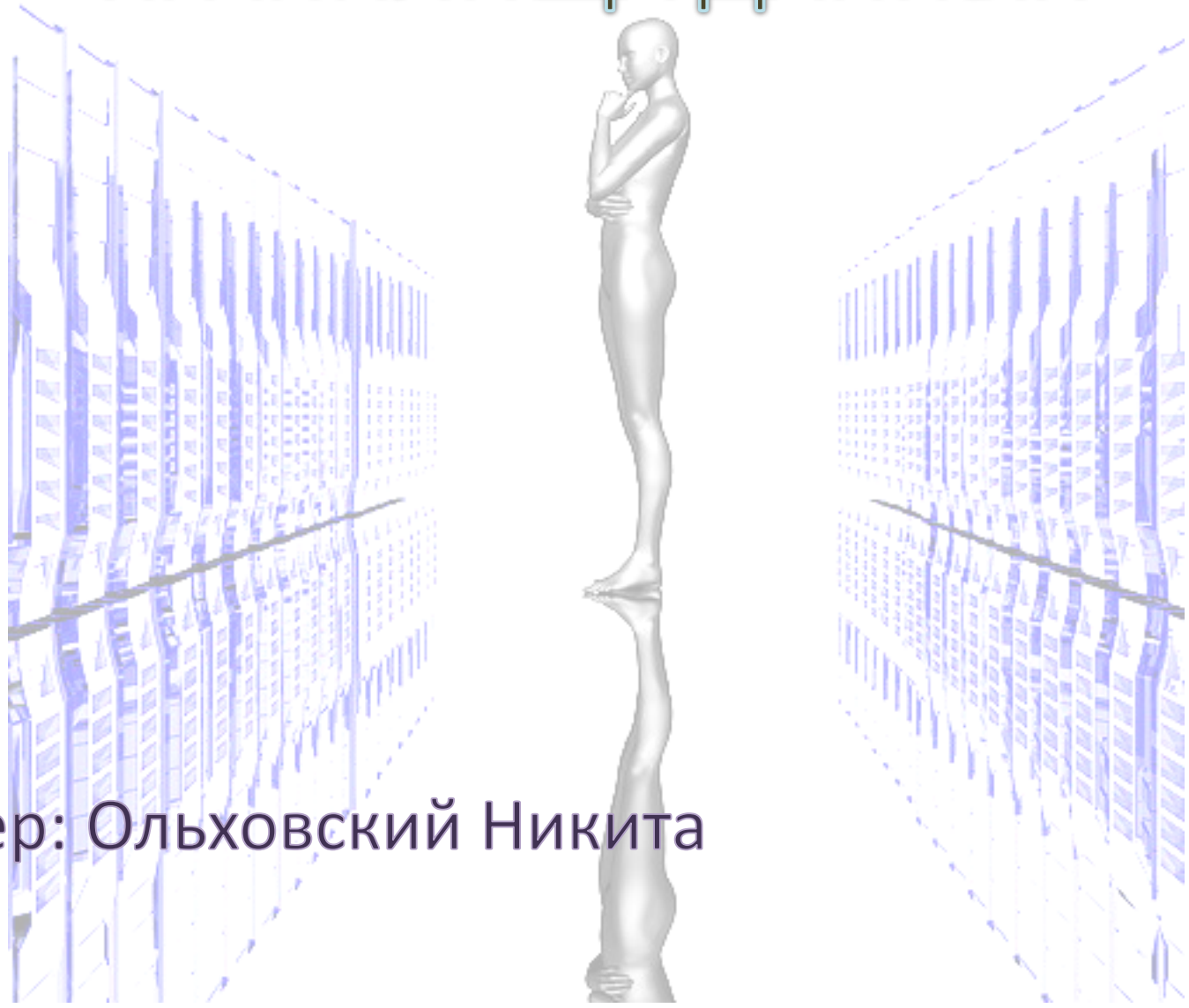
ВВЕДЕНИЕ В BUSINESS INTELLIGENCE. ХРАНИЛИЩА ДАННЫХ



Тренер: Ольховский Никита

Цели курса

По завершению этого тренинга, вы

- получите базовые знания о Business Intelligence
- познакомитесь с понятиями «хранилище данных» и ETL
- рассмотрите основные методы анализа данных

Что такое Business Intelligence (BI)

