**Лекция 4:**

**Модель типового проекта создания хранилища данных**

**Жизненный цикл разработки хранилища данных**

**Для чего создаются хранилища данных**

Главная цель создания ХД состоит в том, чтобы собрать вместе информацию из различных источников и представить эту информацию в формате, который является удобным для принятия решений по основным направлениям деятельности организации. Действия, необходимые для достижения поставленной цели, оказываются значительно более сложными, чем просто собрать данные и получать из них ответы на запланированные вопросы или получить ряд отчетов.

Подготовка данных для таких систем требует всесторонней экспертизы со стороны как технических специалистов, так и *специалистов предметной области*, что обычно включает в себя:

* точную идентификацию бизнес-информации, которая должна храниться в ХД;
* идентификацию предметных областей (бизнес-направлений деятельности организации) и определение их приоритетов, которые составят в итоге набор предметных областей ХД;
* управление границами каждой предметной области, которая будет представлена в ХД на интерактивной основе;
* разработку масштабируемой архитектуры для технического обслуживания ХД и его приложений, точное определение и обоснованный выбор структурных компонент (аппаратной, программной, лингвистической и организационной), которые будут реализовывать ХД;
* определение процедур экстрагирования, очистки, агрегации, преобразования и проверки данных для обеспечения гарантии их точности и согласованности;
* определение корректных уровней суммирования числовых показателей для поддержки бизнес-решений;
* определение и разработку процедур своевременного обновления данных, которые отвечают производственным потребностям, временным и жизненным циклам бизнес-информации;
* разработку дружественных интерфейсов и мощных инструментов аналитической обработки данных для конечных пользователей ХД;
* разработку словаря метаданных для обеспечения общения с ХД широких групп пользователей;
* обучение персонала для работы с приложениями ХД, всестороннее освещение круга решения задач, которые возможно решить в *системе складирования данных*;
* определение и регламентирование процесса сопровождения и обслуживания ХД и т.д.

Таким образом, ХД создаются для информационного обеспечения в решении стратегических задач по определенным направлениям бизнеса (для удовлетворения потребностей руководства и бизнес-аналитиков организации), а не для удовлетворения потребностей в автоматизации ежедневных процедур обработки бизнес-информации сотрудников структурных подразделений организации. В этом и состоит главное и коренное отличие систем складирования данных и систем бизнес-аналитики, использующих технологию ХД, от систем оперативной обработки информации (OLTP-систем, или On-Line Trasactions Proccessing).

**Факторы, влияющие на структуру проекта создания хранилища данных**

В современных экономических условиях стратегический бизнес-план организации разрабатывается на два-три года. Бизнес-процедуры, как правило, также пересматриваются в организации каждые два-четыре года. Это приводит к пересмотру функций и реинженирингу OLTP-систем организации. ХД создаются на пять и более лет, в них накапливается огромный объем информации. ХД предназначено жить, даже если половина бизнес направлений организации будет закрыта, а на их смену придут новые. Информация не теряется, она остается и будет доступна для анализа предыдущего опыта организации. Весьма расточительно планировать ХД так, чтобы через два года его переделывать.

Если язык общения системы бизнес-аналитики с конечными пользователями будет им не понятен, весь смысл и экономический эффект от разработки и внедрения такой системы будет сведен на ноль. Поэтому сбор и систематизация бизнес-терминологии, разработка открытой и гибкой структуры *лингвистического обеспечения*, ее внедрение в структуру ХД (а не в аналитические бизнес-приложения) является важной задачей проекта создания ХД.

Системы бизнес-аналитики создаются для лиц, принимающих решения, или приравненных к ним ведущих специалистов компаний. Эти системы обслуживают в первую очередь стратегические решения. Наивно будет полагать, что любой персонал способен принимать подобные решения для организации. Однако такая постановка вопроса не мешает пользоваться персоналу организации ХД как проверенным информационным источником накопленных компанией знаний.

Можно возразить, приводя примеры систем складирования данных в научных исследованиях. Ими пользуются многие участники научного сообщества — но в качестве источника информации, научного справочника. Однако можно смело утверждать, что научные ХД создаются для определенной группы ученых. Такая группа имеет научного руководителя, который и определяет направленность работы с информацией. В основном это получение новых знаний, разработка и апробирование эмпирических и математических моделей с целью получения новых научных знаний. В этом отношении научные ХД практически ничем не отличаются от ХД коммерческих организаций. Последним также нужны скрытые знания о своем бизнесе.

Таким образом, в проекте ХД вес разработки каждого структурного компонента приблизительно одинаков, ничто: ни аппаратные решения, ни программное, ни сетевое (включая midlleware), ни лингвистическое, ни технологическое обеспечение – не должно быть обойдено вниманием. Недооценка этого факта приводит к краху проекта создания ХД. Поэтому тщательное планирование и организационная структура проекта создания ХД так важна.

**Модель жизненного цикла хранилища данных**

В основу организационной структуры проекта создания или разработки ХД закладывается определенная **модель** (или *представление*) *жизненного цикла ХД* как продукта. Под **жизненным циклом продукта понимается набор определенным образом расположенных во времени фаз (этапов), которые проходит продукт от момента его создания до момента его утилизации**.

В зависимости от квалификации руководителя и масштаба проекта создания ХД используются *модели жизненного цикла* различной степени детализации. В настоящей лекции мы не будем углубляться в *анализ* различных методологических подходов к построению моделей *жизненного цикла ХД*. Все они основаны на обоснованном выборе того или иного способа разбиения процесса разработки ХД на этапы (фазы), идентификации задач каждого этапа и расположения выбранных этапов в пространстве и во времени. Отметим, что в настоящее время различные исследователи предлагают *разбиение* процесса разработки ХД на три, четыре, пять и более этапов. Отметим также, что все *модели жизненного цикла* разработки ХД имеют общие (типовые) элементы, которые в разных моделях могут быть отнесены к различным этапам.

