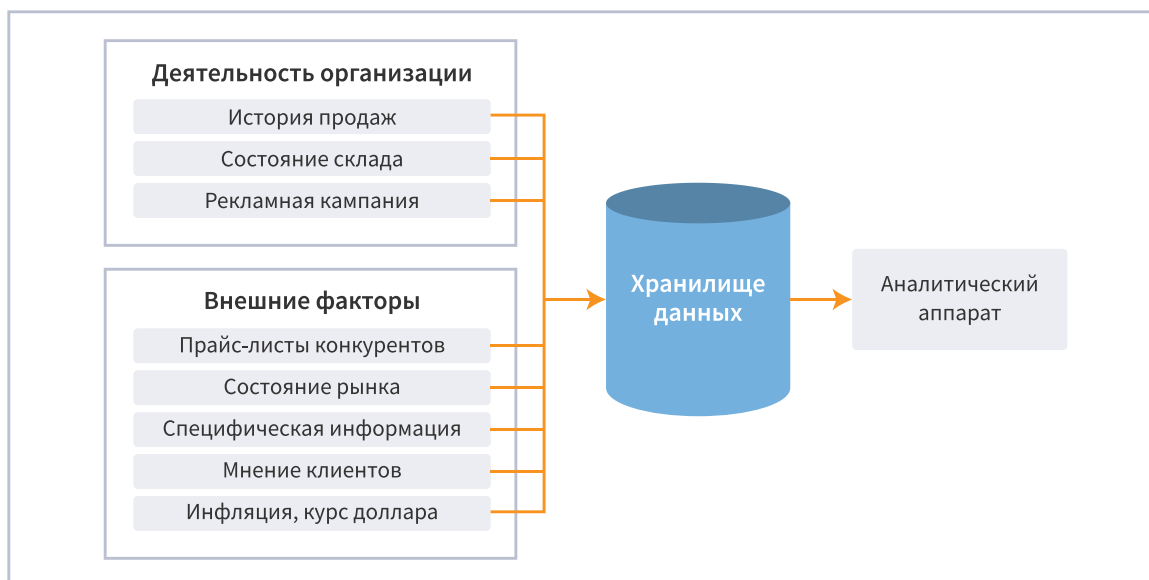


# Хранилище данных (Data Warehouse)

Разделы: [Источники данных](#)

Хранилище данных представляет собой предметно-ориентированный, интегрированный, неизменяемый и поддерживающий хронологию набор данных, организованный для целей поддержки принятия решений. Хранилища данных позволяют эффективнее, быстрее и качественнее предоставлять данные для систем их аналитической обработки, чем обычные СУБД.

- **Предметная ориентированность** означает, что данные в хранилище объединены в соответствии с областями, которые они описывают, а не с приложениями, которые их используют.
- **Интегрированность** означает, что хранилище должно поддерживать совместное хранение данных различной природы, форматов и типов, отражающих различные аспекты предметной области, а не отдельные бизнес-функции. Данные содержатся внутри хранилища в его едином внутреннем формате.
- **Неизменяемость** подразумевает, что для данных в хранилище предусмотрена только операция добавления, а удалять или изменять их нельзя. Если какие-либо изменения все же необходимы, «перегружается» все хранилище целиком. Необходимость такого подхода объясняется тем, что при промышленной эксплуатации хранилища совместно с аналитическими платформами один и тот же запрос к нему, выполняемый в любое время, должен обеспечить предоставление одних и тех же данных. Очевидно, что если бы в хранилище были разрешены изменения, то два одинаковых запроса, выполняемые с некоторым интервалом, в течение которого данные могли измениться, сформируют два различных набора данных, анализ которых может привести к некорректным заключениям и выводам, что недопустимо.
- **Поддержка хронологии** указывает на то, что хранение данных организовано с учетом даты и времени их появления, для чего каждой записи присваивается специальная метка времени (time stamp), что позволяет извлекать данные в хронологическом порядке и анализировать временные последовательности.



Хранилища могут использовать реляционную модель, когда данные в них нормализованы, или многомерную, использующую так называемые измерения. В нормализованных хранилищах данные содержатся в таблицах третьей нормальной формы. Преимущество нормализованных ХД заключается в простоте разработки и управления. Недостатком является необходимость денормализации данных «на лету» при их извлечении из множества таблиц при выполнении сложных аналитических запросов.

При формировании больших выборок это приводит к значительным задержкам в получении данных, а если хранилище и аналитическая платформа интегрированы в информационную систему предприятия, то возрастает нагрузка на всю систему, что может осложнить работу многих пользователей. Данную проблему частично удастся решить, используя в хранилище модель данных, основанную на измерениях. Применяются две разновидности многомерных моделей данных — «звезда» и «снежинка». Все загружаемые в хранилище данные обязательно должны быть определены как измерение, атрибут либо факт.

Кроме собственно данных, описывающих бизнес-процессы компании, в хранилище содержатся метаданные — служебные данные, описывающие структуру хранилища, содержащие информацию о принадлежности данных к тому или иному типу или виду (измерение, атрибут или факт). С помощью метаданных формируется семантический слой, который обеспечивает визуальные средства управления данными и метаданными. Метаданные в хранилищах разделяют на технические (обеспечивают работу самого хранилища), и бизнес-метаданные (описывают структуру данных в рамках заданной бизнес-модели).

В промышленной эксплуатации основными источниками данных для хранилищ являются OLTP-системы. Кроме этого, источниками быть любые файлы в информационной системе предприятия, где содержится структурированная информация, анализ которой, как ожидается, может дать полезные знания. Такие файлы могут иметь различные типы и форматы — электронные таблицы (Excel), настольные СУБД (Access), текст с разделителями (TXT, CSV-файлы), файлы учетных систем (1С:Предприятие, Парус) и т.д. Поэтому для хранилищ данных очень важно иметь развитые средства для загрузки и интегрирования данных из различных типов и форматов.

С хранилищем данных связаны еще два процесса: ETL — extract, transform, load (извлечение, преобразование, загрузка) и ELT — extract, load, transform (извлечение, загрузка, преобразование), реализующие перенос данных из разнородных источников в хранилище.

Автором концепции хранилищ данных в том виде, в каком она существует в настоящее время, считается Билл Инмон, который ввел данный термин в 1970-х. Большой вклад в развитие теории хранилищ данных и практики их использования в области бизнес-анализа и поддержки принятия решений внес Ральф Кимбалл, который также является автором многомерной модели данных.

В аналитической платформе Logiplot возможно импортировать и экспортировать данные из специально разработанного хранилища данных Deductor Warehouse.