МИНИСТЕРСТВО ОБРАЗОВАНИЯ АЗЕРБАЙДЖАНСКОЙ РЕСПУБЛИКИ

АЗЕРБАЙДЖАНСКИЙ ГОСУДАРСТВЕННЫЙ ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ (UNEC)

ЦЕНТР МАГИСТРАТУРЫ

Мамедов Эльнур Сабахаддин оглу

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

на тему

“ЗАДАЧИ ОПЕРАТИВНОГО СИТУАЦИОННОГО АНАЛИЗА, РЕШАЕМЫЕ НА ОСНОВЕ ИНФОРМАЦИОННЫХ ХРАНИЛИЩ”

Шифр:	**060509 – “Компьютерные науки”**
Специалность:	**“Управление информационными технологиями”**
Научный руководитель:	**Руководитель магистерской программы:**
доц. А.Х.АБДУЛЛАЕВ	акад. А.М.АББАСОВ
Заведующий кафедры:	акад. А.М.АББАСОВ

БАКУ – 2020

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы. Облачные технологии произвели революцию в мире бизнеса, позволив компаниям легко получать и хранить ценные данные о своих клиентах, продуктах и сотрудниках. Эти данные используются для информирования важных деловых решений.

Многие глобальные корпорации обратились к хранилищам данных для организации потоковой передачи данных из корпоративных филиалов и операционных центров по всему миру. Для ИТ специалистов важно понимать, как хранилище данных помогает компаниям оставаться конкурентоспособными на быстро развивающемся глобальном рынке.

Хранилище данных становится все более важным инструментом бизнес-аналитики, позволяющим организациям обеспечить последовательность. Хранилища данных запрограммированы на применение единого формата ко всем собранным данным, что позволяет корпоративным лицам, принимающим решения, анализировать данные и делиться ими со своими коллегами по всему миру. Стандартизация данных из разных источников также снижает риск ошибок при интерпретации и повышает общую точность.

Предмет и объект исследования. Предметом исследования является взаимосвязь между моделированием информационных технологий при осуществлении задач предприятия, приобретением и установкой соответствующего программного пакета с рынка в соответствии с деятельностью предприятия. Таким образом, в целом, организация и применение информационных технологий в области, создание интернет-среды в этой области, а также создание информационно-аналитической базы.

Объектом исследования является отраслевая деятельность и рынок информационных технологий. Таким образом, основное внимание здесь уделяется применению информационных технологий в области предприятий,

использованию программных пакетов и изучению развития этих видов деятельности.

Цели и роли исследования. Вскоре после того, как ручные процессы были автоматизированы, руководство всех компаний и всех отраслей промышленности начало запрашивать данные из своих новых автоматизированных систем. Их запросы были услышаны и незамедлительно отложены, поскольку основное внимание в то время все еще уделялось автоматизации процессов. Однако жажда данных была настолько велика, а инструменты для манипулирования этими данными были настолько ограничены, что для обработки данных не потребовалось много времени, чтобы разделиться на два сегмента: операционные системы и системы поддержки принятия решений.

Информационная база и методы исследования. В базе информации исследования использованы теоретические и практические материалы по теме, научные труды зарубежных и местных авторов, библиографические источники, а также журналы, книги, учебники, научные труды, посвященные применению ИКТ для решения задач ситуационного анализа. В ходе исследования использовались научное восприятие, исследование, сравнение, ситуационный анализ, системный подход и другие методы.

Новизна исследования. Исследование отличается своей сложностью с точки зрения принципов и подходов современного рынка информационных технологий. В нем рассматриваются основные вопросы, анализируются вопросы прогнозирования конкурентоспособности системы, а также хранилища данных, используемые для решения вопросов связанных с ситуационным анализом, оценивается система маркетинга на веб-сервере, а также специальное программное обеспечение. предоставлена технология подготовки бизнес решения.

Практическая ценность исследования. Практическая значимость диссертации связана с решением проблем и анализом хранилищ данных,

связанной с развитием информационных технологий. Таким образом, автоматизированная система, соответствующая функциям рынка, может быть организована на разных предприятиях, так как на этих предприятиях целесообразно устанавливать автоматизированные офисные информационные технологии - компьютерные сети и средства и оборудование для передачи данных, а также другие организационные и коммуникационные технологии.

Для достижения качественных результатов маркетинговых исследований необходимо учитывать характеристики продуктов и услуг предприятий, формирование соответствующего сектора экономики, создание информационных систем, формирование комплекса и выбор методов и инструментов исследования. В области рынка ИТ делится на рынок продуктов, компьютеров, устройств и оборудования, программного обеспечения, информации, информации и Интернета. В этом случае эти рынки должны быть изучены или всесторонне или отдельно.

ГЛАВА I. ИНТЕЛЛЕКТУАЛЬНЫЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ

1. Модели представления знаний.

Некоторые факты, которые характеризуют объекты, процессы, явления и свойства в предметной области, называются данными. Выявленные закономерности предметной области (принципы, связи, законы), которые позволяют выполнить задачи в данной области, называются знаниями.