Рассмотрим типовые элементы процесса создания ХД, которые и определяют его *жизненный цикл*. При этом предлагаемая модель не является единственно верной, так как сам процесс разработки не является строго структурированной задачей, она отражает лишь общий взгляд на этот процесс.

После принятия решения о создании ХД, до начала выполнения этапа планирования формируется проектная *команда*, определяются роли и задачи ее участников. Поэтому проектировщику важно представлять, что должны делать другие члены команды, а для этого нужно понимать свое *место* и роль в процессе разработки ХД.

Процесс создания и разработки ХД (жизненный *цикл разработки* хранилища данных) в общем случае можно представить состоящим из следующих основных стадий:

* планирование;
* формулирование требований к *системе складирования данных*;
* анализ;
* проектирование;
* конструирование;
* внедрение;
* поддержка.

Этап поддержки ХД в жизнеспособном состоянии включен для полноты представления жизненного *цикла* разработки ХД. Как правило, *поддержка* ХД осуществляется ИТ-службами заказчика в рамках отдельного проекта и имеет свой *жизненный цикл*. Отметим также, что в настоящее время внедрение также стало рассматриваться как самостоятельный проект — совместный (корпоративный) проект разработчика и заказчика ХД.

Охарактеризуем кратко задачи каждого из этих этапов.

**Планирование**

На **этапе планирования** ([рис. 3.1](https://intuit.ru/studies/courses/599/455/lecture/10159?page=1#image.3.1)) решаются следующие задачи: выбор стратегии реализации и методологии разработки, анализ задач, для решения которых создается ХД, анализ ресурсов разработки с технологической точки зрения и с точки зрения задач бизнеса, выбор архитектуры ХД, определяется бюджет проекта, разрабатываются возможные сценарии использования ХД, начинается сбор метаданных для ХД.



**Рис. 3.1.**Этап планирования разработки хранилища данных

Целью этого этапа является идентификация задач ХД, выбор способа решения этих задач, определения программно-технологического объекта и того, как, в какие сроки и за какие деньги этот объект будет реализоваться.

Выбор стратегии реализации определяет подход, который будет использован при создании ХД. Обычно используют следующие подходы: Top Down ("сверху вниз"), Bottom Up ("снизу вверх"), Middle In ("из середины") и комбинированный подход, который в последнее время становится все более популярным. Подход "сверху вниз" выбирается для вновь создаваемого ХД, т.е. когда "с нуля" принимаются все решения о технологической реализации объекта (аппаратура, программное обеспечение и т.д.). Подход "снизу вверх" используется, когда уже есть определенная вычислительная среда и объекты, из которых можно построить новый объект. Подход "из середины" предполагает эволюционное, поэтапное создание объекта, когда сначала разрабатывается так называемое ядро объекта, которое на следующих этапах наращивается новой функциональностью. Комбинированный подход применяет комбинацию выше перечисленных подходов.

Стратегия реализации определяет, какая концептуальная архитектура ХД будет использоваться и как будет строиться ХД с точки зрения последовательности реализации выбранной концептуальной архитектуры.

Выбор методологии создания ХД определяет, образно говоря, язык проекта, на котором будут разговаривать члены проектной команды, как будет оформлена техническая документация, какие принципы разработки будут использоваться. Это могут быть метод структурного анализа и проектирования (*SADT*), спиральный метод, методы, основанные на применении UML (язык *объектно-ориентированного моделирования*), и т.п.

Методология разработки основывается на использовании типичной для БД концепции трех схем и включает методы анализа спецификаций, методы концептуального, логического и физического проектирования.

Метод структурного анализа и проектирования является хорошо разработанной методикой и опирается на использование стандартов и методик *IDEF* или аналогичных им. Методика создания ХД по спирали реализует концепцию *эволюционного подхода* к созданию системы. Использование методов объектно-ориентированного анализа для реляционных ХД потребует преобразования полученной объектной схемы несущей базы данных в реляционную схему.

Анализ задач ХД предполагает идентификацию и определение объектов бизнеса, информация о которых будет содержаться в ХД. Вот далеко не полный перечень вопросов, на которые следует получить ответ на этой стадии:

* Что является предметной областью для хранилища данных?
* Какие программно-аппаратные платформы используются или какие планируется использовать?
* Какие возможности планируются в терминах свойств, характеристик и функций?
* Что представляют собой источники данных, которые можно или нужно интегрировать в хранилище данных?
* Когда хранилище данных должно начать функционировать?

Определение предметной (тематической) ориентации ХД является одной из самых главных задач на этом этапе. В рамках каждого фрагмента предметной области определяется:

* количество и типы источников данных (подразделения организации);
* число выбранных источников данных;
* данные, которые будут храниться в ХД;
* цели использования данных;
* каким должен быть спектр вопросов, на который должно отвечать ХД;
* каков размер *метамодели* хранилища данных;
* каков размер данных в хранилище данных;
* каковы источники входных данных и их количество;
* насколько востребованными являются данные из источников данных;
* насколько хорошо документированы данные из источников данных;
* каковы уровни управления организации (IFC);
* какие инструментальные средства доступны для логического проектирования;
* доступна ли модель данных в масштабе организации;
* есть ли профессионально подготовленный персонал;
* будет ли хранилище данных реализоваться в рамках существующей программно-аппаратной платформы или на платформе, аналогичной существующей.

На этой стадии планирования решается вопрос, что должно быть реализовано сначала, что можно реализовать в дальнейшем и, самое главное, *что не будет реализовано* в этом проекте создания хранилища данных. Задача состоит в том, чтобы четко ограничить предметную область хранилища данных и понять (переформулировать) спецификации для хранилища данных.

Выбор архитектурных решений должен задать такие ограничения:

* степень использования данных систем оперативной обработки;
* только хранилище данных;
* только *киоски данных*;
* хранилище и *киоски данных*;
* инфраструктура *секционирования* данных;
* использование архитектуры клиент-сервер (двухзвенная или *трехзвенная архитектура*).