Пять стилей BI и их применение для корпоративных приложений

Корпоративная отчетность

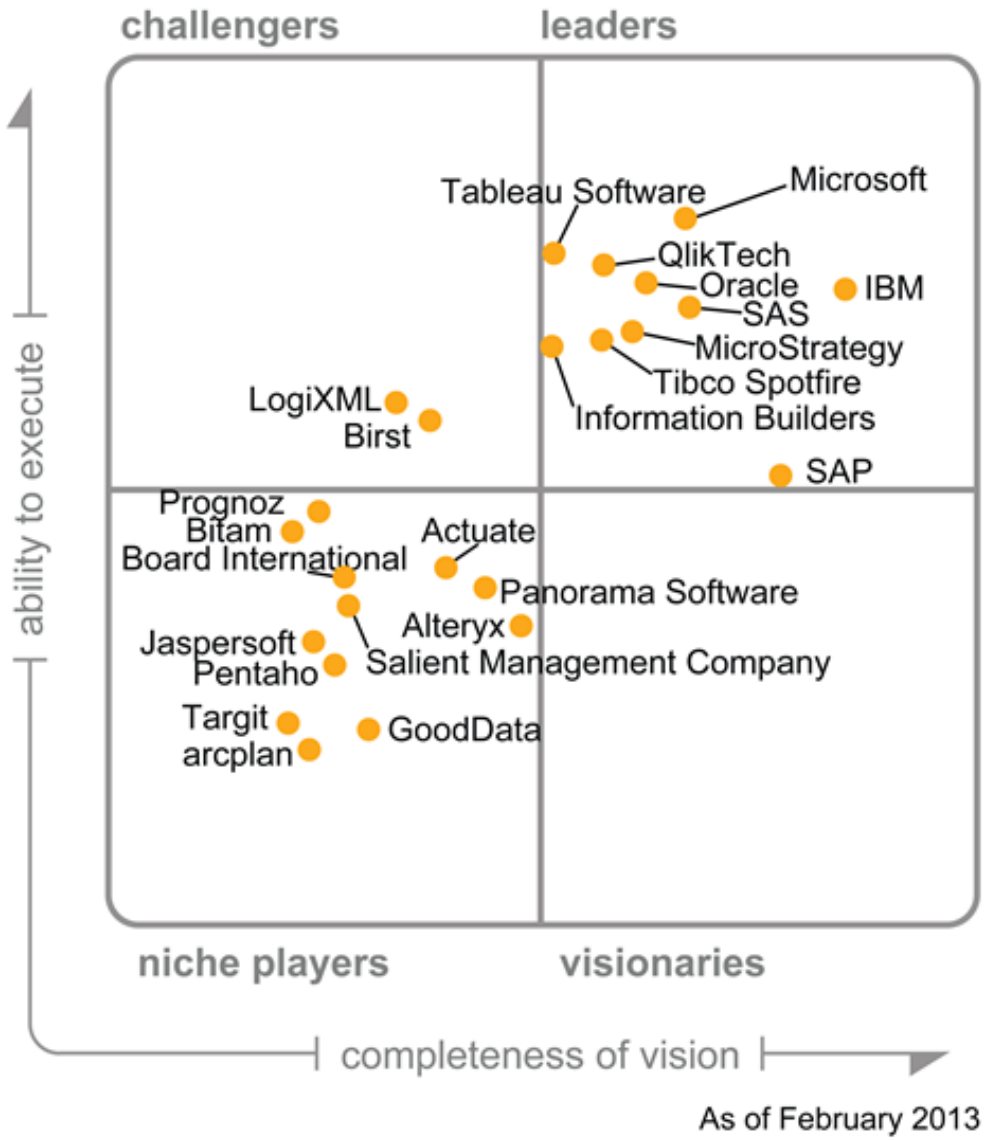
Анализ кубов

Ad-Hoc анализ

Статистика и Data Mining

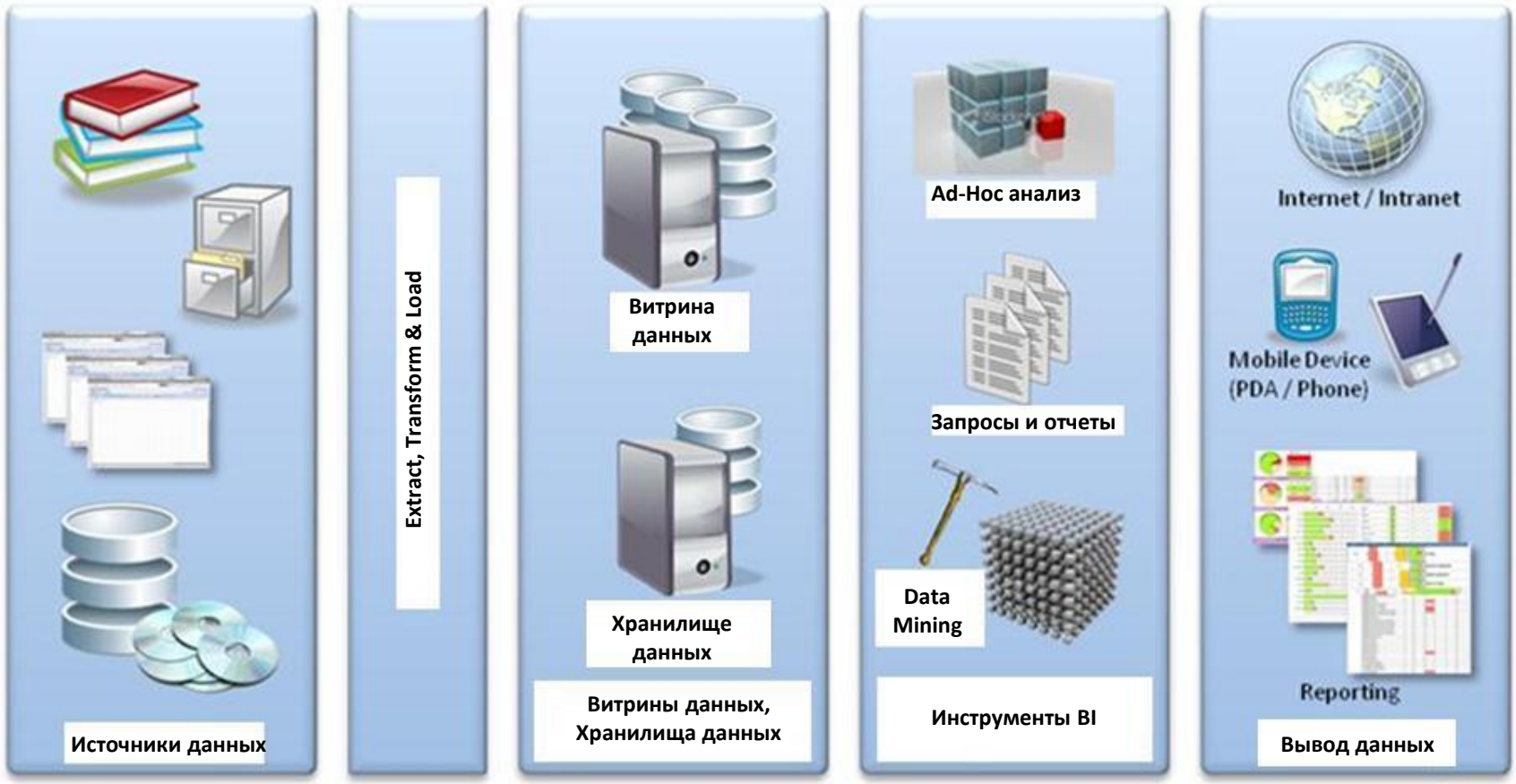
Оповещения и предупреждения

Ключевые игроки рынка BI



- Microsoft
- IBM
- QlikTech
- SAP
- Information Builders
- Tibco Spotfire
- MicroStrategy
- SAS
- Oracle
- Tableau Software

Business Intelligence



Жизненный цикл Business Intelligence

Модель уровней зрелости BI. Этапы развития BI

Управленческие отчеты	Табличные витрины	Витрины данных	Хранилища данных	Корпоративное хранилище данных	BI-сервисы	Технологическая основа модели данных
• система	• сотрудник	•отдел	•филиал	•корпорация	•Международная корпорация	Уровень
• Типовые статические отчеты	• Таблицы (MS Excel, MS Access)	• OLAP •Ad hoc отчеты	• Dashboards (Панели) • Scorecards	• прогнозная (предсказа-тельная) аналитика	•клиентские BI •встроенные BI	Аналитические инструменты

Источник: TDWI (Ассоциация по продвижению технологии Хранилищ данных)

OLTP и OLAP-системы

База данных (БД) - это данные, организованные в виде набора записей определенной структуры и хранящиеся в файлах, где, помимо самих данных, содержится описание их структуры.

Система управления базами данных (СУБД) - это система, обеспечивающая ввод данных в БД, их хранение и восстановление в случае сбоев, манипулирование данными, поиск и вывод данных по запросу пользователя.

Существуют:

- *системы оперативной обработки транзакций (OLTP-системы, Online Transaction Processing);*
- *системы делового анализа (OLAP-системы, Online Analysis Processing).*

Денормализация

REG_ID	REGION
MSK	Moscow
SPB	Saint Petersburg
KRSN	Krasnoyarsk

TAR_ID	TARIFS	MIN_INC	SUBS_FEE
MM500	Manager 500	500	1000
MM300	Manager 300	300	500
PP20	Primary 20	0	0

SUBS_ID	SUBS_FNM	SUBS_SNM	SUBS_REG	SUBS_TAR	SUBS_CUR_ACC
1	Nikita	Mikhalkov	MSK	MM500	540
2	Sergey	Esenin	SPB	MM300	70
3	Anton	Kolnikov	MSK	PP20	120

Денормализация

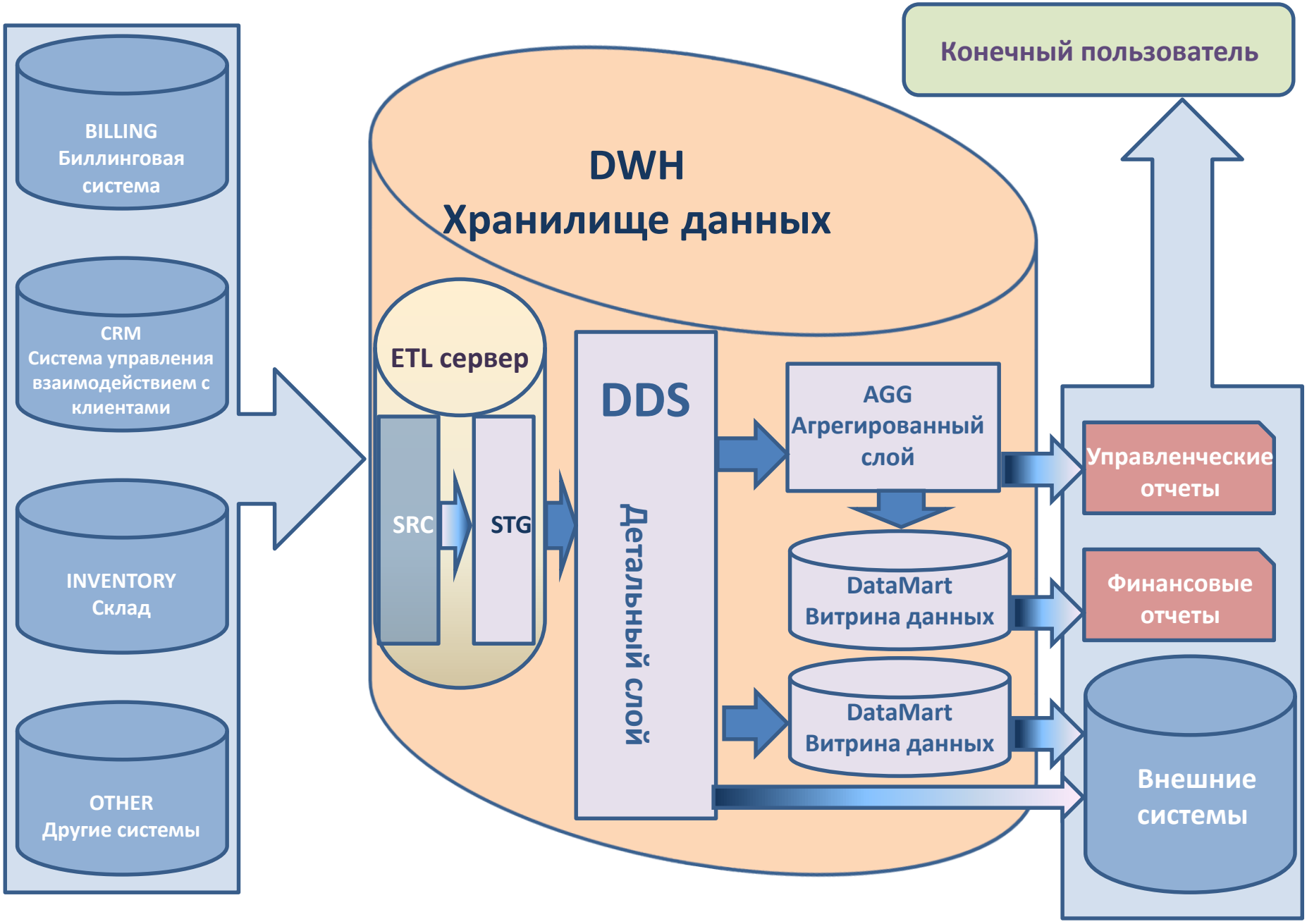
SUBS_ID	SUBS_FNM	SUBS_SNM	SUBS_REG	SUBS_REG_FULL	SUBS_TAR	SUBS_TAR_FULL	MIN_INC	SUBS_FEE	SUBS_CUR_ACC
1	Nikita	Mikhalkov	MSK	Moscow	MM500	Manager 500	500	1000	540
2	Sergey	Esenin	SPB	Saint Petersburg	MM300	Manager 300	300	500	70
3	Anton	Kolnikov	MSK	Moscow	PP20	Primary 20	0	0	120