Представление знаний - это область искусственного интеллекта, основной целью которого является представление знаний таким образом, который облегчает его определение. Одним из главных направлений исследований в сфере искусственного интеллекта являются модели представления знаний. На вопрос почему он является одним из главных и важнейших, ответом служит то, что искусственный интеллект в принципе не может существовать без знаний. В настоящее время разработано достаточное количество разнообразных моделей представления знаний. Каждая из этих моделей обладает своими преимуществами и недостатками, и в связи с этим для каждой конкретной задачи необходимо выбрать именно подходящую модель представления знаний. От этого будет зависеть возможность решения поставленной задачи, а также насколько эффективным является выполнение. В системах искусственного интеллекта используются в основном следующие типы моделей представления знаний: логические, продукционные, фреймы и семантические сети. Подробно опишем каждую из перечисленных моделей представления знаний.

Логическая модель представления знаний.

Для представления математического знания пользуются логическими формализмами исчисления предикатов и исчисления высказываний. Они имеют понятную формальную семантику и для них разработаны способы

вывода. В связи с этим первым логическим языком было исчисление предикатов, который находил применение для формального описания предметных областей, связанных с решением математических задач.

Логические модели представления знаний реализуются средствами логики предикатов. Функция, которая принимает два значения (истина, ложь) называется предикатом. Она как было выше указано предназначена для выражения свойств объектов или связей между ними.

Высказыванием, называется выражение, в котором утверждается или отрицается наличие каких-либо свойств у объекта. Для именования объектов предметной области служат константы.

Приведенные ниже примеры являются логическими моделями представления с помощью предикатов и называются атомарными (элементарными) формулами.

1. Предикат СТОЛИЦА (Анкара) означает следующее: Анкара является столицей.
2. Предикат ДРУЖБА (Эльнур, Самир) означает: Эльнур и Самир друзья.

Достоинства логической модели:

1. Используется аппарат математической логики, методы которой достаточно хорошо изучены;
2. Существуют достаточно эффективные методы вывода
3. В базах знаний можно хранить лишь множество аксиом, а все остальные знания можно получать из них по правилам вывода.

Недостатками модели являются:

- модель применима лишь в исследовательских системах, так как предъявляет высокие требования и ограничения.

Продукционная модель представления знаний.

Одним из распространенных моделей представления знаний являются продукционные модели. Продукционная модель – это модель, которая основана на правилах, позволяющая представить знание в виде предложений типа «Если условие, То действие». Правила могут несколько выражений, объединенных логическими связками (Или, И, Не).

Продукционные системы – системы обработки знаний, которые используют продукционную модель.

Приведем пример продукционных правил:

1. Если «число делится на 2», То «число является четным».
2. Если «число делится на 2» и «число делится на 3», То «число делится и на 6».

В рабочей памяти систем, основанных на продукционных моделях, хранятся пары атрибут и значение, истинность которых установлена в процессе решения конкретно поставленной задачи к некоторому текущему моменту времени. Существуют два типа продукционных систем – это системы с прямыми и обратными выводами. Прямые выводы носят характер от фактов к заключениям. При обратных выводах выдвигаются гипотезы вероятностных заключений. Существуют также системы с двунаправленными выводами.

Достоинства модели:

- Простота представления знаний;
- Организация логического вывода.

Недостатками являются:

- Сложность оценки целостного образа знаний;
- Низкая эффективность обработки знаний;
- Неясность взаимных отношений правил.

При разработке небольших систем заметны положительные стороны продукционных моделей знаний, однако при увеличении объёма знаний более заметными становятся слабые стороны.

Фреймовая модель представления знаний.

Фреймовая модель впервые была выдвинута Марвином Минским профессором Массачусетского технологического института, основателя лаборатории искусственного интеллекта. Данная модель представляет собой систематизированную психологическую модель сознания человека и его памяти.

Фрейм (рамка, каркас) – это структура данных для представления некоторого объекта, минимальное возможное описание сущности данного объекта (процесса, явления). Информация, которая относится к фрейму, содержится в составляющих его слотах.

Слот (щель, прорезь) может быть листом иерархии или представлять собой фрейм нижнего уровня.

Идентификатор присваиваемый фрейму, называется именем фрейма. Фрейм должен иметь уникальное имя. Идентификатор присваиваемый слоту, называется соответственно именем слота. Он также должен иметь уникальное имя во фрейме, к которому данный слот принадлежит.

Для фреймовых моделей иерархического типа имеет место понятие указатели наследования, они указывают, какую информацию об атрибутах слотов во фрейме верхнего фрейма наследуют слоты с нижнего уровня. Указатель типа данных слота – указатель атрибутов.

Процедура, которая автоматически запускается при выполнении некоторого условия, называется демон. Они запускаются при обращении к конкретному слоту. Часто используемые из них:

- IF-NEEDED запускается, если в момент обращения к слоту его значение не было установлено,

- IF-ADDED используется при подстановке в слот некоторого значения,
- IF-REMOVED при стирании значения слота.

