МОДЕЛИРАНЕ И АНАЛИЗ НА СОФТУЕР

Павел Кюркчиев
Ас. към ПУ „Паисий Хилендарски“
https://github.com/pkyurkchiev
@pkyurkchiev

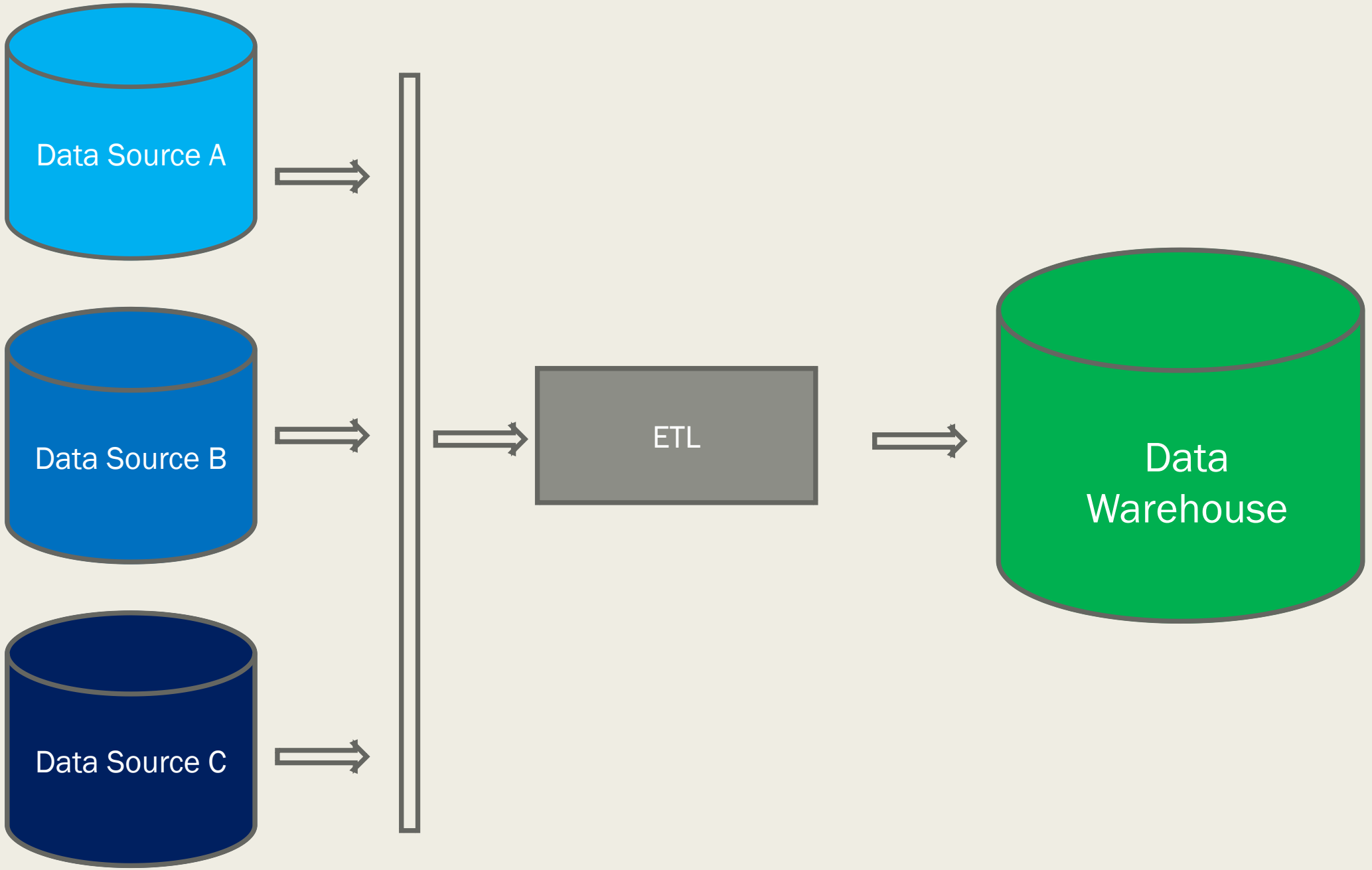
СКЛАД ЗА ДАННИ

Склад данни (Data warehouse)

- Склад за данни се нарича голяма база данни, която може да има достъп до цялата информация на дадена компания или организация. Тази огромна база данни съхранява информацията като хранилище за данни, и предоставя данните за извършване на анализи.