Хранилища данных

Хранилище данных (англ. *Data Warehouse*) — очень большая предметно-ориентированная информационная корпоративная база данных, специально разработанная и предназначенная для подготовки отчётов, анализа бизнес-процессов с целью поддержки принятия решений в организации.

Основные признаки хранилища данных

- Предметная ориентированность
- Унификация
- Временная привязка
- Неразрушаемая совокупность данных


Конечный пользователь
BILLING
Биллинговая система
CRM
Система управления взаимодействием с клиентами
INVENTORY
Склад
OTHER
Другие системы
DWH
Хранилище данных
ETL сервер
SRC
STG
DDS
Детальный слой
AGG
Агрегированный слой
DataMart
Витрина данных
DataMart
Витрина данных
Управленческие отчеты
Финансовые отчеты
Внешние системы

Структура Хранилища Данных

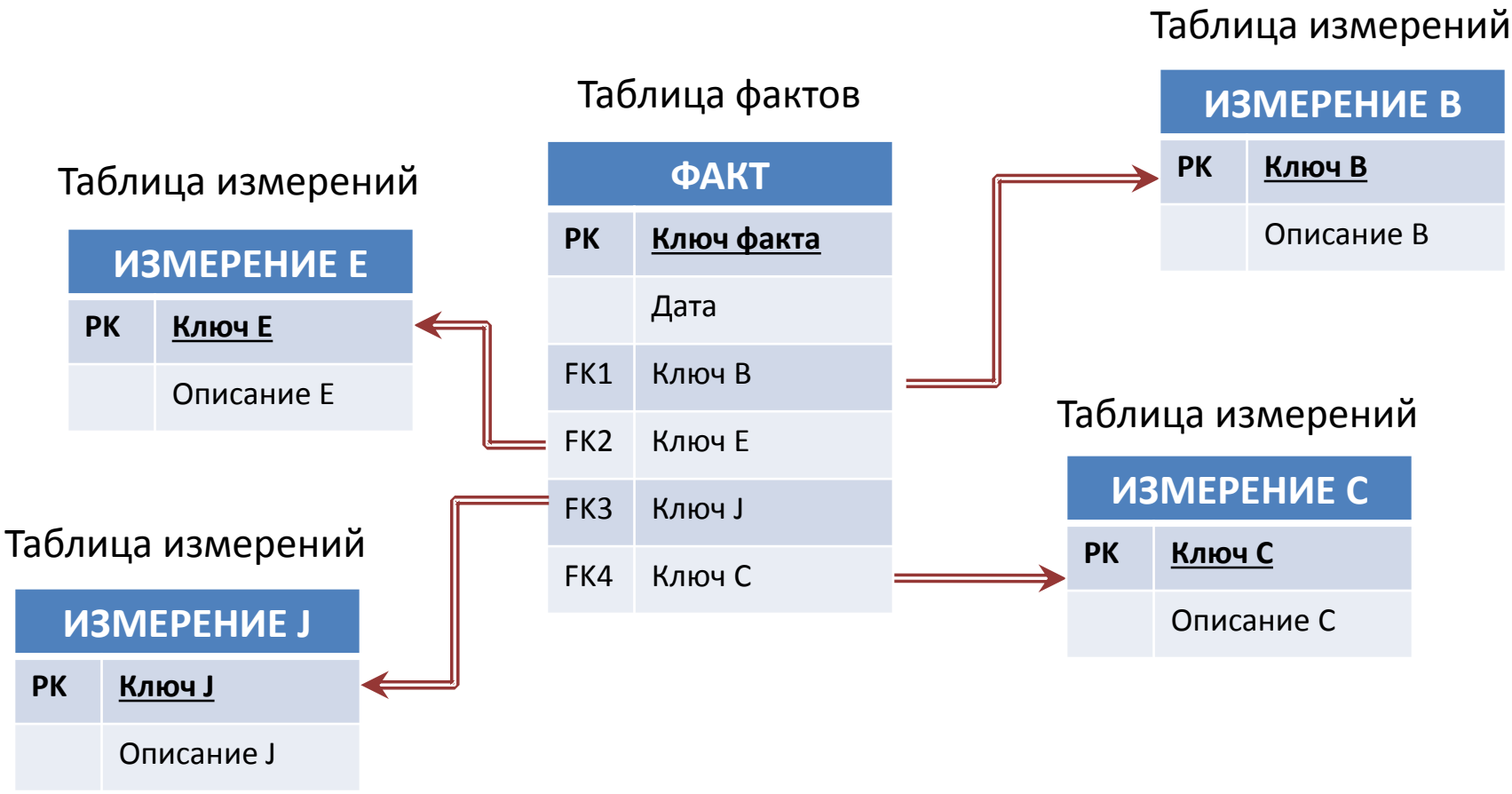


Таблица фактов

Таблица фактов - основная таблица хранилища данных. Содержит сведения об объектах или событиях, совокупность которых будет в дальнейшем анализироваться.

- **факты, связанные с транзакциями (Transaction facts). Пример – снятие денег за телефонный разговор.**
- **факты, связанные с «моментальными снимками» (Snapshot facts). Пример – дневная выручка.**
- **факты, связанные с элементами документа (Line-item facts). Пример – информация о чеке.**
- **факты, связанные с событиями или состоянием объекта (Event or state facts). Пример – смена тарифного плана.**

Таблицы измерений

Таблица измерений – это справочник данных для ХД. В таблицах измерений хранятся данные, описывающие записи из таблицы фактов.

Особенности таблиц измерений:

- **Таблицы измерений содержат неизменяемые либо редко изменяемые данные.**
- **Каждая таблица измерений должна находиться в отношении «один ко многим» с таблицей фактов.**
- **скорость роста таблиц измерений должна быть незначительной по сравнению со скоростью роста таблицы фактов**

Типы измерений

- **Type 0** – неизменяемый тип измерения. Данные, внесенные изначально, с течением времени не изменяются.

- **Type 1** – тип измерения без истории. Данные, вносимые позже, заменяют исходные данные без ведения истории.

CITY	NAME	NUMBER
Москва	Николай	12345

CITY	NAME	NUMBER
Москва	Николай	10456

- **Type 2** – версионный тип измерения. Вносимые данные помечаются версией (дата начала и окончания, либо номера версий) и заносятся в разные строки.