Фреймы образуют иерархию. Иерархия во фреймовых моделях порождает единую многоуровневую структуру, описывающую либо объект, если слоты описывают только свойства объекта, либо ситуацию или процесс, если отдельные слоты являются именами процедур, присоединенных к фрейму и вызываемых при его актуализации.

Фреймы можно разделить на три вида: экземпляр, образец, класс. Подробно опишем каждую из них. Фрейм экземпляр описывает данное состояние в предметной области, конкретную реализацию фрейма. Фрейм образец – это некий шаблон для описания. Фрейм верхнего уровня для представления совокупности соответственно фрейм класс.

Преимуществом фреймовой модели является ее наглядность и гибкость, а также она способна ярко отражать основу организации человеческой памяти.

Семантическая модель представления знаний.

Семантика - это наука, устанавливающая отношения между символами и объектами, которые они обозначают, т.е. наука, определяющая смысл знаков. В переводе слово семантическая означает смысловая. Данная модель впервые была предложена американским ученым Куилианом.

Семантическая сеть представляет собой ориентированный граф, вершинами которого являются понятия, а ребра графа имеют значение отношения между понятиями.

Конкретные или абстрактные объекты в основном играют роль понятий, а отношение обычно это связи принадлежности, части и так далее. В семантической обязательно должно присутствовать следующие три типа отношений:

- класс – элемент класса (оборудование - компьютер);
- свойство – значение (компьютер – вычислительный);

- пример элемента класса (компьютер - персональный).

Семантическую сеть можно разделить на два класса:

- однородная, которая имеет только один тип отношений;
- неоднородная, которая имеет более одного отношения.

Различают по типу отношений семантическую сеть:

- Бинарную (двоичную), которая связывает два объекта;
- Н-арную, предназначенную для связи нескольких объектов.

Преимуществами семантической модели знаний является ее соответствие представлении об основе организации долговременной памяти человека.

Поиск подграфа и вывод семантической сети процедура сложная и требует больших затрат времени, что и является основным недостатком данной модели представления знаний.

1.2. Экспертные системы.

Экспертная система - это компьютерная программа, предназначенная для решения сложных задач и обеспечения способности принимать решения, подобно специалисту-человеку. Он выполняет это путем извлечения знаний из своей базы знаний, используя правила рассуждения и логического вывода в соответствии с запросами пользователя.

Экспертная система является частью искусственного интеллекта, и первая ЭС была разработана в 1970 году, что стало первым успешным подходом к искусственному интеллекту. Он решает самую сложную проблему в качестве эксперта, извлекая знания, хранящиеся в его базе знаний. Система помогает в принятии решений по сложным задачам, используя факты и эвристику, как человеческий эксперт. Он называется так, потому что он

содержит экспертные знания конкретной области и может решить любую сложную проблему этой конкретной области. Эти системы предназначены для определенной области, такой как медицина, наука и т.д.

Эффективность экспертной системы основана на знаниях эксперта, хранящихся в его базе знаний. Чем больше знаний хранится в информационном хранилище, тем больше система улучшает свою производительность. Одним из распространенных примеров экспертных систем является предложение орфографических ошибок при наборе в окне поисковых служб.

Ниже приведены некоторые популярные примеры экспертной системы:

ДЕНДРАЛ: Это был проект по искусственному интеллекту, который был создан в качестве экспертной системы химического анализа. Он использовался в органической химии для обнаружения неизвестных органических молекул с помощью их масс-спектров и базы знаний по химии.

МИЦИН: Это была одна из самых ранних экспертных систем обратной цепочки, которая была разработана для обнаружения бактерий, вызывающих инфекции, такие как бактериемия и менингит. Он также использовался для рекомендации антибиотиков и диагностики заболеваний свертывания крови.

PXDES: Это экспертная система, которая используется для определения типа и уровня рака легких. Чтобы определить заболевание, нужно сделать снимок с верхней части тела, который выглядит как тень. Эта тень определяет тип и степень вреда.

CaDeT: экспертная система CaDet - это диагностическая система поддержки, которая может выявлять рак на ранних стадиях.

Характеристики экспертной системы

Высокая производительность: экспертная система обеспечивает высокую производительность для решения любого типа сложной проблемы конкретной области с высокой эффективностью и точностью.

Понятно, он отвечает так, что пользователь может легко понять его. Он может принимать ввод на человеческом языке и предоставлять вывод таким же образом. Он очень надежен для получения эффективного и точного результата.

Высокая чувствительность: ЭС обеспечивает результат для любого сложного запроса в течение очень короткого периода времени.

Компоненты экспертной системы.

