ВВЕДЕНИЕ В ETL

ПОЧЕМУ ИНТЕГРАЦИЯ ДАННЫХ?

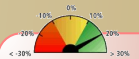
НУЖНО...

Информация там и в том виде в каком необходимо

Business Intelligence

Corporate Performance Management

Business Process Management



Business Activity Monitoring

Интеграция данных

Migration

Data Warehousing

Master Data Management

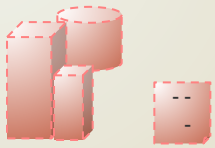
Data Synchronization

Federation

SOA (Messaging)

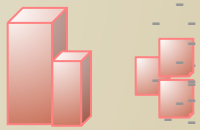
ИМЕЕМ...

Данные в несогласованных источниках

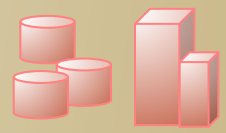


Legacy

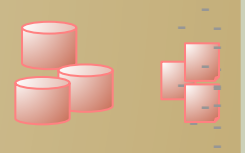
ERP



CRM



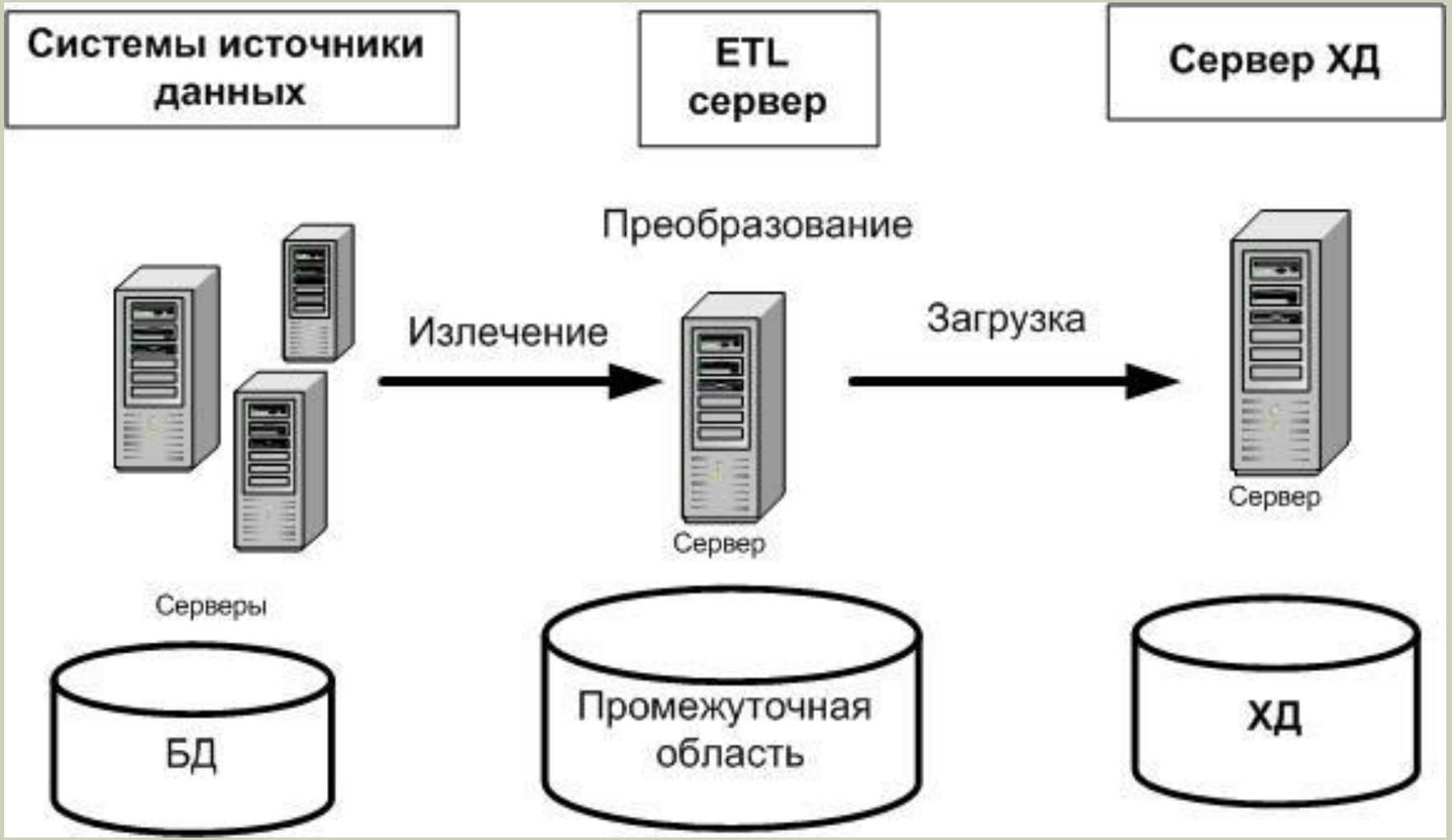
Best-of-breed Applications



ПРОБЛЕМА КОНСОЛИДАЦИИ ДАННЫХ

Извлечение данных из разнотипных источников и перенос их в хранилище данных с целью дальнейшей аналитической обработки связаны с рядом проблем, основными из которых являются нижеследующие.

