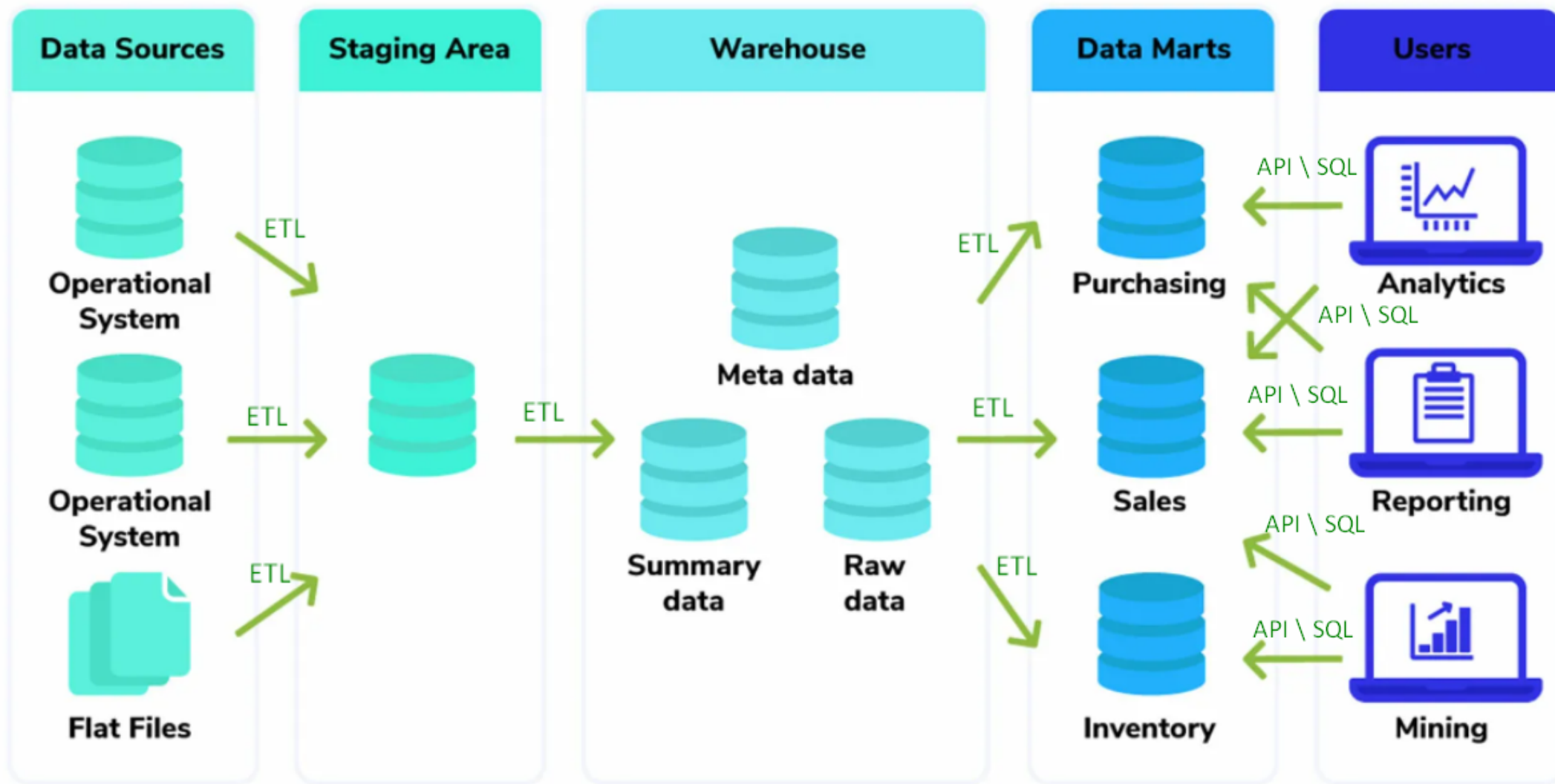


Проектирование Data Warehouse (DWH) - основы

Хранилище данных (Data Warehouse, DWH) - сущность в мире больших данных и бизнес-аналитики. Это может быть что угодно, начиная от компактной базы данных и заканчивая сложной многоуровневой корпоративной системой. В контексте нашего курса, DWH - это не просто база данных, а сложная система, включающая различные процессы и уровни для работы с большими объемами данных из разнообразных источников.

Уровни архитектуры DWH:

1. Источники данных: Это основа DWH, откуда происходит загрузка данных.
2. Staging-область: Здесь данные подвергаются предварительной обработке, чтобы ускорить их загрузку из источников и минимизировать нагрузку на основные системы.
3. Слой хранения структурированных данных: На этом этапе данные уже очищены и структурированы.
4. Витрины данных: Они предоставляют данные для конкретных аналитических нужд.
5. Презентационный уровень: Здесь происходит визуализация данных для отчетов и анализа.
6. ETL-уровень: Обработывает перемещение и преобразование данных между разными слоями.



ETL-процессы (Extract, Transform, Load) играют главную роль в DWH, так как они отвечают за извлечение данных из источников, их преобразование для соответствия стандартам и загрузку в хранилище. При проектировании DWH крайне важно правильно настроить эти процессы для обеспечения эффективности, надежности и консистентности данных.

Пример с библиотекой. Представьте DWH как большую библиотеку, где хранится огромное количество книг (данных) из разных источников. В этой библиотеке:

1. Источники данных: Это аналог книг, поступающих в библиотеку от различных авторов и издательств. Эти книги могут быть в разных форматах, языках и иметь различное содержание.
2. Staging-область: Здесь книги сначала проверяются и каталогизируются. Это подготовительный этап, где определяется, к какому разделу библиотеки они будут относиться. Это необходимо для того, чтобы упростить последующую обработку и хранение.
3. Слой хранения структурированных данных: Это основные полки библиотеки, где книги упорядочены, каталогизированы и доступны для читателей. Здесь данные уже структурированы и легко доступны.
4. Витрины данных: Это специальные стенды или разделы в библиотеке, где представлены книги на конкретные темы для удобства читателей. Аналогично, витрины в DWH предоставляют данные, специально подготовленные для конкретных запросов или отчетов.
5. Презентационный уровень: Это читальные залы, где посетители библиотеки могут просматривать книги, составлять отчеты или вести исследования.
6. ETL-уровень: Это процессы каталогизации, упорядочивания и подготовки книг. В библиотеке сотрудники выполняют эти задачи, так же, как ETL-процессы обрабатывают данные в DWH.