Разработка программы проекта и бюджетного плана имеет целью составить план проекта, построить бюджет проекта, оценить стоимость и обеспечение методов для оценки окупаемости ХД в терминах:

* оценка стоимости восстановления и стоимости хранения;
* оценка возможности организации;
* оценка доходности организации;
* оценка расширения рынка;
* оценка конкурентного преимущества;
* оценка удовлетворения потребностей клиентов.

На этом этапе разработчики ХД должны работать в тесном взаимодействии с экспертами и экономистами.

Проработка сценариев типовой коммерческой деятельности для согласования с конечными пользователями является тонкой задачей, решение которой дает ответ на вопрос, как пользователи видят использование ХД. Решение задачи очень важно не только с точки зрения дальнейшей реализации проекта, но с точки зрения снижения риска неоправданных ожиданий от внедрения ХД.

Обычно анализ сценариев строится по следующей схеме. Каждый сценарий состоит из:

* четко определенного конечного пользователя с точно определенной его ролью в бизнес-процессах;
* тематической области, которая представляется в хранилище или *киоске данных* и будет использоваться данным конечным пользователем;
* одного или более вопросов (задач) по ключевым проблемам тематической области, которые воплощены неудовлетворительно в существующих ИС и полезны для решения следующих задач:
	+ построение адекватных критериев запросов для ХД;
	+ идентификация *метамодели* ХД;
	+ идентификация объема необходимой хронологической информации;
	+ идентификация точек зрения пользователя на данные (измерений);
	+ идентификация потребностей по данным киосков и ХД.

Цель настоящего этапа — очертить круг вопросов, на которые должна отвечать *система складирования данных*.

Цель следующей стадии, сбора метаданных, — сформировать спецификации по описанию данных и описать все необходимые элементы данных ХД в терминах бизнес-процессов и процедур бизнеса.

На этом этап планирования заканчивается, и команда проекта переходит к формированию каталога требований проекта. Без этого документа не имеет смысла продолжать проект, в намеченный срок он не закончится и в планируемый бюджет не вложится.

**Разработка требований**

**Этап разработки требований** к ХД ([рис. 3.2](https://intuit.ru/studies/courses/599/455/lecture/10159?page=2#image.3.2)) включает в себя следующие стадии:

* определение требований владельца ХД;
* определение требований конечных пользователей;
* определение технологических требований;
* определение *архитектурных требований*.



**Рис. 3.2.**Этап разработки требований к хранилищу данных

Требования владельца связаны с уточнением предметно-ориентированной направленности ХД, требованиями потенциального роста предметной области по:

* целям бизнеса;
* масштабу и целям хранилища данных/киосков данных, по клиентам и посредникам;
* требованиям клиентов;
* источникам данных;
* плану — бюджет, календарный план-график, ресурсы;
* влиянию на текущие инвестиции — люди, технология, обучение.

Часть требований бизнеса является спецификацией предметной области для ХД — спецификацией бизнес-функций направлений деятельности определенной категории потенциальных пользователей (идентификация интереса). Рассмотрим, например, отдел маркетинга компании. Отдел маркетинга может интересовать одна из следующих тем (направлений деятельности):

* маркетинговые исследования;
* анализ конкурентоспособности;
* анализ поведения покупателя;
* сегментация (рынка) реализации продукта;
* ценовые и бюджетные решения;
* решения по продукции;
* решения по продвижению;
* решения по каналам сбыта;
* прогнозирование тенденций;
* сертификация (Benchmarking).

Анализ выбранных выше направлений имеет дело со спецификацией предметных областей для отдела маркетинга, которыми в данном случае будут являться "Заказы", "Рынки", "Продажи", "Продвижение продукции", "Временные циклы".

Здесь важен такой параметр, как **степень детализации** (*granularity*) – чем ниже степень детализации, тем больше количество деталей. Это ключевая информация для проектировщика ХД, поскольку она определяет потенциальный объем ХД, скорость его роста, потенциальную производительность обработки запросов и т.д.

На этой стадии идентифицируется, в каких разрезах будут исследоваться данные потенциальными пользователями ХД. Для примера с отделом маркетинга это может быть:

* временной цикл с детализацией "День", "Неделя", "Месяц", "Квартал", "Год";
* группы клиентов с детализацией "Покупатели", "Сегмент рынка", "Рынок", "Отрасль промышленности";
* семейство продуктов с детализацией "Продукция", "Семейство продуктов", "Линия продуктов";
* география и месторасположение с детализацией "Район", "Регион страны", "Страна", "Международный регион";
* структура организации с детализацией "Подразделение", "Отдел", "Филиал", "Организация".

Здесь может быть также учтена специфика как самой организации, так и сектора промышленности и рынка, на которых работает организация.

На стадии определения требований по архитектуре ХД проводится планирование архитектуры в масштабе предприятия и по его структурным подразделениям, определяется архитектура данных в рамках принятой модели (например, ER-модели и т.д.), определяется архитектура приложений ХД, разрабатывается каталог приложений и функции каждого из них. На этой стадии полезно получить *CRUD*- (Create, Read, Update, Delete — создание, чтение, обновление, удаление) матрицу [[44](https://intuit.ru/studies/courses/599/455/literature#literature.44)-[49](https://intuit.ru/studies/courses/599/455/literature#literature.49)], чтобы на стадии проектирования физической структуры оценить частоту использования данных приложениями.

На стадии определения требований разработки создаются:

* спецификация приложений, интерфейсов, компьютеров, баз данных, коммуникаций и форм для конечного пользователя;
* технологические требования;
* требования размещения ПО;
* требования по поддержке ХД в программных продуктах;
* требования для разработчиков и обслуживающего персонала по их квалификации.

На стадии определения требования конечных пользователей оценивается роль ХД в делопроизводстве, соответствие функциональности ХД ежедневной последовательности операций конечных пользователей, формулируются критерии запросов к данным, требования по корпоративной отчетности, требования анализа данных.