Според Уилям Инмон

- Складът данни е „тематично ориентирана, интегрирана, времевариантна, неизменчива съвкупност от данни, подпомагаща вземането на решения“.


Data Source A
Data Source B
Data Source C
ETL
Data Warehouse

Видове хранилища

- Складове за данни (Data warehouses)
- Бази данни (Databases)
- Езеро с данни (Data lakes)

- Складовете за данни съхраняват информация от множество източници и използват предварително дефинирани схеми, предназначени са за анализ на данни.
- Базата данни се използва за улавяне и съхранение на данни от един източник. Базите данни не са проектирани да работят с голям набор от данни.
- Езерото от данни представлява централно хранилище за данни. Може да съдържа всички видове сурови данни, структурирани или неструктурирани, не зависимо от източниците. Схемата не е дефинирана, което позволява различни видове анализи, в сравнение със складовете за данни, които имат дефинирани схеми. Езерото за данни може да се използват за текстови търсения, машинно обучение и анализи в реално време.

Общи характеристики

- Използва тематично ориентиран пространствен модел на данните.
- Съдържа годни за публикуване данни от множество източници.
- Съдържа интегрирани инструменти за отчети;
- Данните в хранилището за данни са само за четене, което означава, че не могат да бъдат актуализирани, създадени или изтрити.

- Данните от складовете за данни представляват данни за дълъг период от време (10 години и повече), което означава, че съхраняват исторически данни.
- В информационния склад данните се обобщават на различни нива. Потребителят може да започне да разглежда общите продажни единици на продукт в един цял регион. След това да разгледа състоянията в този регион. И накрая, може да преглежда отделните магазини в определена държава. Следователно, анализът започва от по-високо ниво и се придвижва надолу към по-ниското ниво на детайлите.

Твърдения и механизъм за достъп

- Складът за данни може да се разпредели на няколко компютъра и да съдържа няколко бази данни, както и информация от многобройни източници в различни формати. Достъпът до склада за данни се осъществява чрез сървър.
- Достъпа до склада за данни е прозрачен за потребителя, който може да използва прости команди, за да открие и анализира необходимата информация.

- Складът данни съдържа и информация за това как е организиран самият склад, къде може да се открие информацията, както и всички връзки между данните.
- Складът данни се обновява пакетно и е конфигуриран за бързи онлайн заявки, даващи кратки и ясни извлечения от данните.

Ползи от употребата на склад за данни

- Интегриране на данни от множество източници в една база данни и един модел за данни. Позволява използването на единичен механизъм за достъп.
- Ограничаване на проблема с блокирането на транзакции на ниво база данни в системата, причинен от опитите за стартиране на големи и продължителни аналитични заявки в бази данни.

- Поддържане история на данните, дори ако системите източници него правят.
- Интегриране на данни от множество системи източници, което позволява централен изглед в предприятието. Тази полза винаги е ценна, особено когато организацията е нараснала чрез сливане.
- Подобрява качеството на данните, като осигурява постоянни кодове и описания, маркиране или дори фиксиране на лоши данни.
- Представя информацията на организацията последователно.
- Осигурява един общ модел за данни за всички данни, представляващи интерес, независимо от източниците им.

- Преструктурира данните така, че да имат смисъл за бизнес потребителите.
- Преструктурира данните така, че да осигуряват отлична ефективност на заявките дори при сложни аналитични заявки, без да се отразяват на операционните системи.
- Добавя стойност към оперативни бизнес приложения, особено системи за управление на взаимоотношенията с клиенти (CRM).
- Прави заявките за поддръжка на решения по-лесни за писане.
- Организира и обособява повтарящи се данни.