CITY	NAME	NUMBER	VALID _DTTM	INVALID_DTTM
Москва	Николай	12345	10.04.2007	31.12.2999

CITY	NAME	NUMBER	VALID _DTTM	INVALID_DTTM
Москва	Николай	12345	10.04.2007	09.06.2010
Москва	Николай	10000	10.06.2010	31.12.2999

Типы измерений

- **Type 3** – версионный тип измерения. Вносимые данные заносятся в одну строку с разделением на оригинальную версию и текущую. При обновлении данных PREV_DATE заменяется на CUR_DATE, а в CUR_DATE ставится дата новой записи

CUR_CITY	CUR_NAME	CUR_NUMBER	CUR_DATE	PREV_DATE	PREV_CITY	PREV_NAME	PREV_NUMBER
Москва	Николай	**12345**	**12.01.2010**	**31.12.2999**	Москва	Николай	**12345**

CUR_CITY	CUR_NAME	CUR_NUMBER	CUR_DATE	PREV_DATE	PREV_CITY	PREV_NAME	PREV_NUMBER
Ростов	Николай	**14785**	**24.06.2010**	**12.01.2010**	Москва	Николай	**12345**

CUR_CITY	CUR_NAME	CUR_NUMBER	CUR_DATE	PREV_DATE	PREV_CITY	PREV_NAME	PREV_NUMBER
Самара	Николай	**17649**	**05.11.2010**	**24.06.2010**	Ростов	Николай	**14785**

Схема «Снежинка»