Например, классифицируются в операциях над ХД действия пользователей, такие как:

* разделение элементов данных по различным аспектам анализа (операции Slice and *dice*);
* развертывание и экспозиция элементов данных по уровням иерархии (операции *Drill*-down up);
* поиск скрытых закономерностей в данных (алгоритмы Data Mining);
* просмотр данных непредусмотренным образом (Datasurfing);
* загрузка и выполнение модификаций данных.

В заключение этой стадии разрабатывается также набор бизнес-моделей использования ХД, после чего принимаются решения по просмотру данных. Например, принимается решение о том, как будут просматриваться различные категории данных:

* с помощью стандартной визуализации в рамках реляционных СУБД (двумерные таблицы);
* с помощью многомерных *кубов данных* (OLAP);
* с помощью предусмотренных отчетов и диаграмм;
* на основе жизненных циклов элементов данных.

Главным результатом этого этапа является создание каталога требований.

**Анализ**

Располагая каталогом требований, можно приступить к реализации **этапа анализа** —получения согласованных по источникам логической модели и определения набора инструментальных средств для работы с ХД. Стадии этого этапа показаны на [рис. 3.3](https://intuit.ru/studies/courses/599/455/lecture/10159?page=2#image.3.3).



**Рис. 3.3.**Этап анализа

Цель первой стадии этого этапа состоит в разработке логической модели данных для ХД и киосков данных. Кроме создания логической модели, необходимо фиксировать модели логической структуры баз данных подающих систем.

На следующей стадии определяются процессы, которые необходимы для связи источников данных и ХД, процедур очистки, преобразования и агрегации данных источников, производится выбор инструментальных средств для загрузки данных в ХД и выбор инструментальных средств конечного пользователя, которые должны быть размещены на клиентских компьютерах.

**Проектирование**

После завершения стадии анализа переходят к реализации **стадии проектирования** ([рис. 3.4](https://intuit.ru/studies/courses/599/455/lecture/10159?page=2#image.3.4)). Цель этого этапа — разработка физической модели ХД, проектирование процедур поступления данных в него и проектирование архитектуры приложений.

Проектирование ХД можно разложить на две равноценных стадии – проектирование архитектуры данных (логическое и физическое проектирование) и архитектуры приложений (анализ запросов и фиксация процессов взаимодействия хранилища данных с внешними источниками и пользователями).



**Рис. 3.4.**Этап проектирования хранилища данных

Детальное проектирование архитектуры данных включает в себя:

* **логическое проектирование**. Разработка *логических моделей данных* для ХД и киосков данных в рабочем пространстве базы данных. Отображение *логических моделей данных* источников данных в логические модели ХД и киосков данных;
* **физическое проектирование**. Отображение логических объектов в физическое описание ХД. *Денормализация* логической модели. Создание *табличного пространства*, *секционирование*, создание индексов и ограничений. Оценка объема физического ввода-вывода.

На стадии логического проектирования рассматриваются логические взаимоотношения между объектами предметной области. Особенностью логического проектирования ХД является тотальная ориентация на потребности и задачи конечного пользователя — бизнес-аналитиков и руководителей организации различных уровней. Они будут анализировать данные и работать с агрегированными данными, а не с данными конкретной бизнес-операции. Этот класс пользователей, как правило, обычно не знает точно, что им все-таки нужно от данных. И здесь важно оценить уровень претензий к данным и очертить круг задач, которые могут возникнуть.

В процессе логического проектирования выделяется набор объектов предметной области с их атрибутами. Объект представляет собой фрагмент информации о предметной области. В реляционных базах данных объект отображается в таблицу, а его атрибуты отображаются в колонки такой таблицы. В логическом проектировании ХД для создания *ER-диаграмм* используется *многомерное моделирование*, которое, грубо говоря, сводится к идентификации информации об объекте в виде фактов (таблицы фактов) и идентификации информации, с помощью которой на эти факты можно посмотреть (таблицы измерений). Результатом стадии логического проектирования является логическая схема ХД и, возможно, отображение логической *схемы источников* данных (подающих систем) на логическую схему ХД.

На стадии физического проектирования рассматриваются задачи размещения данных в БД для эффективной их выборки. На этой стадии (для реляционных хранилищ данных) реально пишутся SQL-операторы. Данные, собранные на стадии логического проектирования, превращаются в описание физической БД, включая таблицы, *уникальные идентификаторы объектов* – в первичные ключи, ограничения по значениям данных, взаимоотношения между объектами – во внешние *ключи, индексы*, *табличные пространства*, разбиения и представления.

Назначение *табличного пространства* состоит в отделении таблиц от их индексов, маленьких таблиц от больших таблиц, т.е. является механизмом для решения задачи оптимального размещения данных (некоторые СУБД решают эту задачу сами, без вмешательства пользователя).

*Секционирование* больших таблиц необходимо для увеличения производительности обработки запросов, т.е. является одним из механизмов решения задачи оптимизации выборки, так же, как и создание индексов.

Ограничения в ХД отличаются от ограничений в OLTP-системах. В *системах складирования данных* целостность и проверка данных обеспечивается на стадии загрузки данных. Поэтому роль ограничений в ХД не столь уж велика. Типичным ограничением в ХД является ограничение **NOT NULL**.

Параллельно с проектированием архитектуры данных начинается стадия детального проектирования структуры приложений, которая включает в себя определение:

* процессов, которые являются внутренними для источников данных и связаны с процедурами очистки и частичной экстракции информации;
* процессов, которые связывают источники данных с ХД и киосками данных;
* процессов, которые являются внутренними по отношению к ХД и предназначены исключительно для обслуживания данных в хранилище;
* процессов, которые связывают ХД и *киоски данных*;
* процессов, которые являются внутренними по отношению к *киоскам данных* и предназначены исключительно для обслуживания данных в них;
* процессов, которые связывают хранилище и *киоски данных* со средствами конечного пользователя;
* процессов, которые являются внутренними на рабочих станциях конечных пользователей и используются для установки связи с ХД и запуском средств анализа;
* процессов для поддержки управления и администрирования внутренних задач ХД как системы.