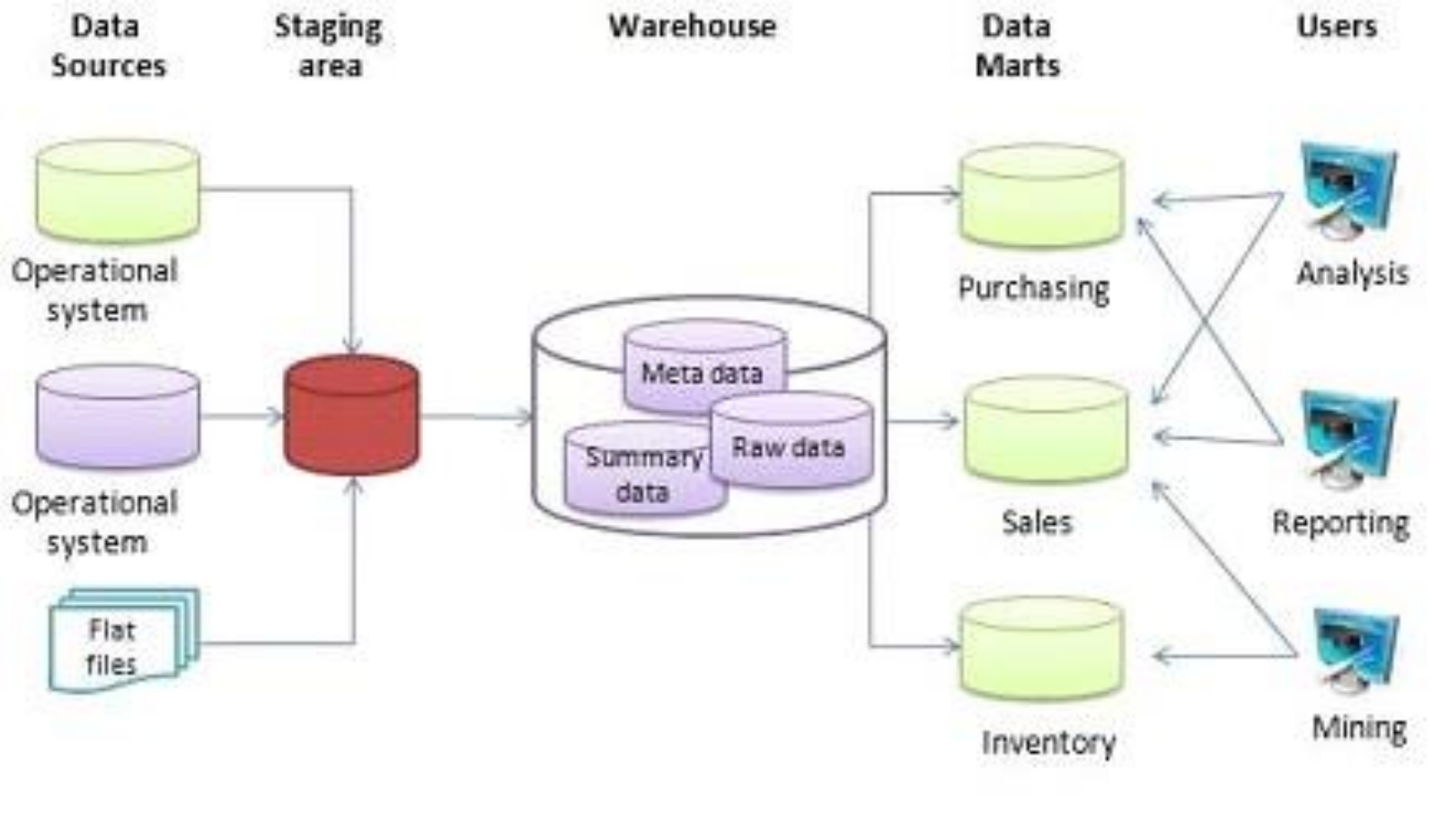
Видове Data warehouse

- ETL based Data warehousing
- ELT based Data warehousing

ETL based Data warehousing

- ETL Data warehouse включва staging, data integration, и access layers. Staging layer съхранява извлечената информация от хетерогенни източници. Integration layer трансформира информацията и след това я премества в data warehouse. Access layer помага на потребителя да достъпва информацията.

