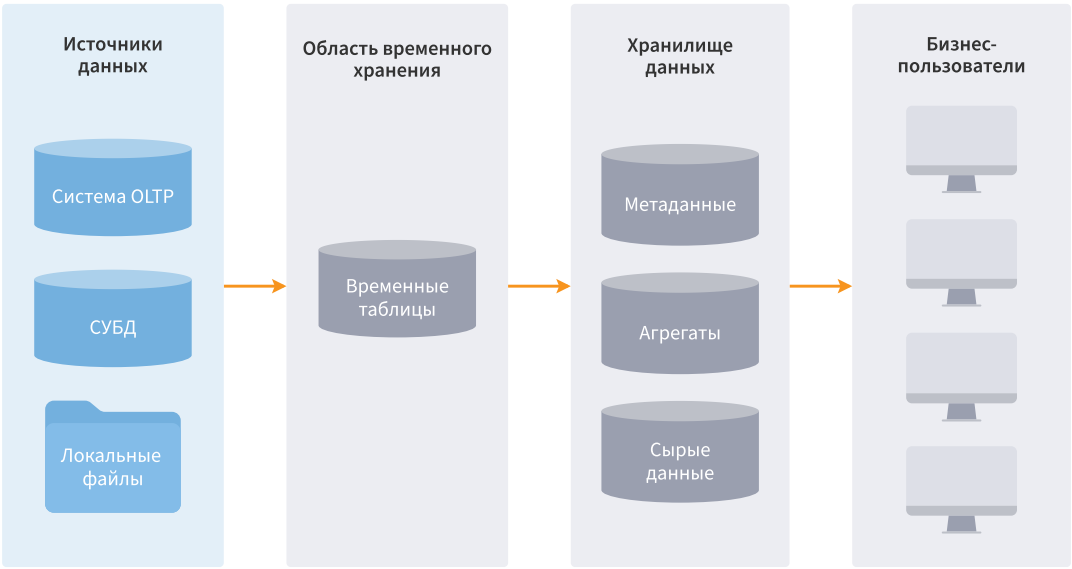
Область временного хранения (Staging area)

> Синонимы: Промежуточная область, Зона временного хранения, Landing zone
> Разделы: Источники данных

Область временного хранения — это технический уровень между источниками данных и хранилищем (ХД), формируемый в процессе ETL. Она содержит виртуальные таблицы, в которые временно загружаются сырые данные, извлеченные из источников. Здесь данные проходят предобработку перед загрузкой в физические структуры хранилища или витрины.



Как правило, данные из области временного хранения удаляются сразу после загрузки в структуры ХД. В этом случае такая область называется транзитной (transient staging area, TSA). Альтернативный подход — постоянная область хранения (persistent staging area, PSA), в которой сохраняется вся история изменений исходной таблицы, обычно для архивирования или восстановления данных после сбоев процесса ETL.

Области временного хранения реализуются в виде таблиц в реляционных структурах, текстовых или XML файлах. Они могут использоваться для различных целей, однако их основным назначением являются повышение эффективности процессов ETL, обеспечение целостности и качества данных.

Обычно области временного хранения поддерживают следующие функции:

- консолидация — объединение данных из множества источников с добавлением метаданных, указывающих на их происхождение, и временных меток, используемых для ведения истории в целевом ХД;
- выравнивание — стандартизацию поступающих из различных источников справочных данных, а также проверку взаимосвязей записей из разных исходных таблиц;
- минимизация рассогласований — снижение риска появления дублирующих или противоречивых записей за счет поточной, а не блочной выгрузки данных из источников.
- независимость потоков обработки — обеспечение возможности параллельной и независимой обработки данных, поступающих в разное время (например, из источников в разных часовых поясах) или предназначенных для различных целевых систем;
- захват изменений данных — обнаружение новых записей в источниках для предотвращения повторной загрузки одной и той же информации в ХД;
- очистка данных — обнаружение и исключение проблем в данных, снижающих их качество: пропусков, выбросов, дубликатов, противоречий и ошибок. Для этого используются бизнес-правила и технические ограничения;
- предварительное агрегирование данных — вычисление агрегатов в промежуточной области в соответствии с бизнес-правилами, что снижает нагрузку на целевое ХД;
- архивирование и устранение последствий сбоев — временное хранение данных после их загрузки в ХД для архивирования исторической информации или восстановления после технических сбоев в процессе ETL.

Следует отметить, что если инфраструктура ХД использует процесс ELT вместо ETL (когда предобработка данных производится в структурах хранилища), область промежуточного хранения не используется.