Целью этой стадии является получение спецификаций всех приложений ХД.

Все результаты стадии проектирования тщательно документируются, поскольку являются исходными и рабочими документами команды проекта и постоянно используются в ходе дальнейшей работы над проектом.

**Построение хранилища данных**

Следующий этап проекта создания ХД — это **построение ХД** ([рис. 3.5](https://intuit.ru/studies/courses/599/455/lecture/10159?page=2#image.3.5)).



**Рис. 3.5.**Построение хранилища данных

Цель этого этапа состоит в разработке программ и собственно физической БД под ХД. Выполнение этого этапа проекта, помимо создания собственно ХД, включает в себя разработку и отладку приложений ХД, а именно следующих групп программ:

* программы, которые создают и модифицируют БД для ХД и киосков данных;
* программы, которые экстрагируют данные из источников данных;
* программы, которые выполняют преобразования данных, такие как интеграция, суммирование и агрегация;
* программы, которые выполняют обновление реляционных БД;
* программы, которые реализуют поиск в очень больших БД.

Результатом выполнения этого этапа является комплекс программ, работающих с ХД.

**Внедрение**

Следующим этапом реализации проекта создание является **внедрение** в опытную эксплуатацию ([рис. 3.6](https://intuit.ru/studies/courses/599/455/lecture/10159?page=2#image.3.6)). Это очень ответственный и трудоемкий этап. Начинается он со стадии начальной инсталляции, включающей начальную загрузку хранилища из источников данных, и тестирования процедур обновления и синхронизации данных.



**Рис. 3.6.**Этап внедрения хранилища данных

Далее составляются и утверждаются планы постадийного тестирования, тренировки и обучения персонала для всех классов пользователей, реализации обновления несущих платформ для ХД, информационных потребностей и т. д.

Выполняется обеспечение администрирования системы и данных о пользователях, возможностей архивирования и резервного копирования, возможностей восстановления, управления доступом и безопасностью, полных возможностей и процесса управления системой и ее структурными компонентами, информационного каталога/директории. Проводится целостная интеграция хранилища данных в существующую инфраструктуру организации.

Результатом выполнения этого этапа является всесторонняя *подготовка перехода* ХД в промышленную эксплуатацию.

**Поддержка**

Этап поддержки ХД в работоспособном состоянии, как уже отмечалось выше, является самостоятельным проектом. Это последний этап *жизненного цикла ХД*. По его завершении происходит либо уничтожение ХД как продукта, либо его реинжениринг. Это связано с быстрыми изменениями сферы бизнеса организации, которые приводят к необходимости пересмотра стратегических и тактических планов организации. ХД, так же, как и другие элементы инфраструктуры организации, должно соответствовать генеральному бизнес-плану организации. В зависимости от структуры генерального бизнес-плана организации, эксплуатирующей ХД, продолжительность этого этапа может составлять два-четыре года, реже пять и более лет.

На этом этапе ИТ-служба организации обеспечивает:

* поддержку работоспособности и масштабируемости программно-аппаратного обеспечения ХД;
* сбор, очистку, преобразование, загрузку и актуализацию данных в соответствии с установленными бизнес-процедурами;
* поддержку автоматизированных мест пользователей.

Этот этап может также включать в себя техническую поддержку со стороны разработчика ХД, в частности консультирование, обучение пользователей или абонентское обслуживание.

Этот этап является также испытанием ХД "на прочность". Конечные пользователи, как правило, не являются ИТ-профессионалами. Они решают свои задачи. Они будут использовать ХД только в том случае, если оно будет эффективно помогать им решать их задачи. Поэтому ИТ-служба организации должна постоянно отслеживать активность пользователей на ХД и вести постоянный аудит мнений пользователей о полезности ХД в их профессиональной деятельности, отсекая при этом "зерна от плевел".

**Временные затраты на реализацию этапов**

Таким образом, учитывая замечание о поддержке ХД как о самостоятельном проекте, в жизненном цикле создания ХД можно выделить шесть основных этапов. Независимо от того, как заказчик ХД понимает процесс его создания, эти этапы должны быть пройдены и заложены в бюджет проекта. Они обычно выполняются в последовательном порядке, хотя некоторые из них могут разрабатываться параллельно.

Выше была рассмотрена классическая схема организации проекта разработки программного обеспечения в приложении к ХД. Руководитель проекта может придерживаться иной схемы разбиения на этапы. Например, если перед ним ставится задача *только* создать ХД, то можно использовать разбиение жизненного цикла создания ХД, как в [табл. 3.1](https://intuit.ru/studies/courses/599/455/lecture/10159?page=2#table.3.1).

В данном случае разбиение также включает шесть основных этапов. Этап планирования сформулирован более узко, а этап анализа сведен к выбору технологии складирования данных. Из этапа проектирования ХД исключено проектирование приложений. Этап конструирования ХД представлен в виде трех самостоятельных этапов. Этап внедрения рассматривается вне проекта создания ХД.

В [табл. 3.1](https://intuit.ru/studies/courses/599/455/lecture/10159?page=2#table.3.1) дана также приблизительная оценка временных затрат на каждый этап для типового проекта создания ХД *средней сложности*.

|  |
| --- |
| Таблица 3.1. Временные затраты на реализацию этапов проекта создания ХД |
| **Наименование этапа** | **Время** | **Результат** |
| Документирование подающих систем и источников данных | 1-2 мес. | Каталог требований |
| Выбор технологии складирования данных | 0,5 мес. | Архитектура ХД |
| Проектирование модели хранилища данных | 1 мес. | Логическая модельФизическая модельСкрипт |
| Разработка программ извлечения и *очистки данных* | 2 мес. | Инструментальные cредства *ETL* (extraction, transformation, loading) |
| Построение репозитория метаданных | 4 мес. | Тезаурус бизнес-терминовКаталог бизнес-процедур обработки данных |
| Создание многомерных отчетов и *кубов данных* | 2 мес. | Приложения ХД |

**Бизнес-модель типового проекта создания хранилища данных**

**Типовая модель бизнес-процессов разработки хранилища данных**

В предыдущем разделе настоящей лекции была изложена абстрактная модель жизненного цикла разработки ХД. Ниже будет рассмотрена типовая модель бизнес-процессов разработки ХД, которая может быть положена в основу реализации такого проекта. Эта модель содержит минимально достаточное число обязательных этапов для реализации небольшого или среднего по масштабу проекта. Описание бизнес-модели приведено в нотации *IDEF0*.