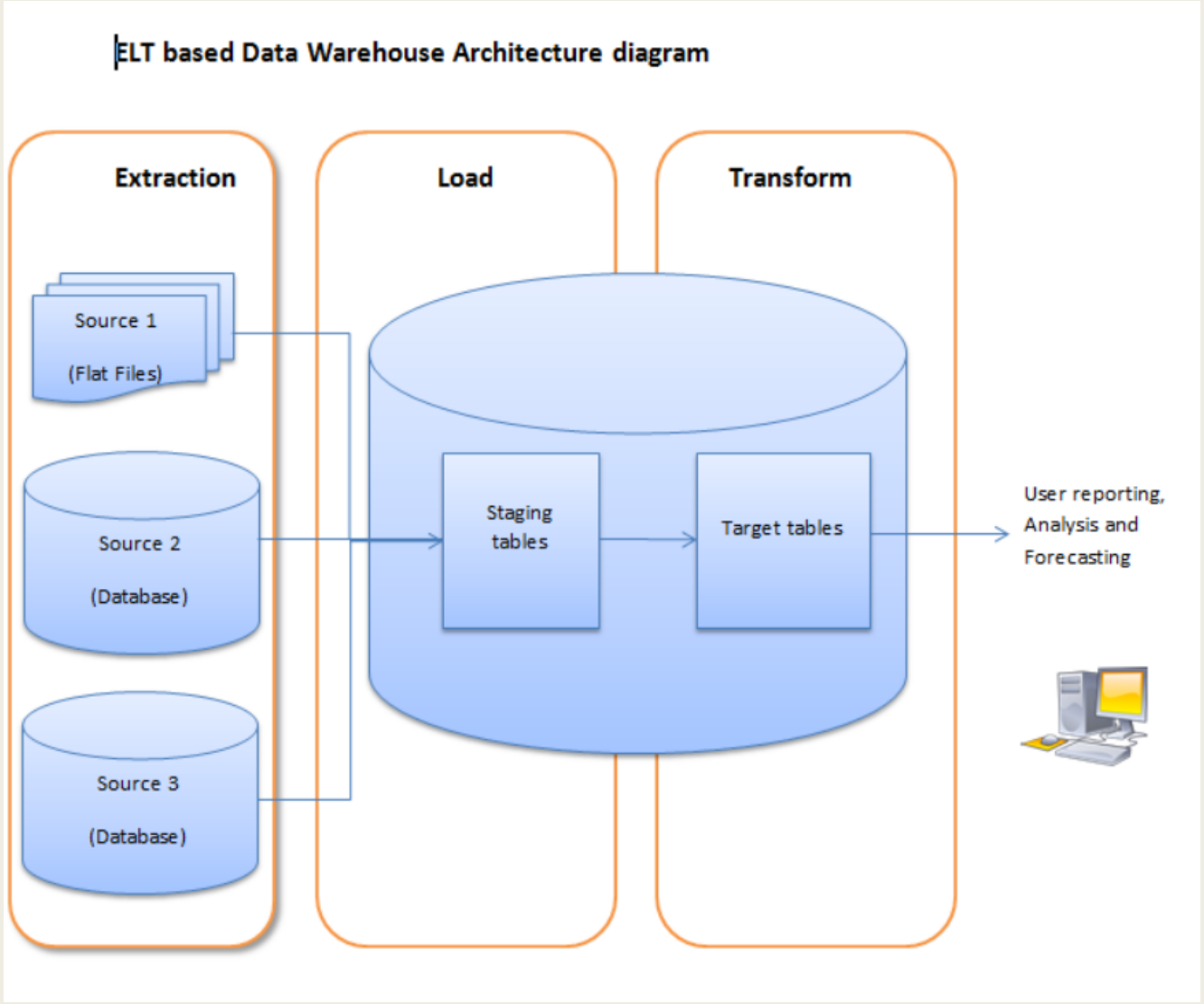
ETL based Data warehousing



ELT based Data warehousing

- ELT data warehouse не съдържа ETL инструменти. Вместо това той поддържа зона за поставяне в самия склад на данни. При този подход данните се извличат от хетерогенни източници на системи и след това се зареждат директно в хранилището на данни, преди да се извърши каквато и да е трансформация. След това всички необходими трансформации се осъществяват в самия склад на данни. Накрая, манипулираните данни се зареждат в целеви таблици.

ELT based Data warehousing



Свързани системи

- Матрица с данни (Data mart)
- Онлайн аналитичната обработка (OLAP)
- Онлайн обработката на транзакции (OLTP)
- Предсказуем анализ (Predictive analytics)

Матрица с данни (Data mart)

- Матрица с данни е проста форма на склад за данни, която е фокусирана върху един обект (или функционална област), поради което черпи данни от ограничен брой източници като продажби, финанси или маркетинг. Матриците за данни често се изграждат и контролират от един отдел в рамките на една организация. Източниците могат да бъдат вътрешни операционни системи, централен склад за данни или външни данни. Денормализирането е основна техника за моделиране на данни в тази система.