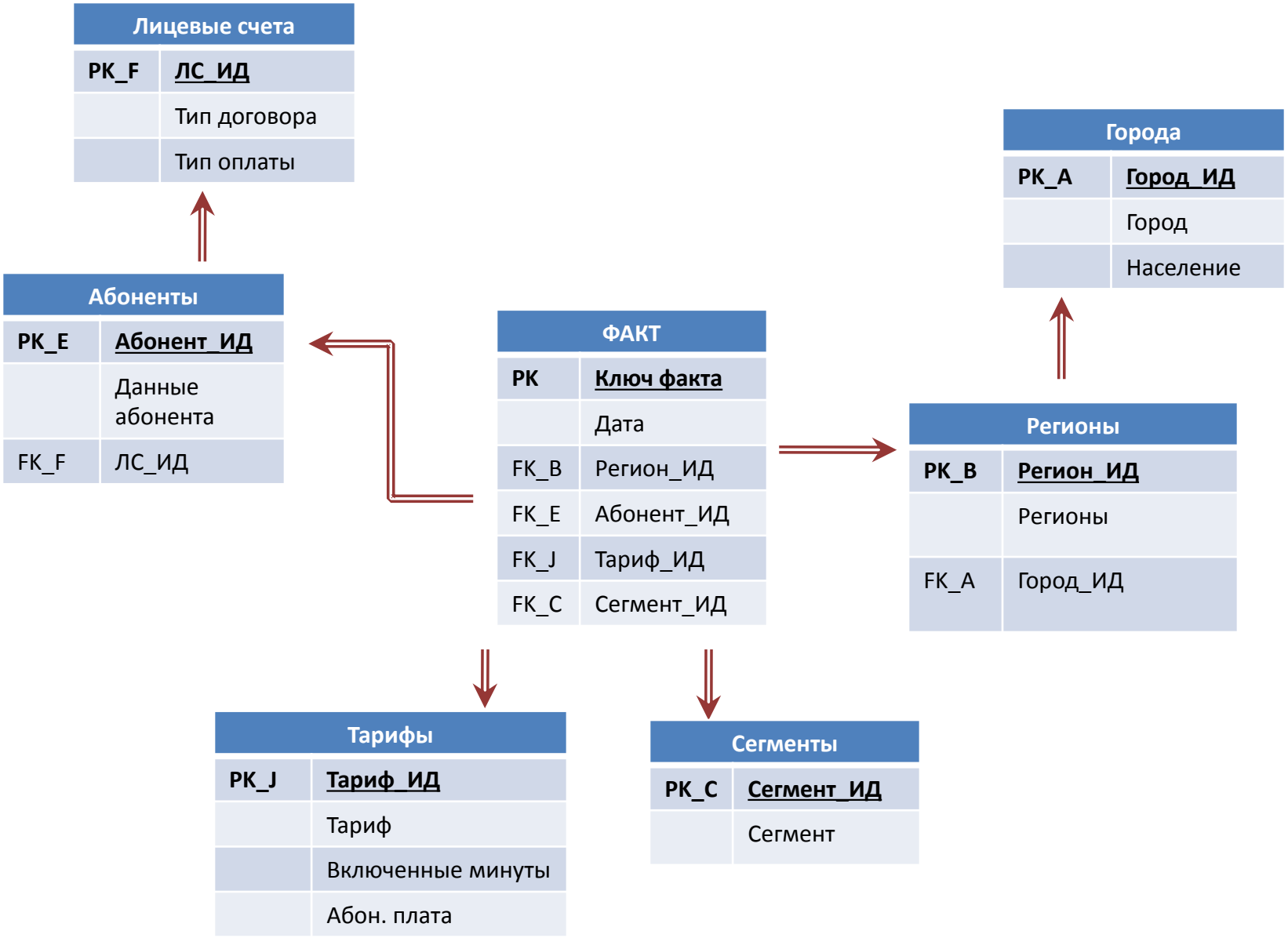
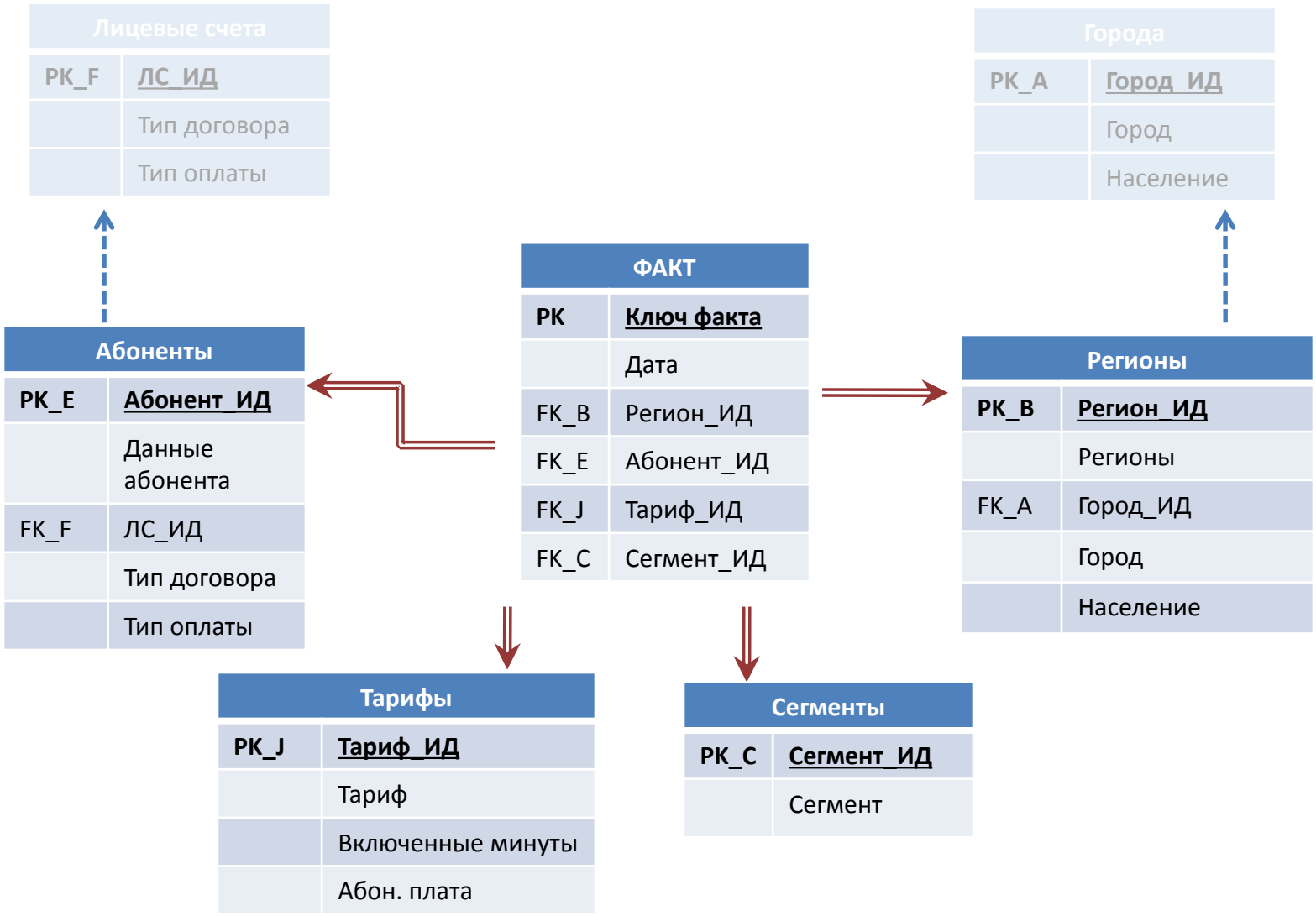
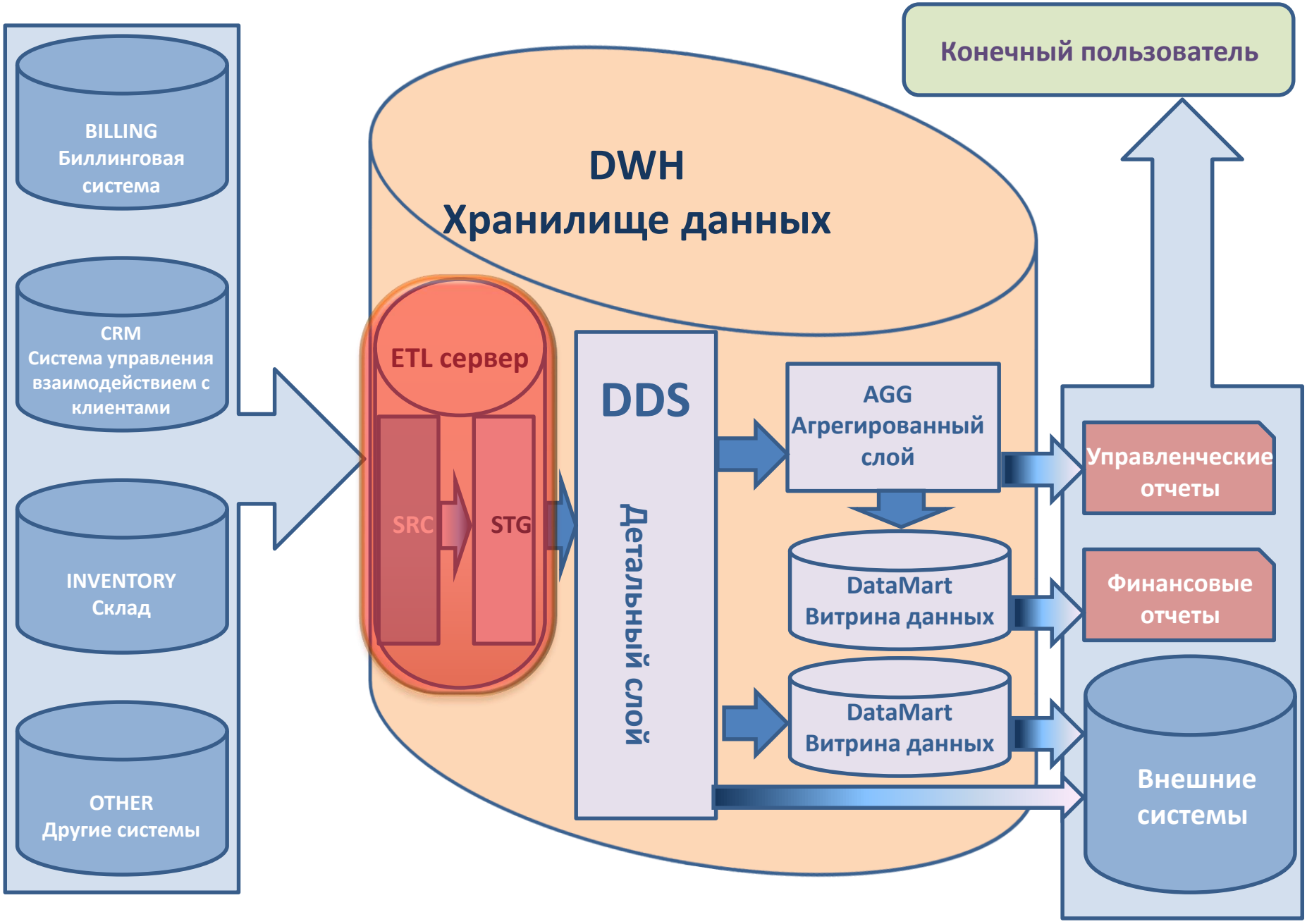
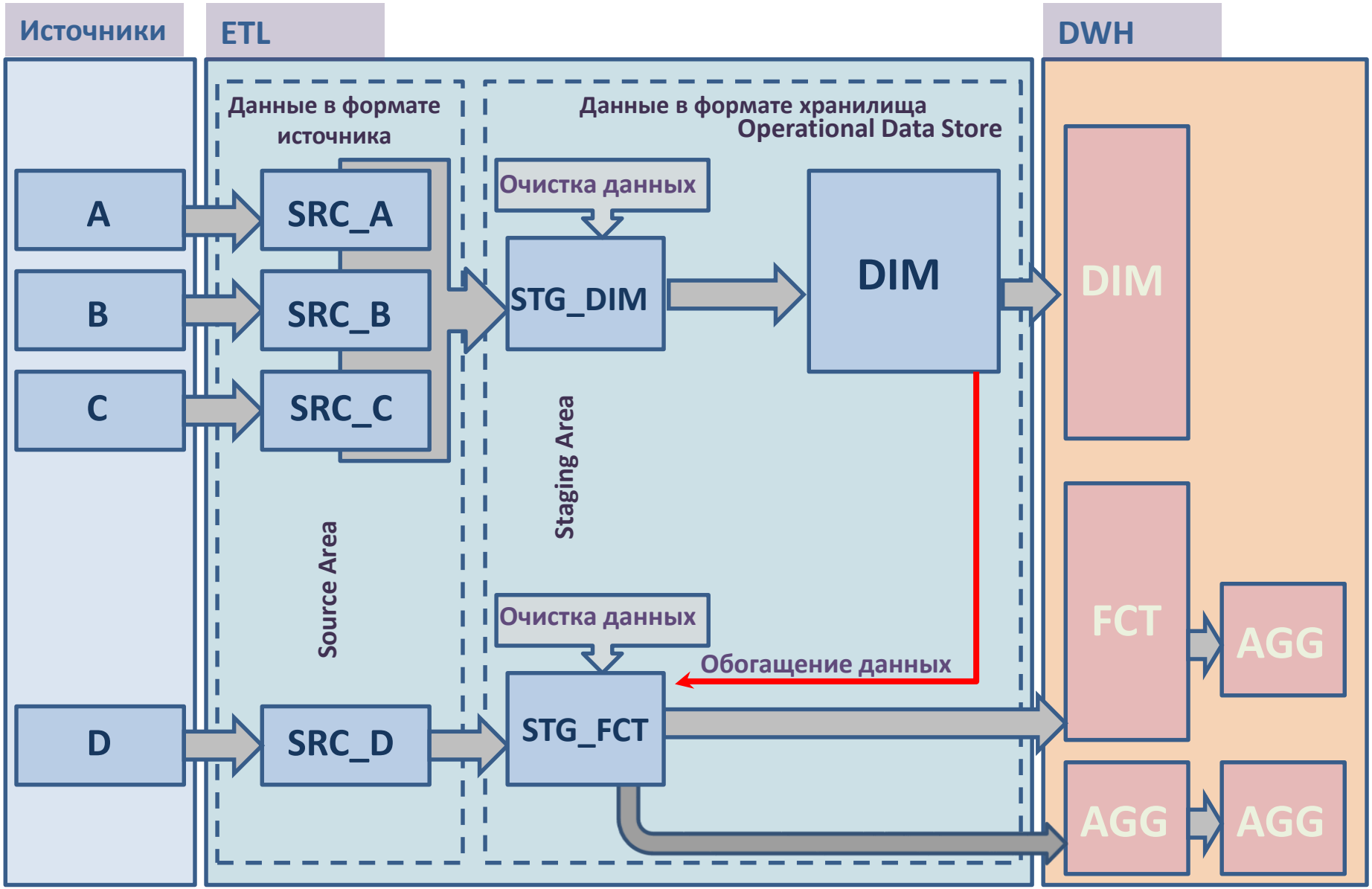


Схема «Звезда»