Экспертная система в основном состоит из трех компонентов:

1. Пользовательский интерфейс
2. Механизм логического вывода
3. База знаний

1. Пользовательский интерфейс

С помощью пользовательского интерфейса экспертная система взаимодействует с пользователем, принимает запросы в качестве входных данных в удобочитаемом формате и передает их в механизм вывода. Получив ответ от механизма логического вывода, он отображает вывод для пользователя. Другими словами, это интерфейс, который помогает неопытному пользователю общаться с экспертной системой, чтобы найти решение.

2. Механизм логического вывода

Механизм вывода известен как мозг экспертной системы, поскольку он является основным процессором системы. Он применяет правила вывода к базе знаний, чтобы получить заключение или вывести новую информацию. Это помогает получить безошибочное решение запросов, задаваемых пользователем.

С помощью механизма вывода система извлекает знания из базы знаний.

Существует два типа механизма вывода:

Детерминированный механизм вывода: выводы, сделанные из этого типа механизма вывода, считаются верными. Он основан на фактах и правилах.

Механизм вероятностного вывода. Этот тип механизма вывода содержит неопределенность в выводах и основывается на вероятности.

Механизм логического вывода использует следующие режимы для выведения решений.

Прямая цепочка: она начинается с известных фактов и правил и применяет правила вывода, чтобы добавить их заключение к известным фактам.

Обратная цепочка: это метод обратной рассуждения, который начинается с цели и работает в обратном направлении, чтобы доказать известные факты.

3. База знаний

База знаний - это тип хранилища, в котором хранятся знания, полученные от разных экспертов конкретной области. Это считается большим хранилищем знаний. Чем больше база знаний, тем точнее будет экспертная система.

Это похоже на базу данных, которая содержит информацию и правила определенного домена или субъекта.

Можно также просмотреть базу знаний в виде коллекций объектов и их атрибутов. Например, Лев - это объект, и его атрибуты - это млекопитающее, это не домашнее животное и т.д.

Компоненты базы знаний:

Фактические знания. Знания, основанные на фактах и принятые специалистами по знаниям, подпадают под фактические знания.

Эвристическое знание: это знание основано на практике, умении угадывать, оценке и опыте.

Представление знаний: используется для формализации знаний, хранящихся в базе знаний, с использованием правил условия.

Приобретение знаний: Это процесс извлечения, организации и структурирования знаний в предметной области, определения правил получения знаний от различных экспертов и сохранения этих знаний в базе знаний.

Разработка Экспертной Системы.

Здесь мы объясним работу экспертной системы на примере MYCIN ES. Ниже приведены некоторые шаги для создания MYCIN:

Во-первых, ЭС следует подпитывать экспертными знаниями. Что касается MYCIN, специалисты-люди, специализирующиеся в медицинской области бактериальной инфекции, предоставляют информацию о причинах, симптомах и другие знания в этой области.

КБ MYCIN успешно обновлен. Для того, чтобы проверить это, доктор создает новую проблему для него. Проблема состоит в том, чтобы идентифицировать присутствие бактерий, вводя данные пациента, включая симптомы, текущее состояние и историю болезни.

ЭС потребуется анкета, которую пациент должен заполнить, чтобы узнать общую информацию о пациенте, такую как пол, возраст и т.д.

Теперь система собрала всю информацию, поэтому она найдет решение проблемы путем применения правил if-then с использованием механизма логического вывода и фактов, хранящихся в компонентах базы.

В конце концов, он предоставит ответ пациенту с помощью пользовательского интерфейса.

Участники разработки экспертной системы.

В построении Экспертной Системы есть три основных участника:

Эксперт: Успех ЭС во многом зависит от знаний, предоставленных человеческими экспертами. Эти эксперты - те люди, которые специализируются в этой конкретной области.

Инженер знаний: Инженер знаний - это человек, который собирает знания у экспертов в области и затем систематизирует эти знания в системе в соответствии с формализмом.

Конечный пользователь: это конкретное лицо или группа людей, которые не могут быть экспертами, и для работы в экспертной системе необходимо решение или совет для его запросов, которые являются сложными.

Прежде чем использовать какую-либо технологию, мы должны иметь представление о том, зачем использовать эту технологию и, следовательно, то же самое для ЭС. Несмотря на то, что у нас есть специалисты в каждой области, тогда какова необходимость разработки компьютерной системы. Ниже приведены пункты, которые описывают необходимость ЭС:

Нет ограничений памяти: он может хранить столько данных, сколько требуется, и может запомнить их во время применения. Но для людей-экспертов есть некоторые ограничения, чтобы запомнить все вещи в любое время.

Высокая эффективность: если база знаний обновляется с правильными знаниями, то она обеспечивает высокоэффективный результат, который может быть невозможен для человека.

Экспертиза в области: в каждой области много экспертов-людей, и все они имеют разные навыки, разный опыт и разные навыки, поэтому получить конечный результат для запроса непросто. Но если мы поместим знания, полученные от человеческих экспертов, в экспертную систему, то это даст эффективный результат, смешав все факты и знания

Не подвержен влиянию эмоций. Эти системы не подвержены влиянию эмоций человека, таких как усталость, гнев, депрессия, беспокойство. Следовательно, производительность остается постоянной.