- Исходные данные расположены в источниках самых разнообразных типов и форматов, созданных в различных приложениях, и, кроме того, могут использовать различную кодировку, в то время как для решения задач анализа данные должны быть преобразованы в единый универсальный формат, который поддерживается хранилищем и аналитическим приложением.
- Данные в источниках обычно излишне детализированы, тогда как для решения задач анализа в большинстве случаев требуются обобщенные данные.
- Исходные данные, как правило, являются «грязными», то есть содержат различные факторы, которые мешают их корректному анализу.


Системы источники данных
ETL сервер
Сервер ХД
Преобразование
Излечение
Загрузка
Сервер
Сервер
Серверы
БД
Промежуточная область
ХД

ETL

- Поэтому для переноса исходных данных из различных источников в ХД следует использовать специальный инструментарий, который должен извлекать данные из источников различного формата, преобразовывать их в единый формат, поддерживаемый ХД, а при необходимости — производить очистку данных от факторов, мешающих корректно выполнять их аналитическую обработку.
- Такой комплекс программных средств получил обобщенное название ETL (от англ. extraction, transformation, loading — «извлечение», «преобразование», «загрузка»). Сам процесс переноса данных и связанные с ним действия называются ETL-процессом, а соответствующие программные средства — ETL-системами.

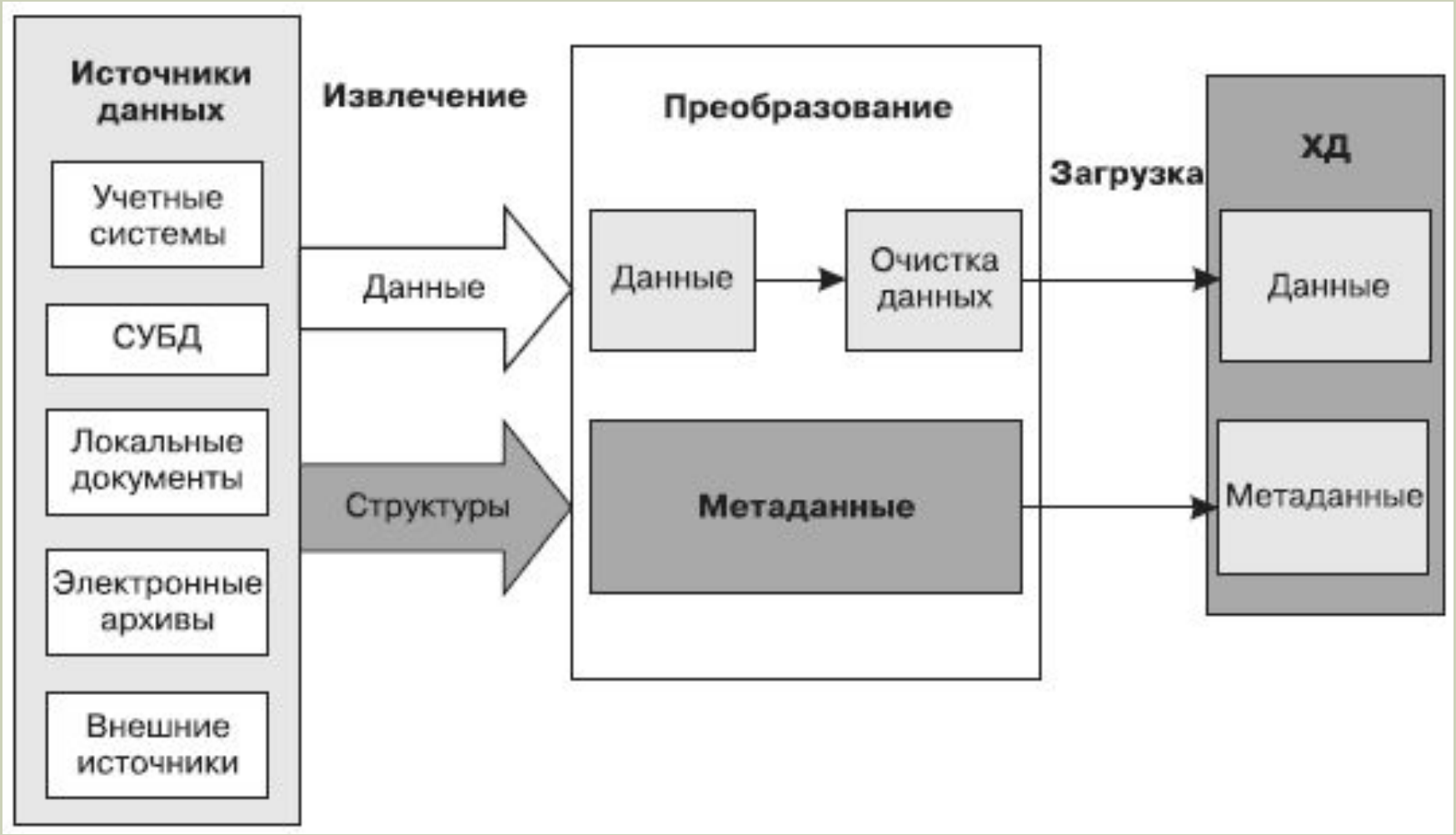
ETL — комплекс методов, реализующих процесс переноса исходных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных.