Разлика между Data warehouse и Data Mart

Атрибути	Data warehouse	Data mart
Обхват на данните	цяло предприятие	цял отдел
Брой тематични области	многоброен	единичен
Колко е трудно за изработка	труден	лесен
Колко време е необходимо за изработка	много	малко
Размер на паметта	голямо	лимитирано

Онлайн аналитичната обработка (OLAP)

- Онлайн аналитичната обработка се характеризира със сравнително малък обем транзакции. Запитванията често са много сложни и включват обединения. За OLAP системите времето за реакция е мярка за ефективност. OLAP приложенията се използват широко от техниките за извличане на данни. OLAP базите данни съхраняват обобщени исторически данни в многомерни схеми.

Онлайн обработката на транзакции (OLTP)

- Онлайн обработката на транзакции се характеризира с голям брой кратки онлайн транзакции (INSERT, UPDATE, DELETE). Системите OLTP наблягат на много бърза обработка на заявки и поддържане на целостта на данните в среда с множествен достъп.

Предсказуем анализ (Predictive analytics)

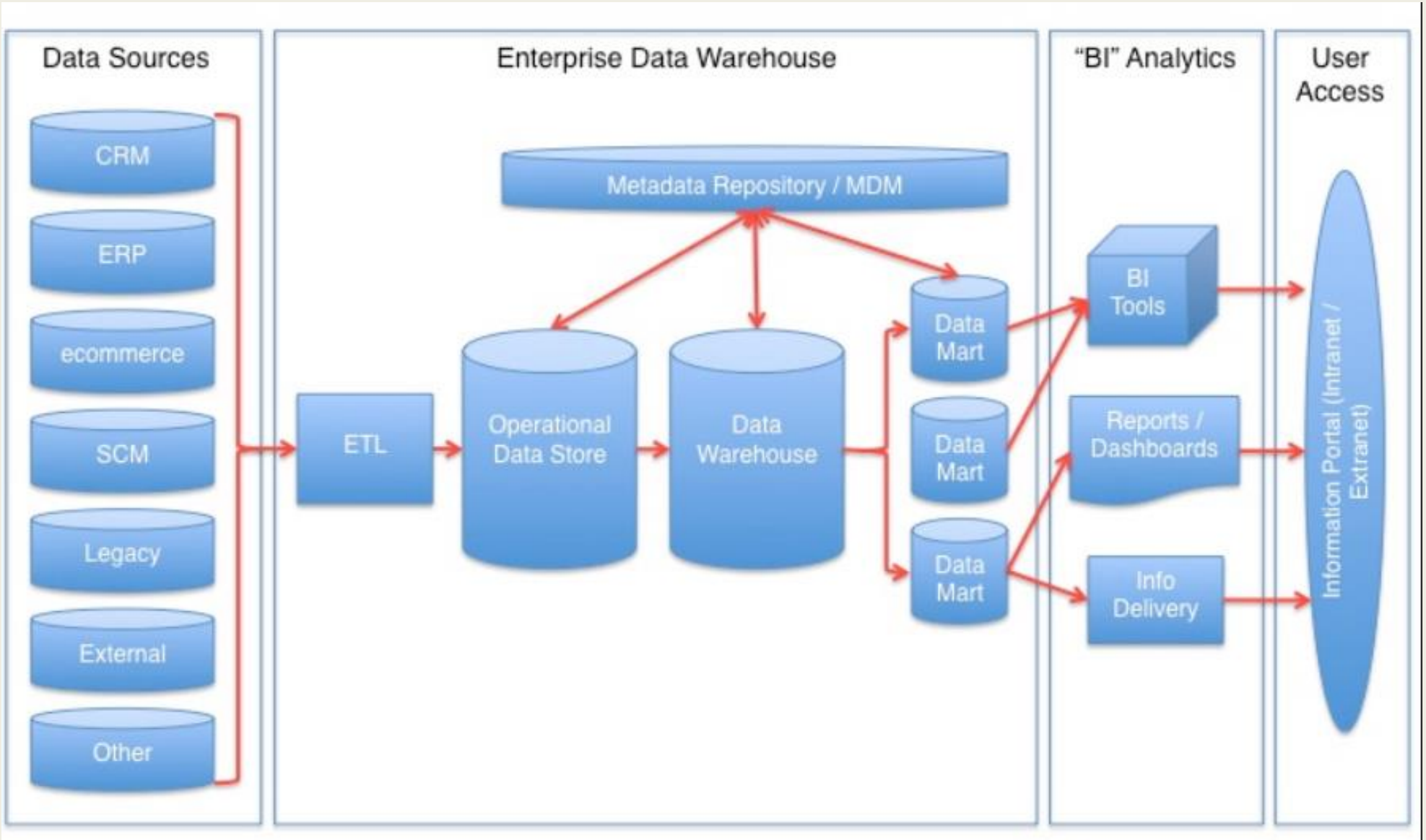
- Предсказуемия анализ служи за намиране и количествено определяне на скрити шаблони в данните, използвайки сложни математически модели, които могат да бъдат използвани за прогнозиране на бъдещите резултати. Предсказуемият анализ е различен от OLAP, тъй като OLAP се фокусира върху исторически анализ на данните и е реактивен по своя характер, докато предсказуемият анализ се фокусира върху бъдещето. Тези системи се използват и за управление на взаимоотношенията с клиенти (CRM).