Высокий уровень безопасности: эти системы обеспечивают высокий уровень безопасности для разрешения любого запроса.

Учитывает все факты. Чтобы ответить на любой запрос, он проверяет и учитывает все доступные факты и соответственно предоставляет результат. Но возможно, что человеческий эксперт не может рассмотреть некоторые факты по любой причине.

Регулярные обновления повышают производительность: если в результате, предоставляемом экспертными системами, возникает проблема, мы можем улучшить производительность системы, обновив базу знаний.

Возможности экспертной системы.

Ниже приведены некоторые возможности экспертной системы:

Консультирование: оно способно консультировать человека по запросу любого домена из конкретной ЭС.

Предоставляют возможности для принятия решений. Он обеспечивает возможность принятия решений в любой области, например, для принятия любых финансовых решений, решений в области медицинской науки.

Демонстрация устройства: оно способно демонстрировать любые новые продукты, такие как его функции, спецификации, как использовать этот продукт.

Решение проблем: у него есть возможности решения проблем.

Объяснение проблемы: оно также способно предоставить подробное описание проблемы ввода.

Интерпретация ввода: он способен интерпретировать ввод, предоставленный пользователем.

Прогнозирование результатов: его можно использовать для прогнозирования результата.

Диагностика: ЭС, разработанный для медицинской области, способна диагностировать заболевание без использования нескольких компонентов,

поскольку она уже содержит различные встроенные медицинские инструменты.

Преимущества экспертной системы.

Эти системы очень воспроизводимы.

Их можно использовать в опасных местах, где присутствие человека небезопасно.

Производительность этих систем остается стабильной, поскольку на нее не влияют эмоции, напряжение или усталость.

Они обеспечивают очень высокую скорость ответа на конкретный запрос.

Ограничения экспертной системы.

Ответ экспертной системы может быть неправильным, если база знаний содержит неверную информацию.

Как и человек, он не может создавать творческий результат для разных сценариев.

Затраты на его обслуживание и разработку очень высоки.

Приобретение знаний для проектирования намного сложнее.

Для каждого домена нам требуется определенный ES, который является одним из больших ограничений.

Он не может учиться у себя и, следовательно, требует ручного обновления.

1.3. Нейронные сети.

Нейронные сети представляют собой глубокое обучение с использованием искусственного интеллекта. Некоторые прикладные сценарии слишком тяжелы или выходят за рамки традиционных алгоритмов машинного обучения. Нейронные сети, как известно создают такие сценарии и заполняют

пробел. Искусственные нейронные сети основаны на биологических нейронах человеческого тела, которые активируются при определенных обстоятельствах, что приводит к связанному действию, выполняемому организмом в ответ.

Искусственные нейронные сети состоят из различных слоев взаимосвязанных искусственных нейронов, приводимых в действие функциями активации, которые помогают включать или выключать их. Подобно традиционным машинным алгоритмам, здесь также есть определенные значения, которые нейронные сети изучают на этапе обучения.

Вкратце, каждый нейрон получает умноженную версию входных данных и случайных весов, которые затем добавляются со значением статического смещения (уникальным для каждого слоя нейрона), а затем передаются в соответствующую функцию активации, которая решает, какое окончательное значение будет дано из нейрона.

Существуют различные функции активации, доступные в зависимости от характера вводимых значений. После того, как выходные данные сгенерированы из конечного слоя нейронной сети, вычисляется функция потерь (входные данные против выходных данных) и выполняется обратное распространение, где веса корректируются, чтобы сделать минимум потерь. Поиск оптимальных значений весов - это то, на чем сфокусирована вся операция.

Входной слой представляет размеры входного вектора.

Скрытый слой представляет собой промежуточные узлы, которые делят входное пространство на области с (мягкими) границами. Он принимает набор взвешенного ввода и производит вывод через функцию активации.

Выходной слой представляет собой выход нейронной сети.

Веса - это числовые значения, которые умножаются на входные данные. При обратном распространении они модифицируются для уменьшения

потерь. Проще говоря, веса - это машинные значения, полученные из нейронных сетей. Они само настраиваются в зависимости от разницы между прогнозируемыми результатами и результатами обучения.

Функция активации - это математическая формула, которая помогает нейрону включать или выключать.

Существует множество типов нейронных сетей, которые могут находиться на стадии разработки. Они могут быть классифицированы в зависимости от их структуры, потока данных, используемых нейронов и их плотности, слоев и их фильтров активации глубины. Есть 7 типов нейронных сетей:

1. Нейронная сеть с прямой связью

Простейшая форма нейронных сетей, где входные данные перемещаются только в одном направлении, проходя через искусственные нейронные узлы и выходя через выходные узлы. Там, где скрытые слои могут присутствовать или отсутствовать, присутствуют входные и выходные слои. Исходя из этого, они могут быть далее классифицированы как однослойные или многослойные нейронные сети с прямой связью. Количество слоев зависит от сложности функции. Он имеет однонаправленное прямое распространение, но не имеет обратного распространения. Веса здесь статичны. Функция активации подается на входы, которые умножаются на весовые коэффициенты. Для этого используется классифицирующая функция активации или функция пошаговой активации. Например, нейрон активируется, если он превышает пороговое значение (обычно 0), и нейрон выдает 1 в качестве выхода. Нейрон не активируется, если он ниже порога (обычно 0), который считается -1. Они довольно просты в обслуживании и оснащены для работы с данными, которые содержат много шума.