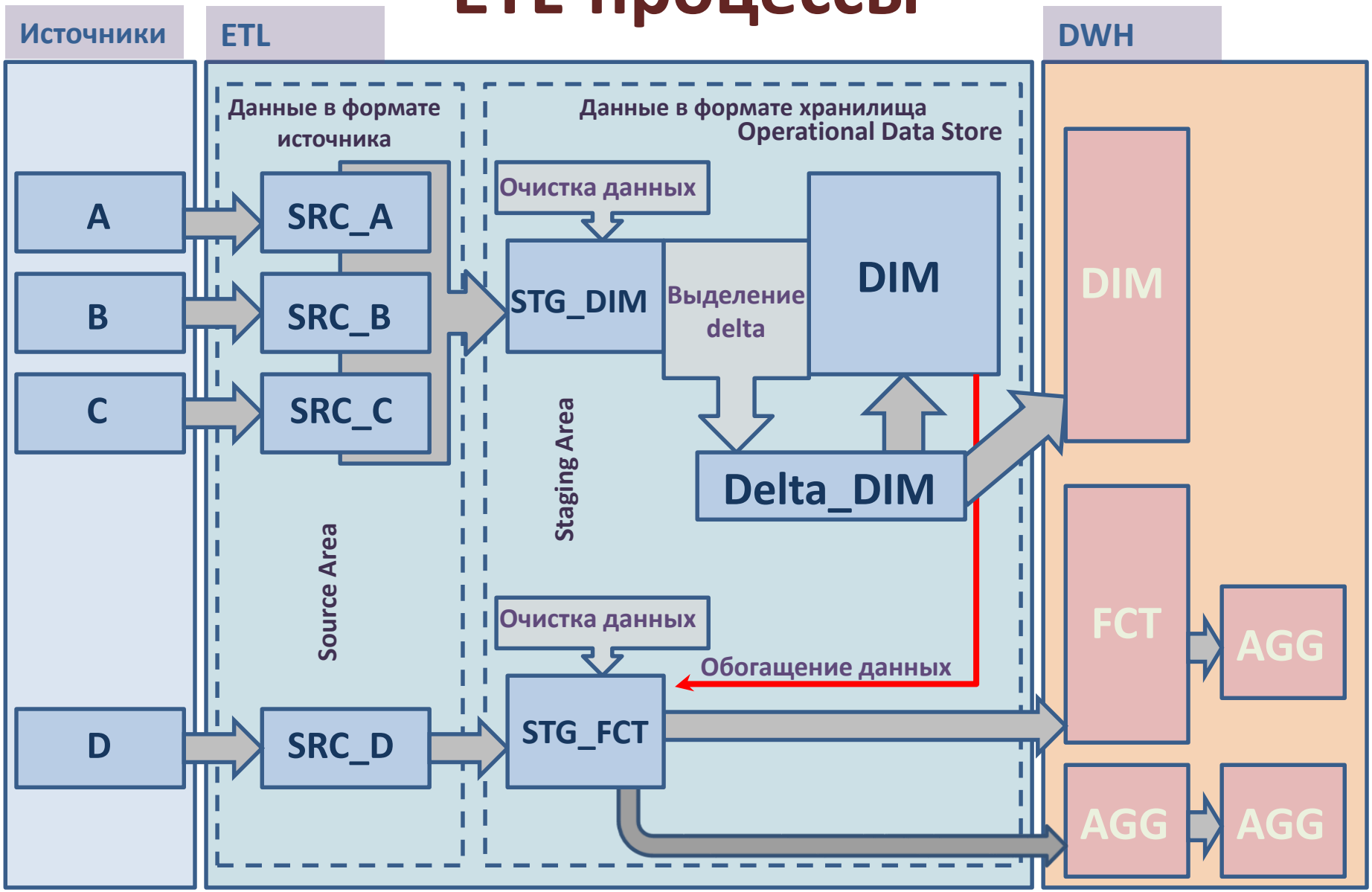



Конечный пользователь
DWH
Хранилище данных
BILLING
Биллинговая система
CRM
Система управления взаимодействием с клиентами
INVENTORY
Склад
OTHER
Другие системы
ETL сервер
SRC
STG
DDS
Детальный слой
AGG
Агрегированный слой
DataMart
Витрина данных
DataMart
Витрина данных
Управленческие отчеты
Финансовые отчеты
Внешние системы

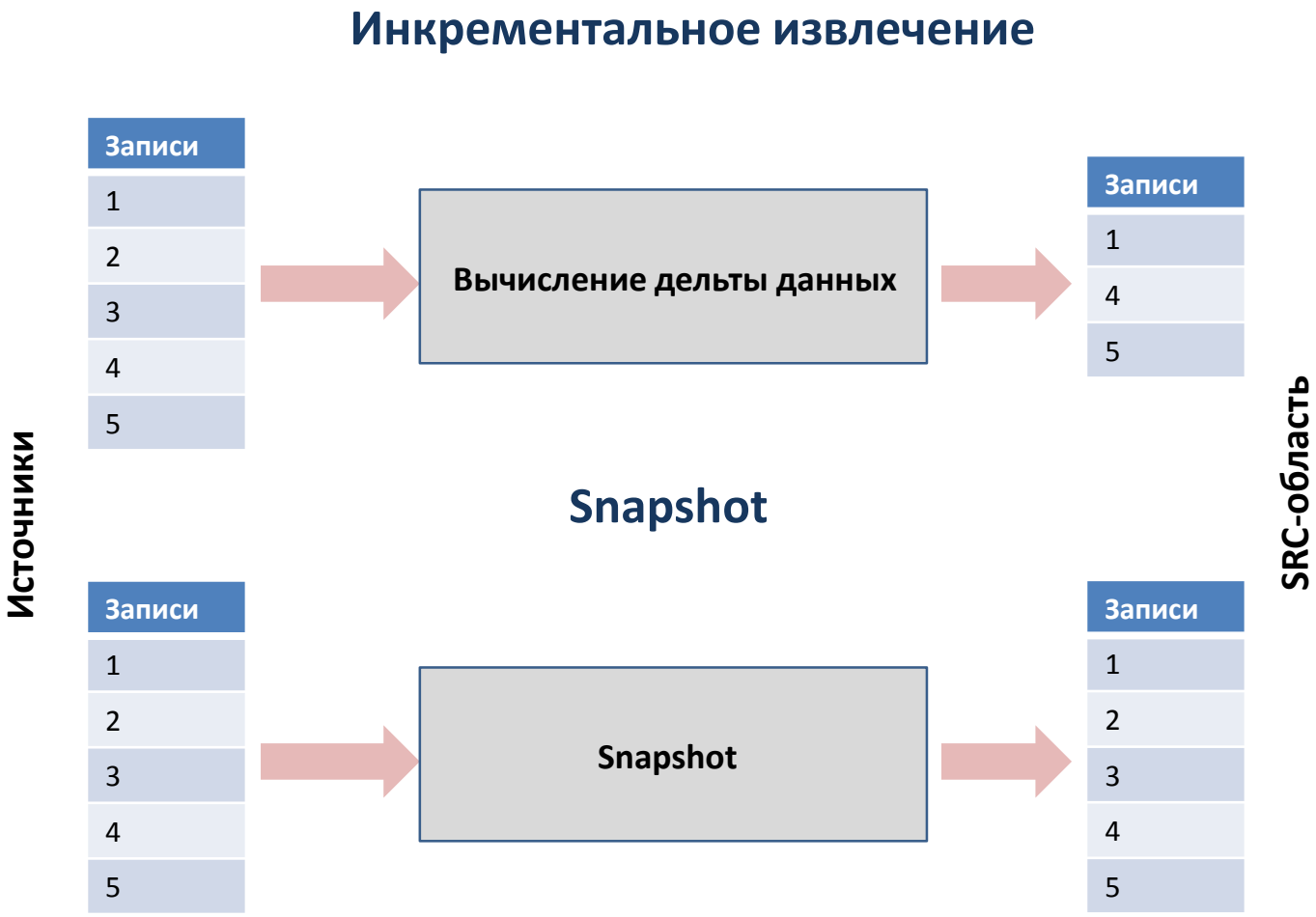
Первичная загрузка данных в хранилище. ETL-процессы