Архитектура и архитектурни елементи

- Данните в склада данни се зареждат чрез процеса ETL /extraction, transformation, loading/, който включва извличане на данните от първичните източници, почистване и форматиране, проверка за дублиране, проверка за съответствие с ограниченията, зареждане в склада.

- Инструментите за аналитична онлайн обработка OLAP в общи линии са проектирани да работят с „денормализирани“ бази данни, въпреки че има инструменти, които работят със специални схеми за складове данни, съхранени в трета нормална форма (third normal form), т.е. нормализирани бази данни.
- Работата със складове данни често се нарича аналитична онлайн обработка (OLAP), за разлика от онлайн обработката на транзакции (OLTP), която се използва за обикновени бизнес дейности. Данните от системите за планиране на ресурсите на предприятието (Enterprise resource planning, ERP) и други свързани софтуерни бизнес системи, периодично се внасят в складовете данни за по-нататъшна обработка.

Склад за данни архитектура



Методи на проектиране

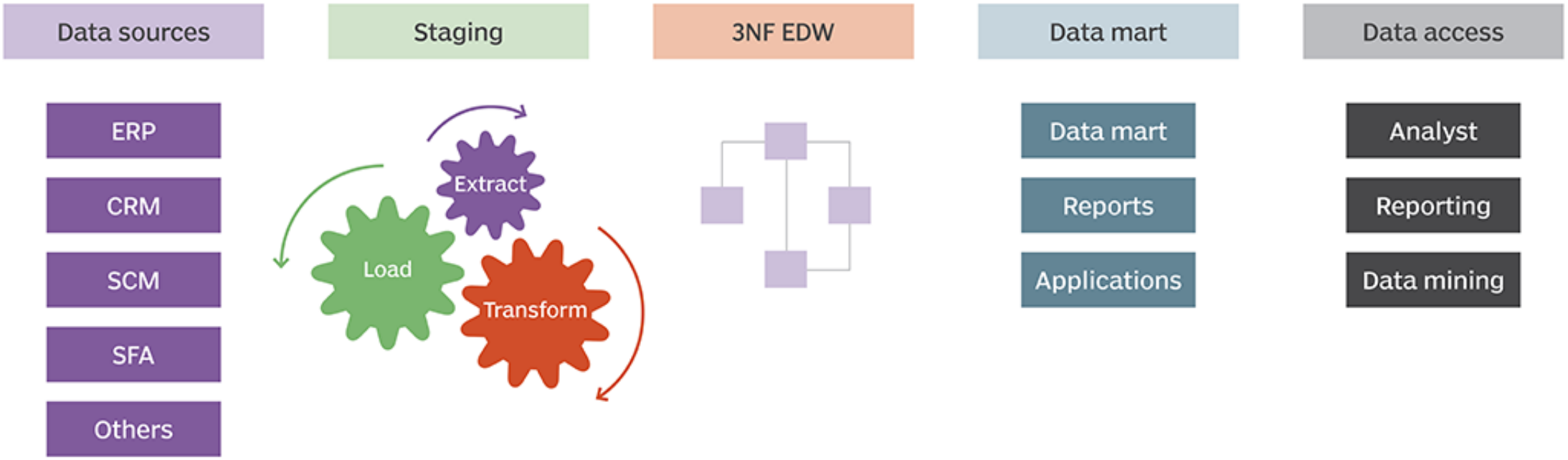
- Метод отгоре-надолу (Top-down method)
- Метод отдолу-нагоре (Bottom-up approach)
- Хибриден метод (Hybrid method)

Метод отгоре-надолу (Top-down method)

- Методът “отгоре-надолу” изисква първо да се изгради склад за данни. Данните се извличат от външни оперативни системи като могат да бъдат валидирани в зоната на създаване, преди да бъдат интегрирани в нормализиран модел за данни. Матриците данни се създават от данните, съхранявани в хранилището за данни.