Преимущества нейронных сетей прямой связи является ее менее сложность, простой в разработке и обслуживании. К недостаткам можно

отнести, то что не может использоваться для глубокого обучения, из-за отсутствия плотных слоев и обратного распространения.

2.Многослойный персептрон

Точка входа в сложные нейронные сети, где входные данные проходят через различные слои искусственных нейронов. Каждый отдельный узел связан со всеми нейронами в следующем слое, что делает его полностью подключенной нейронной сетью. Входные и выходные слои имеют несколько скрытых слоев, то есть, по крайней мере, три или более слоев в общей сложности. Он имеет двунаправленное распространение, то есть прямое распространение и обратное распространение. Входы умножаются на весовые коэффициенты и передаются в функцию активации, а при обратном распространении они модифицируются для уменьшения потерь. Проще говоря, веса - это машинные значения, полученные из нейронных сетей. Они само настраиваются в зависимости от разницы между прогнозируемыми результатами и результатами обучения.

Преимущества многослойного персептрона является использование для глубокого обучения из-за наличия плотных полностью связанных слоев и обратного распространения. Недостатки то сравнительно сложно проектировать и поддерживать, а также сравнительно медленно, зависит от количества скрытых слоев.

3. Свертка нейронной сети

Сверточные нейронные сети (рис 1.1.) содержат трехмерное расположение нейронов вместо стандартного двумерного массива. Первый слой называется сверточным слоем. Каждый нейрон в сверточном слое обрабатывает информацию только из небольшой части поля зрения.

Сеть понимает изображения по частям и может вычислять эти операции несколько раз, чтобы завершить полную обработку изображения. Обработка включает в себя преобразование изображения из шкалы RGB или HSI в шкалу

серого. Дальнейшее изменение значения пикселя поможет обнаружить края, и изображения можно классифицировать по разным категориям.

Распространение является однонаправленным, где CNN содержит один или несколько сверточных слоев с последующим пулированием и двунаправленным, когда вывод сверточного слоя направляется в полностью подключенную нейронную сеть для классификации изображений. Фильтры используются для извлечения определенных частей изображения.

В MLP входы умножаются на веса и подаются на функцию активации. Сверточные нейронные сети показывают очень эффективные результаты в распознавании изображений и видео, семантическом разборе и обнаружении перефразирования.

Преимущества данного типа нейронной сети, ее использование для глубокого обучения с несколькими параметрами, меньше параметров для изучения по сравнению с полностью подключенным слоем. Недостатки в данном случае то сравнительно сложно проектировать и поддерживать, сравнительно медленно.

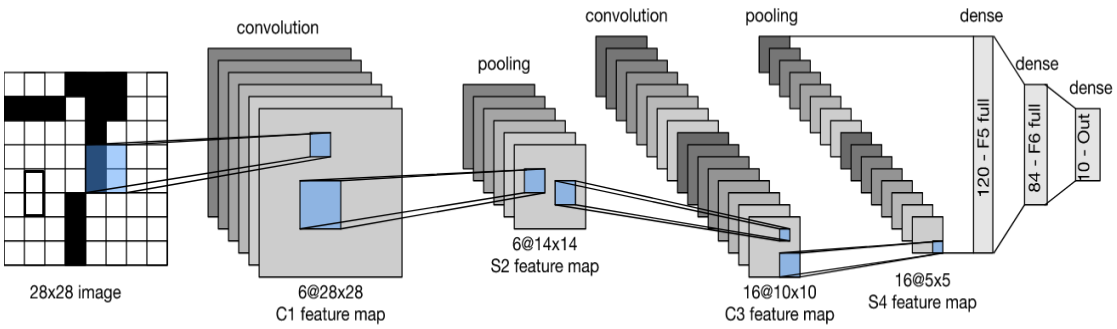

Рис. 1.1. Структура модели свертки

4. Радиальные базисные функции нейронных сетей

Радиальная базисная функциональная сеть состоит из входного вектора, за которым следует слой нейронов RBF и выходной слой с одним узлом на категорию. Классификация выполняется путем измерения сходства входных данных с точками данных из обучающего набора, где каждый нейрон хранит прототип. Это будет один из примеров из учебного набора. Когда необходимо классифицировать новый входной вектор (n-мерный вектор, который вы пытаетесь классифицировать), каждый нейрон вычисляет евклидово расстояние между входом и его прототипом. Например, если у нас есть два класса, то есть класс A и класс B, то новый классифицируемый ввод более близок к прототипам класса A, чем прототипам класса B. Следовательно, он может быть помечен или классифицирован как класс A. Каждый нейрон RBF сравнивает входной вектор со своим прототипом и выдает значение в диапазоне, которое является мерой сходства от 0 до 1. Поскольку входной сигнал равен прототипу, выход этого Нейрон RBF будет равен 1, и с увеличением расстояния между входом и прототипом отклик экспоненциально падает до 0. Кривая, генерируемая из отклика нейрона, стремится к типичной кривой колокола. Выходной слой состоит из набора нейронов (по одному на категорию).

