год) |
|---|---|---|---|---|
| 24.04.2020 | 17 | 04 Апрель | 2 | 2020 |
| 12.08.2020 | 33 | 08 Август | 3 | 2020 |
| 30.12.2020 | 53 | 12 Декабрь | 4 | 2020 |

# Операции с датой и временем

| Дата | Кол-во |
|---|---|
| 02.01.2017 | 250 |
| 03.01.2017 | 230 |
| 04.01.2017 | 345 |
| 05.01.2017 | 215 |
| 06.01.2017 | 312 |
| 07.01.2017 | 124 |
| 08.01.2017 | 321 |
| 09.01.2017 | 234 |
| 10.01.2017 | 243 |
| 11.01.2017 | 312 |
| 12.01.2017 | 321 |
| 13.01.2017 | 267 |
| 14.01.2017 | 351 |
| 15.01.2017 | 216 |
| 16.01.2017 | 187 |
| 17.01.2017 | 179 |
| 18.01.2017 | 261 |
| 19.01.2017 | 305 |
| 20.01.2017 | 156 |


| Дата | Дата (Год + Неделя, Первый день) | Кол-во |
|---|---|---|
| 02.01.2017 | 02.01.2017 | 250 |
| 03.01.2017 | 02.01.2017 | 230 |
| 04.01.2017 | 02.01.2017 | 345 |
| 05.01.2017 | 02.01.2017 | 215 |
| 06.01.2017 | 02.01.2017 | 312 |
| 07.01.2017 | 02.01.2017 | 124 |
| 08.01.2017 | 02.01.2017 | 321 |
| 09.01.2017 | 09.01.2017 | 234 |
| 10.01.2017 | 09.01.2017 | 243 |
| 11.01.2017 | 09.01.2017 | 312 |
| 12.01.2017 | 09.01.2017 | 321 |
| 13.01.2017 | 09.01.2017 | 267 |
| 14.01.2017 | 09.01.2017 | 351 |
| 15.01.2017 | 09.01.2017 | 216 |
| 16.01.2017 | 16.01.2017 | 187 |
| 17.01.2017 | 16.01.2017 | 179 |
| 18.01.2017 | 16.01.2017 | 261 |
| 19.01.2017 | 16.01.2017 | 305 |
| 20.01.2017 | 16.01.2017 | 156 |

# Операции  с датой и временем

| Время | Время (часы) | Время (минуты) | Время (секунды) |
|---|---|---|---|
| 12:10:39 | 12 | 10 | 39 |
| 13:15:20 | 13 | 15 | 20 |
| 16:20:44 | 16 | 20 | 44 |
| 18:09:56 | 18 | 09 | 56 |
| 19:15:30 | 19 | 15 | 30 |
| 21:22:50 | 21 | 22 | 50 |
| 23:12:40 | 23 | 12 | 40 |

**Проверим, насколько хорошо вы усвоили материал.**

**Количество вопросов**: 3

ETL-процесс состоит из следующих этапов:

1. [ ]
2. Преобразование
3. Загрузка

Какие функции агрегации **недоступны** для строковых значений?

- [ ] медиана
- [ ] максимум
- [ ] среднее
- [ ] сумма
- [ ] количество

К дате **27.05.2020** было применено преобразование **Год+Квартал, Первый день**. Выберете вариант, соответствующий результату преобразования.

- 01.05.2020
- 25.05.2020
- 01.04.2020
- 27.05.2020

# 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 Концепция хранилища данных. Понятие хранилища данных

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!