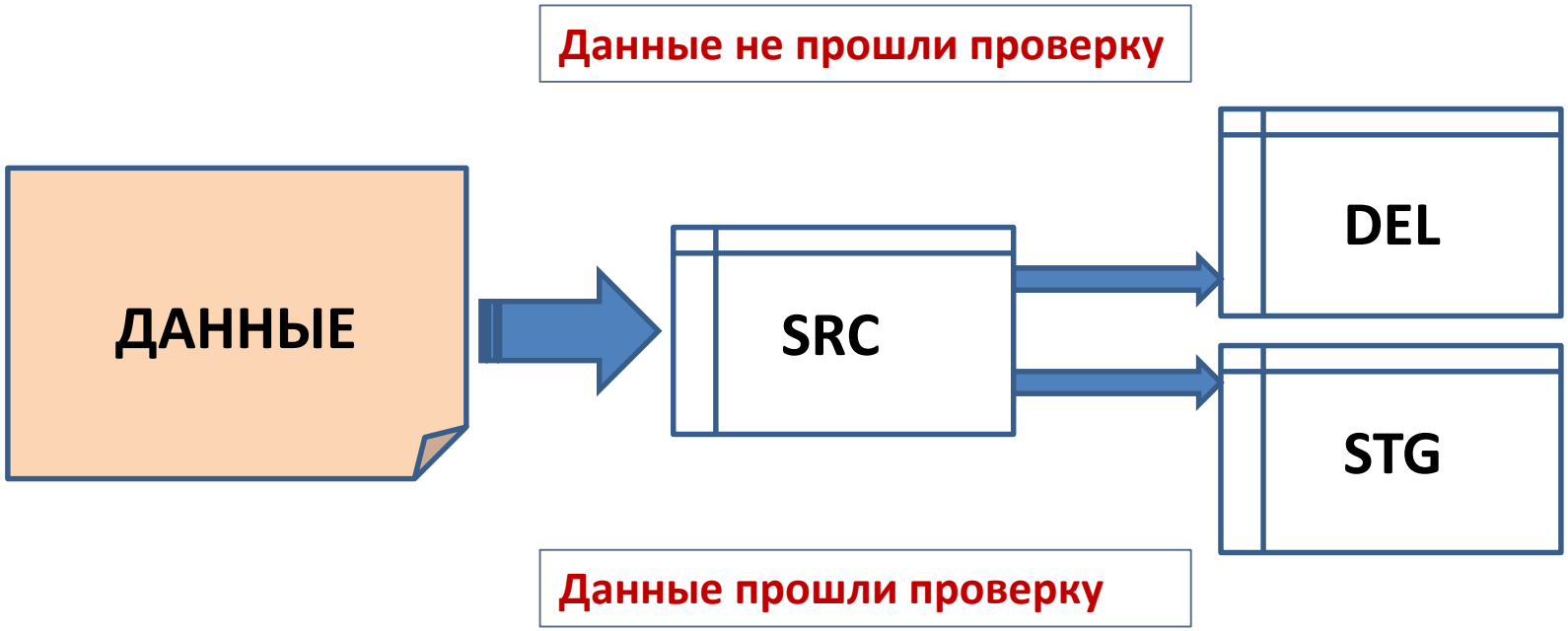
Загрузка данных в хранилище. ETL-процессы



Извлечение данных



Обработка данных



Почему данные не проходят проверку

Категории критериев оценки качества данных:

По критичности:

- Критичные ошибки в данных
- Некритичные ошибки в данных
- Качественные данные

По проверяемым объектам:

- Корректность форматов и представлений данных
- Уникальность первичных и альтернативных ключей
- Полнота данных
- Полнота связей
- Соответствие данных аналитическим ограничениям

Обогащение данных

SUBSCRIBERS			
Абонент	Фамилия	Имя	Счет
SUBS_KEY	SURNAME	NAME	ACCOUNT
122	Колосков	Кирилл	11246
244	Игнатьев	Илья	14677
746	Сергеев	Антон	45666

TARIFS			
Тариф	Название	Регион	Сегмент
TAR_KEY	NAME	REGION_KEY	SEGMENT_KEY
MAN500	Менеджер-500	MSK	HI
PP10	Потреб-10-10	SPB	LO
MAN300	Менеджер-300	MSK	MED

FACTS				
Дата	№	Абонент	Тариф	Списание
DATE	PR_KEY	SUBS_KEY	TAR_KEY	CHARGE
01.09.2010	144	122	MAN500	3250
01.09.2010	145	244	PP10	4500
01.09.2010	146	746	MAN300	550

DDS								
Дата	№	Абонент	Счет	Тариф	Название	Регион	Сегмент	Списание
DATE	PR_KEY	SUBS_KEY	ACCOUNT	TAR_KEY	NAME	REGION_KEY	SEGMENT_KEY	CHARGE
01.09.2010	144	122	11246	MAN500	Менеджер-500	MSK	HI	3250
01.09.2010	145	244	14677	PP10	Потреб-10-10	SPB	LO	4500
01.09.2010	146	746	45666	MAN300	Менеджер-300	MSK	MED	550

Слои данных

- Детализированное представление данных
- Агрегированное представление данных
- Витрины данных (DataMart)

Детализированное представление данных

FCT_USAGE						
Абонент	Регион	Тарифный план	Сегмент	Списание	Длительность	Дата звонка
SUBS_KEY	REGION_KEY	PLAN_KEY	SEGMENT_KEY	CHARGE	CALL_DURATION	CALL_DTTM
121	MSK	MAN500	HI	50	5	01.09.2010
112	SPB	PP10	LO	10	2	01.09.2010
146	SPB	MAN300	MED	21	1	01.09.2010
876	MSK	MAN500	HI	43	1	01.09.2010
786	SPB	MAN500	HI	8	3	01.09.2010
121	MSK	MAN300	MED	35	4	01.09.2010
458	SPB	PP10	LO	3	7	01.09.2010
211	SPB	PP10	MED	6	9	01.09.2010
453	MSK	MAN300	MED	17	3	01.09.2010
906	SPB	MAN300	MED	8	4	01.09.2010
112	MSK	PP10	LO	21	7	01.09.2010
458	MSK	MAN300	MED	14	11	01.09.2010