5.Рекуррентные нейронные сети

Предназначенная для сохранения выходных данных слоя, рекуррентная нейронная сеть возвращается на вход, чтобы помочь в прогнозировании результата слоя. Первый уровень обычно представляет собой нейронную сеть с прямой связью, за которой следует рекуррентный уровень нейронной сети, где некоторая информация, которая была у него на предыдущем временном шаге, запоминается памятью. В этом случае осуществляется прямое распространение. Он хранит информацию, необходимую для его будущего использования. Если прогноз неверен, скорость обучения используется для внесения небольших изменений. Следовательно, постепенно увеличивая его в сторону правильного прогноза во время обратного распространения.

Преимущества рекуррентных нейронных сетей является, последовательные данные модели, где можно предположить, что каждая выборка зависит от исторических, является одним из преимуществ. Используется со сверточными слоями для увеличения эффективности пикселей. Недостатками рекуррентных нейронных сетей являются проблемы градиента исчезновения и взрыва, обучение повторяющихся нейронных сетей может быть сложной задачей, а также сложно обрабатывать длинные последовательные данные.

6.Последовательность к моделям последовательности

Модель последовательности к последовательности состоит из двух рекуррентных нейронных сетей. Здесь существует кодер, который обрабатывает ввод, и декодер, который обрабатывает вывод. Кодер и декодер работают одновременно - либо используя один и тот же параметр, либо разные. Эта модель, в отличие от фактического RNN, особенно применима в тех случаях, когда длина входных данных равна длине выходных данных. Несмотря на то, что они обладают аналогичными преимуществами и ограничениями RNN, эти модели обычно применяются в основном в чат-роботах, машинных переводах и системах ответов на вопросы.

7. Модульная нейронная сеть

Модульная нейронная сеть (рис. 1.2.) имеет ряд различных сетей, которые функционируют независимо и выполняют подзадачи. Различные сети на самом деле не взаимодействуют и не сигнализируют друг другу в процессе вычислений. Они работают независимо для достижения результата.

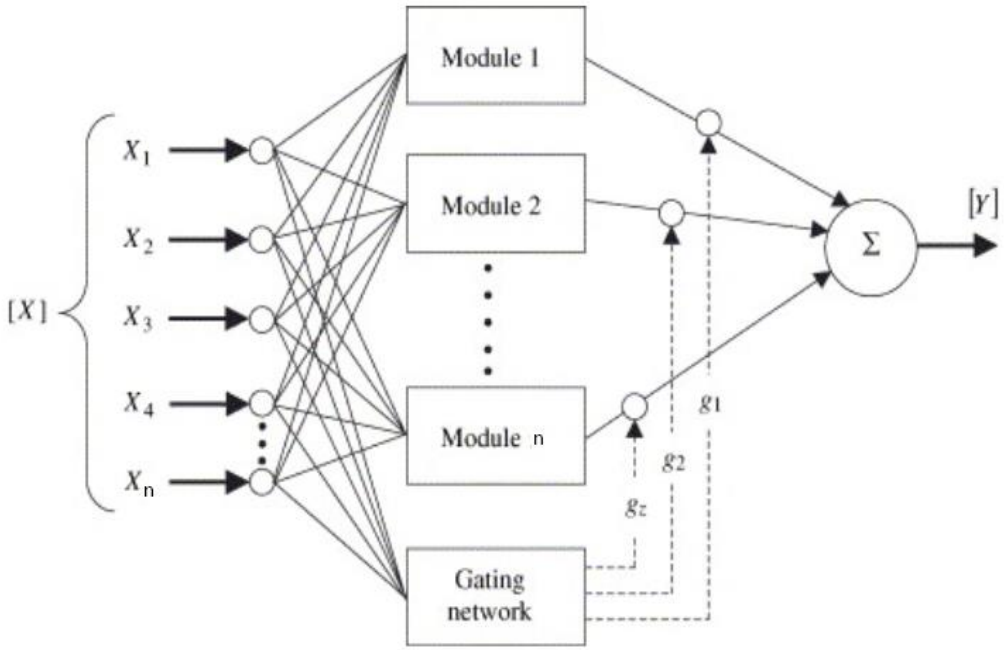

Рис. 1.2. Схема модульной нейронной сети

В результате большой и сложный вычислительный процесс выполняется значительно быстрее, разбивая его на независимые компоненты. Скорость вычислений увеличивается, потому что сети не взаимодействуют и даже не связаны друг с другом.