Inmon's approach

In Bill Inmon's approach to data warehouse design—commonly known as top-down design—data is extracted from operational and third-party systems, transformed, and then loaded into the data warehouse.

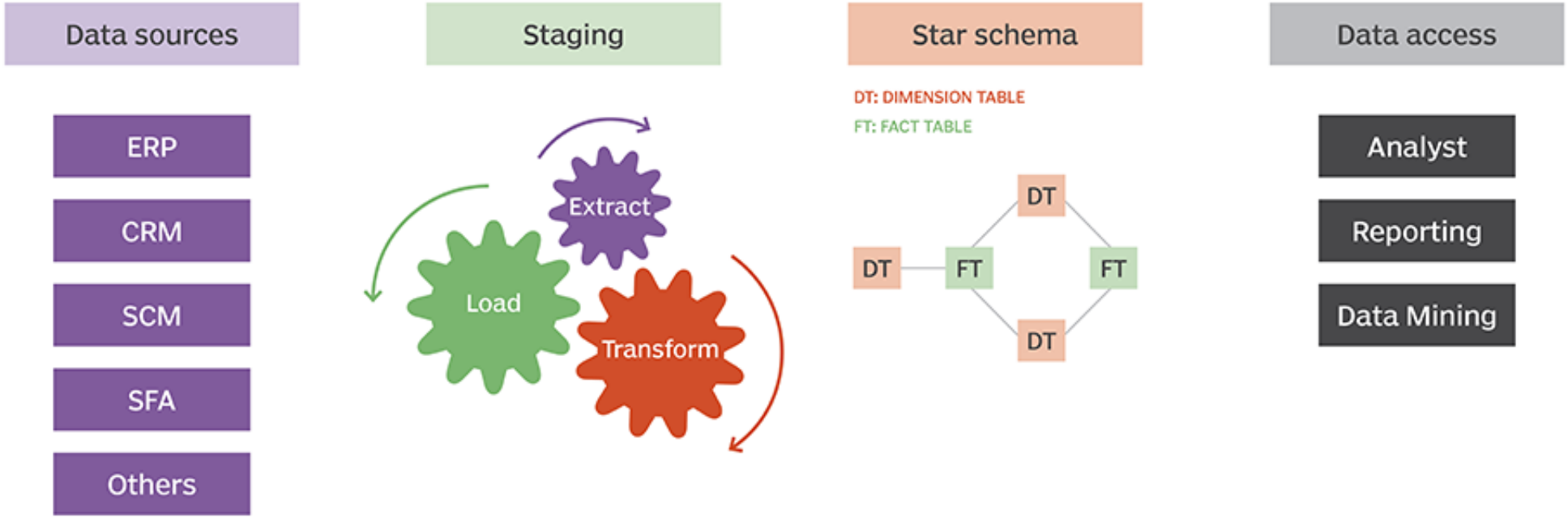


Метод отдолу-нагоре (Bottom-up approach)

- Архитектурата “отдолу-нагоре” за съхранение на данни изисква първо да се създадат матрици данни за измеренията. Данните се извличат от оперативни системи, преместват се в зоната на създаване и се моделират в схема звезда, с една или повече таблици с факти, свързани към една или повече таблици с измерения. Данните след това се обработват и зареждат в матрици данни, всяка от които се фокусира върху конкретен бизнес процес. Матриците данни се интегрират чрез архитектурата на складовата база данни, за да се създаде корпоративен склад за данни.

Kimball's approach

Ralph Kimball's approach to data warehouse design—referred to as bottom-up design—calls for data to be extracted from operational and third-party systems, transformed, and then loaded into data marts that are integrated into data warehouses.



Хибриден метод (Hybrid method)

- Хибридните подходи към дизайна на складове за данни включват аспекти от методите “отгоре-надолу” и “отдолу-нагоре”. Организациите често се стремят да комбинират скоростта на подхода “отдолу-нагоре” с интеграцията, постигната в дизайна “отгоре-надолу”.

ВЪПРОСИ ?