Проект разработки ХД начинается после того, как выбраны инструменты разработки и сформирована команда проекта. Как видно из [рис. 3.7](https://intuit.ru/studies/courses/599/455/lecture/10159?page=3#image.3.7), в проектный цикл разработки ХД обычно включаются следующие *типовые процессы* (этапы):

* формулирование требований;
* создание вычислительной среды;
* моделирование данных;
* определение процедур извлечения преобразования и загрузки данных;
* проектирование аналитических отчетов;
* разработка приложений ХД;
* настройка производительности;
* проверка качества;
* передача *системы складирования данных* в эксплуатацию.



**Рис. 3.7.**Укрупненная бизнес-модель (IDEF0) процесса разработки хранилища данных

Опишем каждый типовой этап разработки ХД по следующей схеме.

* **Описание задачи**. Что обычно должно быть достигнуто в течение данного этапа разработки ХД.
* **Временные требования**. Приблизительная оценка количества времени для решения задачи данного этапа.
* **Результат выполнения этапа**. В конце каждого этапа оформляются один или более документов, которые описывают шаги и результаты решения данной задачи.
* **Потенциальные опасности**.

Теперь перейдем к рассмотрению отдельных компонентов представленной бизнес-модели.

**Формулирование требований**

**Задача этапа**. Главной задачей данного этапа является идентификация требований заказчика ХД и оформление их в виде документа "Каталог требований". Обычно сбор требований осуществляется путем опроса групп потенциальных пользователей ХД на специальных совещаниях и переговорах. Конечные пользователи, как правило, не знакомы с концепцией ХД и процессом складирования данных. Поэтому для успешного решения этой задачи важна помощь лиц, принимающих решения (ЛПР), т.е. руководителей организации.

На этом этапе определяются:

* масштаб проекта создания ХД (границы предметной области);
* перечень и содержание отчетов;
* требования по анализу данных (перечень задач анализа данных);
* требования к аппаратному обеспечению;
* требования к системному программному обеспечению;
* значения базовых параметров *системы складирования данных* (скорость обработки запросов, объемы используемых данных, актуализация данных, производительность системы и т.д.);
* требования к квалификации персонала (программа обучения персонала);
* перечень источников данных;
* конкретизация плана реализации проекта разработки ХД (определяется дата завершения проекта).

В дополнение, основываясь на собранной выше информации, составляется план восстановления *системы складирования данных* в случае аварийных сбоев. Разрабатывается стратегия архивирования и восстановления ХД.

**Временные требования**. Время выполнения этапа — от двух недель до двух месяцев.

**Результатом выполнения этапа** являются каталог требований, утвержденный заказчиком, и уточненный план проекта, который точно определяет используемые ресурсы и даты контрольных точек проверки хода выполнения проекта.

**Потенциальные опасности**. Этап формулирования требований часто оказывается одним из самых узких мест проекта создания ХД. Причина состоит в конфликте внутрикорпоративных интересов и в необходимости наладить коммуникации для успешного выполнения и этапа и проекта в целом.

По определению складирование данных предполагает интеграцию в ХД данных из нескольких источников (подразделений организации). Позиция и взгляды подразделений на бизнес-информацию (производство данных, их верификацию, поставку данных в ХД, разграничение доступа к данным и т.д.) могут сильно расходиться и даже быть диаметрально противоположными. Даже если команда разработчиков предложит блестящее решение по созданию ХД, его реализация может споткнуться о нежелание определенных групп потенциальных пользователей предоставлять данные в ХД или конструктивно участвовать в определении требований к *системе складирования данных*.

Если не удается наладить коммуникации между участниками процесса, то все усилия по созданию ХД будут потрачены впустую и проект никогда не будет завершен в установленные сроки, а в худшем случае просто будет провален.

Одним из способов избежать такой ловушки является непосредственное вовлечение руководителей организации в процесс реализации проекта. Следует заручиться поддержкой влиятельного покровителя из числа высшего руководства, чтобы можно было влиять на позицию подразделений сотрудничать с командой разработчиков ХД.

**Создание вычислительной среды для хранилища данных**

**Задача этапа**. Главной задачей этого этапа является создание информационно-вычислительной среды, в которой будет разрабатываться ХД. Это достаточно важный этап. Совершенно ясно, что создание ХД организации на двух компьютерах разработчиков — затея весьма опасная и нереальная. Вычислительная среда разработки ХД данных должна достаточно точно моделировать ту вычислительную среду, в которой реально будет эксплуатироваться ХД, т.е. она должна быть масштабируемой по аппаратному решению.

После того как требования к программно-аппаратной части определены, нужно установить серверы приложений, серверы БД, клиентские рабочие станции и автоматизированные места разработчиков.

На этом этапе должна быть достигнута полная ясность в том, как будут выполняться технологические работы по созданию ХД. Из-за особенностей работы конкретной организации возможно применение нескольких вычислительных сред. Например, для кредитных организаций, которые работают в режиме реального времени, может быть использовано три различных вычислительных среды: программно-аппаратный комплекс разработчиков, программно-аппаратный комплекс для тестирования и вычислительная среда организации, в которой будет эксплуатироваться ХД.