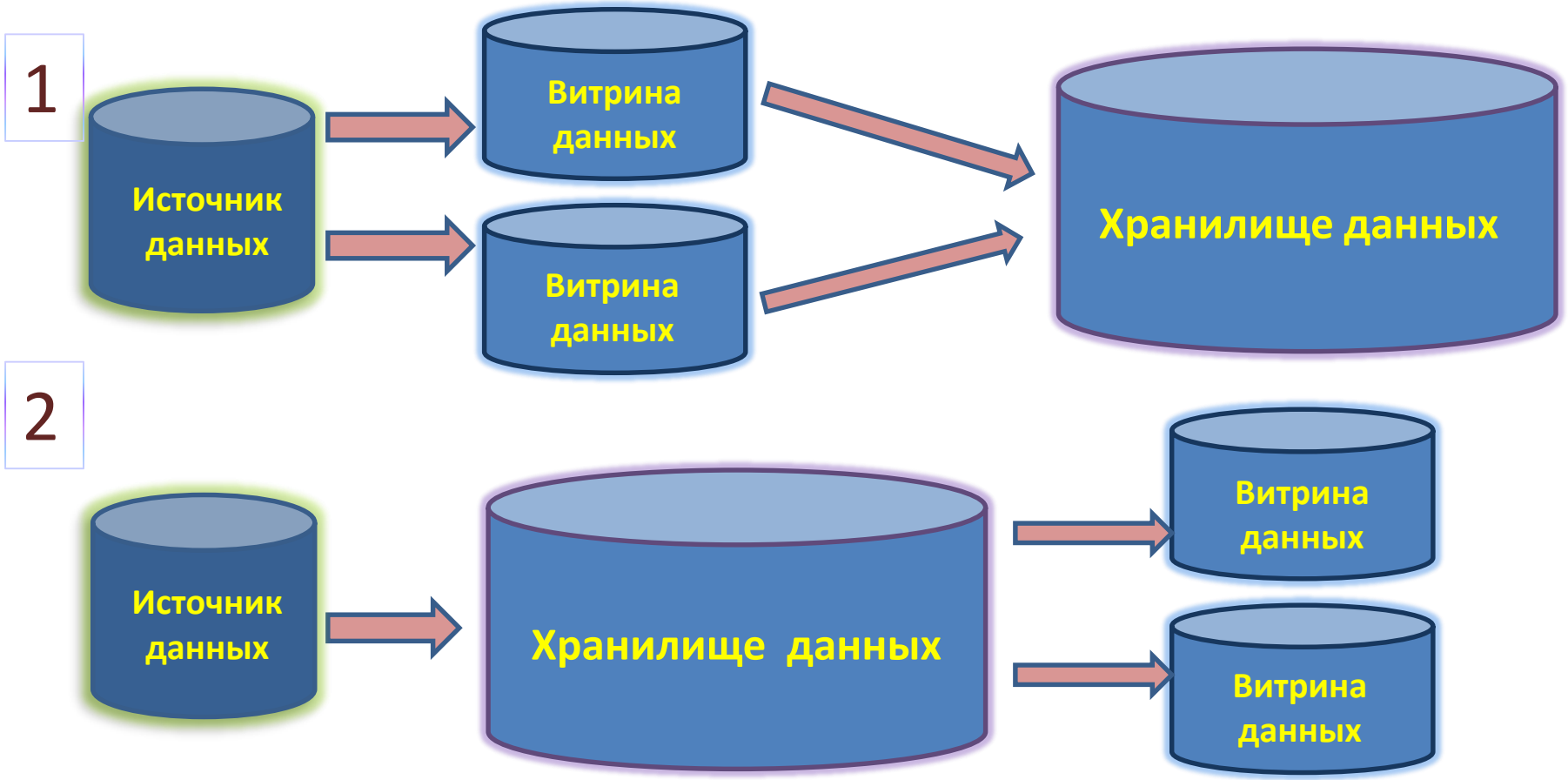
Агрегированное представление данных

FCT_USAGE						
Абонент	Регион	Тарифный план	Сегмент	Списание	Длительность	Дата звонка
SUBS_KEY	REGION_KEY	PLAN_KEY	SEGMENT_KEY	CHARGE	CALL_DURATION	CALL_DTTM
121	MSK	MAN500	HI	50	5	01.09.2010
112	SPB	PP10	LO	10	2	01.09.2010
146	SPB	MAN300	MED	21	1	01.09.2010
876	MSK	MAN500	HI	43	1	01.09.2010
786	MSK	MAN500	HI	8	3	01.09.2010
121	MSK	MAN300	MED	35	4	01.09.2010

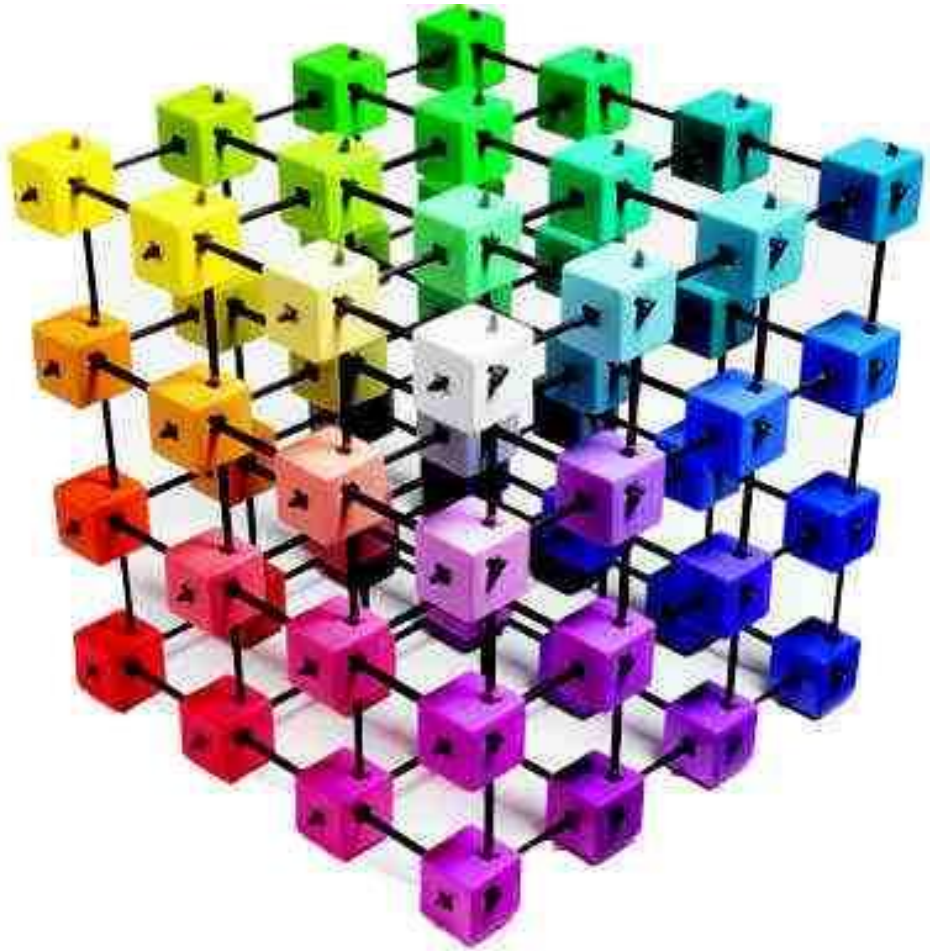
AGG_USAGE					
Дата звонка	Регион	Тарифный план	Сегмент	Сумма списаний	Общая длительность
CALL_DTTM	REGION_KEY	PLAN_KEY	SEGMENT_KEY	CHARGE_AMT	CALL_DURATION_AMT
01.09.2010	MSK	MAN500	HI	101	9
01.09.2010	SPB	PP10	LO	10	2
01.09.2010	SPB	MAN300	MED	21	1
01.09.2010	MSK	MAN300	MED	35	4

Витрины данных

Витрина данных — срез хранилища данных, представляющий собой массив тематической, узконаправленной информации, ориентированный на пользователей одной рабочей группы или департамента.



OLAP (MOLAP, ROLAP, HOLAP)



Тест FASMI

Fast	Быстрый
Analysis	Анализ
of Shared	Разделяемой
Multidimensional	Многомерной
Information	Информации

Классификация продуктов OLAP по способу представления данных

- *Многомерный OLAP (MOLAP)*
- *Реляционный OLAP (ROLAP)*
- *Гибридный OLAP (HOLAP)*

Многомерный OLAP (MOLAP)

В специализированных СУБД, основанных на многомерном представлении данных, данные организованы не в форме реляционных таблиц, а в виде упорядоченных многомерных массивов:

1) гиперкуб

2) поликуб

Плюсы

- **Быстрый поиск**
- **Многомерные СУБД легко справляются с задачами включения в информационную модель разнообразных встроенных функций**

Минусы

- **Многомерные СУБД не позволяют работать с большими базами данных**
- **Многомерные СУБД очень неэффективно используют внешнюю память**

Реляционный OLAP (ROLAP)

- В реляционных OLAP-системах структура куба данных хранится в реляционной базе данных. Меры самого нижнего уровня остаются в реляционной витрине данных, служащей источником данных для куба. Предварительно обработанные агрегаты также хранятся в реляционной таблице.

Плюсы

- **Обычно ХД – реляционные БД. инструменты ROLAP позволяют производить анализ непосредственно над ними.**
- **ROLAP - системы с динамическим представлением**
- **Высокий уровень защиты данных и возможности разграничения прав доступа**

Минусы

- **Меньшая производительность**
- **Для приемлемой производительности требуется тщательная проработка схемы базы данных и настройки индексов**

Гибридный OLAP (HOLAP)

- В HOLAP-системах структура куба и предварительно обработанные агрегаты хранятся в многомерной базе данных. Это позволяет обеспечить быстрое извлечение агрегатов из структур MOLAP. Значения нижнего уровня иерархии в HOLAP остаются в реляционной витрине данных, которая служит источником данных для куба.

Плюсы	• **Комбинирование технологии ROLAP для разреженных данных и MOLAP для плотных областей**
Минусы	• **Необходимость поддерживания ROLAP и MOLAP**

Интеллектуальный анализ данных

- **Интеллектуальный анализ данных**— это процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей (шаблонов информации)

- *Выявление закономерностей (свободный поиск)*
- *Использование выявленных закономерностей для предсказания неизвестных значений (прогностическое моделирование)*
- *Анализ исключений, предназначенный для выявления и толкования аномалий в найденных закономерностях*

DSS-системы

Уровень пользователя

- Активные
- Пассивные
- Кооперативные

Концептуальный уровень

- Управляется сообщениями
- Управляется данными
- Управляется документами
- Управляется знаниями
- Управляется моделями

Технический уровень

- СППР предприятия
- Настольная СППР

Уровень данных

- Оперативные
- стратегические

Программное обеспечение, используемое в блоке BI

Хранилища данных	ETL-инструменты	BI-системы	OLAP-продукты
• Oracle • TeraData • DB2 • SyBase • SAS BIS • SAS TIS	• Informatica • IBM DataStage • Oracle Data Integrator • SAS Data Integration Studio • SAP Data Integrator	• Business Objects • MicroStrategy • Cognos • Oracle BI • SAS BI	• Oracle Hyperion/EssBase • Microsoft OLAP • SAS OLAP Studio

Итоги курса

Прослушав этот тренинг, вы

- ✓ **получили базовые знания о Business Intelligence**
- ✓ **познакомились с понятиями «хранилище данных» и ETL**
- ✓ **рассмотрели основные методы анализа данных**