3 ПРОЦЕССА

Независимо от особенностей построения и функционирования ETL-система должна обеспечивать выполнение трех основных этапов процесса переноса данных (ETL-процесса).

- **Извлечение данных**. На этом этапе данные извлекаются из одного или нескольких источников и подготавливаются к преобразованию. Следует отметить, что для корректного представления данных после их загрузки в ХД из источников должны извлекаться не только сами данные, но и информация, описывающая их структуру, из которой будут сформированы метаданные для хранилища.
- **Преобразование данных.** Производятся преобразование форматов и кодировки данных, а также их обобщение и очистка.
- **Загрузка данных** — запись преобразованных данных в соответствующую систему хранения.

ПЕРЕМЕЩЕНИЕ ДАННЫХ В ПРОЦЕССЕ ETL



ИЗВЛЕЧЕНИЕ ДАННЫХ В ETL

Процесс извлечения данных в рамках ETL существенно зависит от типов и структуры источников данных. Можно выделить три разновидности источников данных, с которыми чаще всего сталкиваются организаторы аналитических проектов.

- *Базы данных* (SQL Server, Oracle, Firebird, Access и т.д.). В большинстве случаев извлечение данных из баз данных не вызывает проблем, поскольку структура данных в них жестко задана
- *Структурированные файлы различных форматов*. Такие файлы очень широко распространены, поскольку средства их создания общедоступны и не требуют высокой квалификации персонала и высокой производительности систем. К таким источникам относятся текстовые файлы с разделителями, файлы электронных таблиц (например, Excel, CSV-файлы, HTML-документы и т.д.). Здесь проблем больше, поскольку пользователь может допускать ошибки, пропуски, вводить противоречивые данные, терять фрагменты данных и т.д. Единственным плюсом является то, что для доступа к типовым структурированным данным можно применять такие стандартные средства, как ODBC и ADO.
- *Неструктурированные источники*.Если избежать использования неструктурированных источников не получается, нужно применить специальные средства их преобразования в структурированный вид. Когда источник невелик, возможно, это удастся сделать вручную. Но в большинстве случаев приходится разрабатывать специальный инструментарий, учитывающий особенности организации данных в источнике и то, какую структуру из них следует создать.

ОЧИСТКА ДАННЫХ В ETL. ДВА УРОВНЯ ОЧИСТКИ ДАННЫХ

- Наличие «грязных» данных — одна из важнейших и трудно формализуемых проблем аналитических технологий вообще и ХД в частности.
- Очистка данных обязательна при их перегрузке в хранилище, и при разработке стратегии ETL этому уделяется большое внимание.
- Следует отметить, что, помимо очистки данных перед их загрузкой в хранилище, пользователь может выполнить дополнительную очистку средствами аналитической системы уже после выполнения запроса к ХД.
- Существует несколько проблем, из-за которых данные нуждаются в очистке. Наиболее широко распространены проблемы, связанные с нарушением структуры данных:
  - корректность форматов и представлений данных;
  - уникальность первичных ключей в таблицах БД;
  - полнота и целостность данных;
  - полнота связей;
  - соответствие некоторым аналитическим ограничениям и т.д.

ПРЕОБРАЗОВАНИЕ ДАННЫХ В ETL

Цель этого этапа — подготовка данных к размещению в ХД и приведение их к виду, наиболее удобному для последующего анализа. При этом должны учитываться некоторые выдвигаемые аналитиком требования, в частности, к уровню качества данных. В процессе преобразования данных в рамках ETL чаще всего выполняются следующие операции

преобразование структуры данных:

- агрегирование данных;
- перевод значений;
- создание новых данных;
- очистка данных.

ЗАГРУЗКА ДАННЫХ В ХРАНИЛИЩЕ

- Процесс загрузки заключается в переносе данных из промежуточных таблиц в структуры хранилища данных. От продуманности и оптимальности процесса загрузки данных во многом зависит время, требуемое для полного цикла обновления данных в ХД, а также полнота и корректность данных в хранилище.
- Первыми в процессе загрузки данных в ХД обычно загружаются таблицы измерений, которые содержат суррогатные ключи и другую описательную информацию, необходимую для таблиц фактов.
- Иногда не все записи могут быть загружены из-за несоответствия структуре, например