Причина, по которой лучше применять отдельную вычислительную среду для разработки, состоит в том, что внесение изменений и тестирование разрабатываемого ХД могут привести к аварийным сбоям в существующей вычислительной среде организации, что повлечет дополнительные эксплуатационные издержки.

**Временные требования**.

* Закупка и установка серверов — до двух недель.
* Монтажные работы на установке компьютерной сети — две-четыре недели.

**Результатом выполнения этапа** являются спецификации на программно-аппаратное обеспечение, а также скрипты и установки для программного обеспечения.

**Потенциальные опасности**. Самой большой опасностью является использование одного сервера БД для моделирования различных вычислительных сред, например, вычислительной среды разработки и вычислительной среды тестирования, или, что еще хуже, для вычислительной среды разработки и вычислительной среды эксплуатации ХД, особенно если на этом сервере работает уже существующая информационная система.

В этом случае могут возникнуть конфликты между различными участниками проекта создания ХД и ИТ-службой организации в случае непредусмотренной остановки сервера БД.

Кроме того, в тестовой среде будет невозможно промоделировать ожидаемые показатели нагрузки на систему и оценить ее производительность.

**Моделирование данных**

**Задача этапа**. Главной задачей этапа является разработка логической и *физической моделей данных* для ХД. Этот этап — один из самых важных для проекта создания ХД. Ошибки и просчеты, допущенные на этом этапе, будет очень сложно исправить на последующих этапах. Кроме того, подобные просчеты могут привести к пересмотру всего проекта и, следовательно, к его фактическому провалу.

При выполнении этого этапа сначала строится *логическая модель данных*, которая впоследствии преобразуется в физическую модель.

Одной из подзадач этого этапа являются идентификация и описание источников данных, которые также могут стать подзадачей этапа определения процедур извлечения, преобразования и загрузки данных в ХД (*ETL*-процессов).

**Временные требования**. Время выполнения этого этапа — от двух недель до двух месяцев.

**Результатом выполнения этапа** являются перечень источников данных и их описание, а также логическая и физическая модели данных.

**Потенциальные опасности**. Самой большой опасностью при выполнении этого этапа является самоуверенность проектировщиков ХД. Во многих случаях даже опытные проектировщики допускают от двух до пяти ошибок в структуре данных (потерь существенных семантических зависимостей в данных) на проект. Причиной таких ошибок часто становится недостаточная осведомленность проектировщиков о предметной области ХД и низкое качество информации, поставляемой аналитиками предметной области.

Бизнес-аналитики предметной области могут не предоставить полной информации о функциональных зависимостях в данных, что приведет в результате к проектированию частично неправильной схемы данных. В этом случае схема правится "на лету" на последующих этапах выполнения проекта, что может привести к пересмотру архитектуры приложений и процессов *ETL*. Иногда ошибки носят принципиальный характер и приводят к полному пересмотру фрагмента схемы данных. Например, как в случае, когда пропущена зависимость вложения на подмножествах в данных (типа "часть-целое").

Хорошей предупредительной мерой для предотвращения подобных ситуаций является привлечение квалифицированных экспертов, особенно в случае если проект разработки ХД выполняется силами самой организации.

**Определение процедур извлечения, преобразования и загрузки данных**

**Задача этапа**. Главной задачей этапа является идентификация и определение процедур извлечения, очистки (фильтрации), преобразования и загрузки данных. Это весьма трудоемкий этап, не столько по временным затратам, сколько по усилиям, предпринимаемым по анализу структур данных источников и подающих систем.

Исключительно редким является случай, когда ХД данных создается на голом месте, т.е. в подразделениях отсутствуют автоматизированные подсистемы обработки данных. Как правило, данные уже существуют (в том или ином виде). Их нужно собрать, согласовать, привести к единому формату, агрегировать и загрузить в ХД. По этой причине этот этап является характерным для проекта создания ХД.

Также следует помнить о том, что сам процесс подготовки и загрузки данных в создаваемое ХД может занять более половины времени, отведенного на реализацию проекта, особенно в проектах большого масштаба.

**Временные требования**. Время выполнения этапа — от одной недели до полутора месяцев.

**Результатом выполнения этапа** являются схема соответствия данных подающих систем и ХД, программы или *ETL*-инструменты.

**Потенциальные опасности**. Первой потенциальной опасностью при выполнении этого этапа является недооценка временных параметров. Обычно на выполнение этапа выделяют немного времени. Процедуры извлечения, преобразования и загрузки данных в ХД оказывают непосредственное влияние на качество и полноту предоставляемой информации конечным пользователям. Недостаточное внимание к разработке этих процедур может вызвать негативные реакции у пользователей после получения ими отчетов.

Второй потенциальной опасностью является стремление команды разработчиков сделать процесс *ETL* как можно более всеобъемлющим, мотивируя свои действия стремлением обеспечить качество данных. Следует помнить, что главная цель *ETL*-процесса — оптимизации скорости загрузки данных в ХД.

**Проектирование аналитических отчетов**

**Задача этапа**. Главной задачей выполнения этого этапа является проектирование и разработка аналитических отчетов на спроектированной структуре данных. Это также специфичный этап для проекта ХД.

Перечень требуемых отчетов содержится в "Каталоге требований", и их разработка решает в целом поставленную задачу. Однако следует помнить, что потенциальные пользователи не всегда точно знают, что они хотят увидеть в отчетах. С другой стороны, как правило, собранные данные предоставляют большие возможности по формированию разнообразных отчетов, чем это зафиксировано в "Каталоге требований". Здесь должен быть найден разумный компромисс.

**Временные требования**. Время выполнения этого этапа зависит от числа разрабатываемых отчетов. В зависимости от сложности отчета его разработка занимает от 4 часов до двух недель.

**Результатом выполнения этапа** станут спецификация *кубов данных* (измерения и метрики) и разработанные отчеты.

**Потенциальные опасности**. Потенциальной опасностью при выполнении этого этапа является то, что не уделяется достаточного внимания оптимизации времени получения отчета. Конечный пользователь не любит долго ждать. А если он получал такой отчет на своей настольной системе быстрее, то мнение о ХД у него будет, мягко говоря, неадекватное.

Хороший способ избежать такой опасности — тестирование разработанных отчетов с целью минимизации времени их получения.

**Разработка приложений хранилища данных**

**Задача этапа**. Главной задачей данного этапа является формирование программной среды, в которой пользователи будут извлекать данные из ХД и просматривать предопределенные отчеты.

Каковы бы ни были инструментальные средства OLAP, необходимо продумать, как будут происходить визуализация отчетов и их доставка конечному пользователю.

**Временные требования**. Срок выполнения этого этапа — от одной недели до месяца. При использовании тонкого клиента типа браузера следует исходить из того, что в среднем на разработку и отладку одной веб-страницы тратится полдня.

**Результатом выполнения этапа** будет документация, описывающая механизм доставки пользователям отчетов и спецификации экранных форм.

**Потенциальные опасности**. Самой большой потенциальной опасностью является ложное представление о достаточной квалификации пользователей ХД для работы с ИТ- технологиями. Конечные пользователи хотят быстро получать данные, необходимые для решения своих конкретных задач, а не изучать многотомные инструкции по работе с программным обеспечением.

Если интерфейс интуитивно не понятен или не содержит ясных и кратких инструкций по работе с электронной формой (встроенных в форму Help), он не будет использовать систему складирования данных с должной эффективностью, что приведет к постепенному ее вымиранию.

**Настройка производительности**

**Задача этапа**. Главная задача выполнения этого этапа — добиться оптимальной производительности *ETL*-процессов, производства отчетов и их доставки конечному пользователю.

Извлечение, преобразование и загрузка данных, как правило, является длительным процессом, который выполняется в пакетном режиме с не очень высоким приоритетом. Пока данные не загружены в ХД, новые отчеты не могут быть получены.

При обработке некоторых запросов захватывается большое количество данных, на которых выполняются динамические операции агрегирования и суммирования, что приводит к значительному времени производства отчета.

На доставку отчетов пользователю может влиять загрузка локальной вычислительной сети и большой объем сетевого трафика.

**Временные требования**. Время выполнения этого этапа не должно превышать одну-две недели.

**Результатом выполнения этапа** является перечень рекомендаций по настройке производительности.

**Потенциальные опасности**. Потенциальной опасностью этого этапа считается использование вычислительной среды разработки ХД, которая не масштабируется к вычислительной среде эксплуатации ХД. Из-за этого рекомендации по настройке производительности не будут соответствовать реальным условиям эксплуатации ХД.

**Проверка качества**

**Задача этапа**. Главной задачей этапа — убедиться, что ХД готово к эксплуатации. Как правило, проверка качества выполняется отдельной группой специалистов, не входящих в состав команды разработчиков.

**Временные требования**. Срок выполнения этого этапа — от одной до четырех недель.

**Результатом выполнения этапа** являются план тестирования ХД и заключение о готовности ХД к эксплуатации.

**Потенциальные опасности**. Самый большой риск любого проекта — это люди: их квалификация, амбиции, заинтересованность в деле, мотивы и т.д. Поскольку люди, проверяющие ХД на этом этапе, не входят в проектную команду, возникает потенциальная опасность, связанная с недостаточной образованностью этих людей в области складирования данных.

Разумным решением при выполнении этого этапа является привлечение организацией сторонних специалистов высокой квалификации по проверке качества ХД.

**Передача системы складирования данных в эксплуатацию**

**Задача этапа**. Главная задача этапа — передача *системы складирования данных* заказчику и представление ее конечным пользователям.

**Временные требования**. Срок выполнения этого этапа — от одного дня до нескольких недель.

**Результатом выполнения этапа** является акт о приемке-сдаче программного продукта.

**Потенциальные опасности**. Самая большая потенциальная опасность для этого этапа — неготовность потенциальных пользователей к работе с ХД. Хорошим правилом здесь будет проведение нескольких презентаций и коротких обучающих курсов по работе с ХД.

**Сопровождение и модификация хранилища данных**

Обычно в проектный цикл включают еще два этапа — этап сопровождения ХД и этап его модификации. Это важные этапы в жизни ХД. Однако, как показывает опыт, целесообразно выделять их в самостоятельные проекты.

Было бы большой ошибкой поручать разработчикам ХД его сопровождение. Процессы сопровождения ХД требуют от ИТ-специалистов иной квалификации, чем процессы его разработки. Чтобы сопровождать, не обязательно уметь писать программы, но обязательно нужно уметь их настраивать и использовать.

Если необходимость в модернизации ХД возникает спустя несколько месяцев после сдачи его в эксплуатацию, это говорит о том, что проект не был успешным. Потребность в модернизации реально может сформироваться спустя шесть месяцев после интенсивной его эксплуатации, когда проверены его возможности, прочувствована отдача и видны новые перспективы использования, т.е. когда ХД стало частью производственного процесса. При этом не факт, что та же команда разработчиков будет заниматься его модернизацией.

**Резюме**

Процесс разработки ХД может быть представлен в виде модели бизнес-процессов. Обычно руководители ИТ-проектов не создают *бизнес-модель* процесса разработки ХД. А напрасно! *Бизнес-модель* процесса разработки позволяет:

* отобразить субъективное мнение руководителя ИТ и некоторых участников команды на процесс проектирования конкретного ХД;
* учесть особенности ИТ-проекта, в рамках которого проектируется ХД;
* достаточно быстро составить план проектирования конкретного ХД;
* просчитать длительность проектных работ (создать временную модель проектирования).

Представленная в настоящей лекции укрупненная *бизнес-модель* процесса разработки ХД является достаточно общей. В нее не включены многие детали и риски. *Конкретизация* и *детализация* процесса разработки ХД — это основная задача руководителя проекта разработки ХД. Проектировщик ХД должен понимать, какие документы он вправе ожидать от членов команды, какие документы должен создать и кому разработанные документы должен передать.