Преимущества модульной нейронной сети эффективное независимое обучение, прочность, а недостатком модульной нейронной сети является перемещение цели проблемы.

ГЛАВА 2. МЕТОДЫ И МОДЕЛИ АНАЛИЗА ДАННЫХ: OLAP И DATA MINING

2.1. Хранилища данных

Главные проблемы, которые связаны с анализом информации, как известно, обуславливается уровнем готовности и качеством данных, т.е. когда отсутствуют агрегаты и вычисляемые показатели. Именно поэтому в настоящее время требуемой технологией, которая используется при выполнении интеллектуальной информационной системы, составляют хранилища данных, благодаря которых выполняется задача собрания, очищения и преобразования первичной информации.

Главные идеи, которые лежат на основании теории хранилище данных, являются:

1) Интегрирование разобранных детализированных данных, описывающие определенные точные факты, свойства, процессы и другие в едином хранилище данных;

2) Отделение наборов данных и приложений, которые используются чтобы провести оперативную обработку и решать задачу анализа.

Во время бурного развития регистрирующей информационной системы появилось понятие ограниченности возможности использования база данных для анализа данных и построения на их основании систем поддержки и выбора решений. Эти системы создались для автоматизации рутинных процессов, для введения бизнеса, т.е. выписки счетов, оформлении соглашений, проверки состояния склада и другие. Главный линейный персонал являлся пользователем этих регистрирующих систем.

Главным требованием, предъявляемое к этим системам было обеспечение транзакционности включаемых изменений и увеличение до максимума скорости их решения.

В этой системе информация важна только на время перехода к БД, в последующий момент по тому же запросу возможно получить совершенно иной ответ. Интерфейс РС рассчитывается на жесткое выполнение некоторых операций. Возможность получения результата на нерегламентированный запрос в значительной степени ограничен. Благодаря настройки системы управления базы данных на решение коротких транзакций и неизбежных замедлений выполнения других пользователей малы возможности обработки больших массивов данных.

Для выполнения этих требований появилось новая технология организации без данных, т.е. технология хранилища данных. Система, которая содержит непротиворечивый предметно-ориентированный комплекс исторических данных огромной корпорации или другой подготовки для того, чтобы поддержать принятие стратегических задач, называется хранилищем данных. Благодаря извлечению моментальных выделений базы данных ОИС организации и разных наружных материалов создаются информационные ресурсы хранилища данных.

Если эффективно применять хранилища данных, она может быть главным источником точной информации для руководящих и профессионалов всех классов организации. Благодаря хранилищу данных собирается, очищается, загружается, агрегируется, хранятся данные и предоставляется к этим данным быстрый доступ. Хранилище данных, можно сказать, что это комплекс источника данных, в котором собираются информации для последующей обработки, процессов извлечения, преобразования и загрузки данных. Физически ХД представляет собой реляционную БД. Но по сравнению с базой данных КИС это хранилище имеет принципиально другую структуру. База данных корпоративных информационных систем в отличие от

хранилища данных включает детализированные данные. Период сохранения этих данных относительно малый.

Классическую архитектуру хранилища данных составляет следующие элементы:

1. Реляционная БД
2. Многомерная БД
3. Средство извлечения
4. Очистка и загрузка данных
5. Средство визуализации данных и генерации отчетов

На рис 2.1. показана схема концептуальной модели ХД.

Рис. 2.1. Концептуальная модель хранилища данных

Данные из разных источников помещаются в хранилище данных. Применяя разные инструменты визуализации, а также содержимое

репозитория, конечный пользователь анализирует данные в хранилище. Информация, представляющая собой готовые отчеты, найденные скрытые закономерности, какие – либо прогнозы, является результатом его работы. Поскольку средство деятельности конечного пользователя с ХД моет быть самым разнообразным, то их выбор, согласно теории, не должен оказывать влияние на его структуру и функции его сохранения в актуальном положении.

Свойства ХД связаны со свойствами задач, которыми являются следующие:

1. Аналитическая оперативная обработка информации
2. Сложные для оперативных БД SQL – запросы

На основании ХД создаются подмножества данных – OLAP-кубы, которые являются многомерными иерархическими структурами данных. Они содержат много показателей:

1) Дата (время);

2) Область деятельности, к которой относится данные;

3) Субъект управления;

4) Вид ресурса и т.д.

Благодаря этим показателям агрегируется данные с помощью совокупности показателей и решения статистических оценок. Когда осуществляется анализ информации, образуется новое знание, которое полезно для задач управления.

Здесь появляется вопрос, откуда данные попадают в хранилище? Конечно же от оперативных систем, предназначенные для автоматизации бизнес-процессов. Кроме этого, хранилище может пополнен внешними источниками, например, статистическими отчетами.

На рис 2.2. показаны компоненты, которые входят в типичное хранилище.

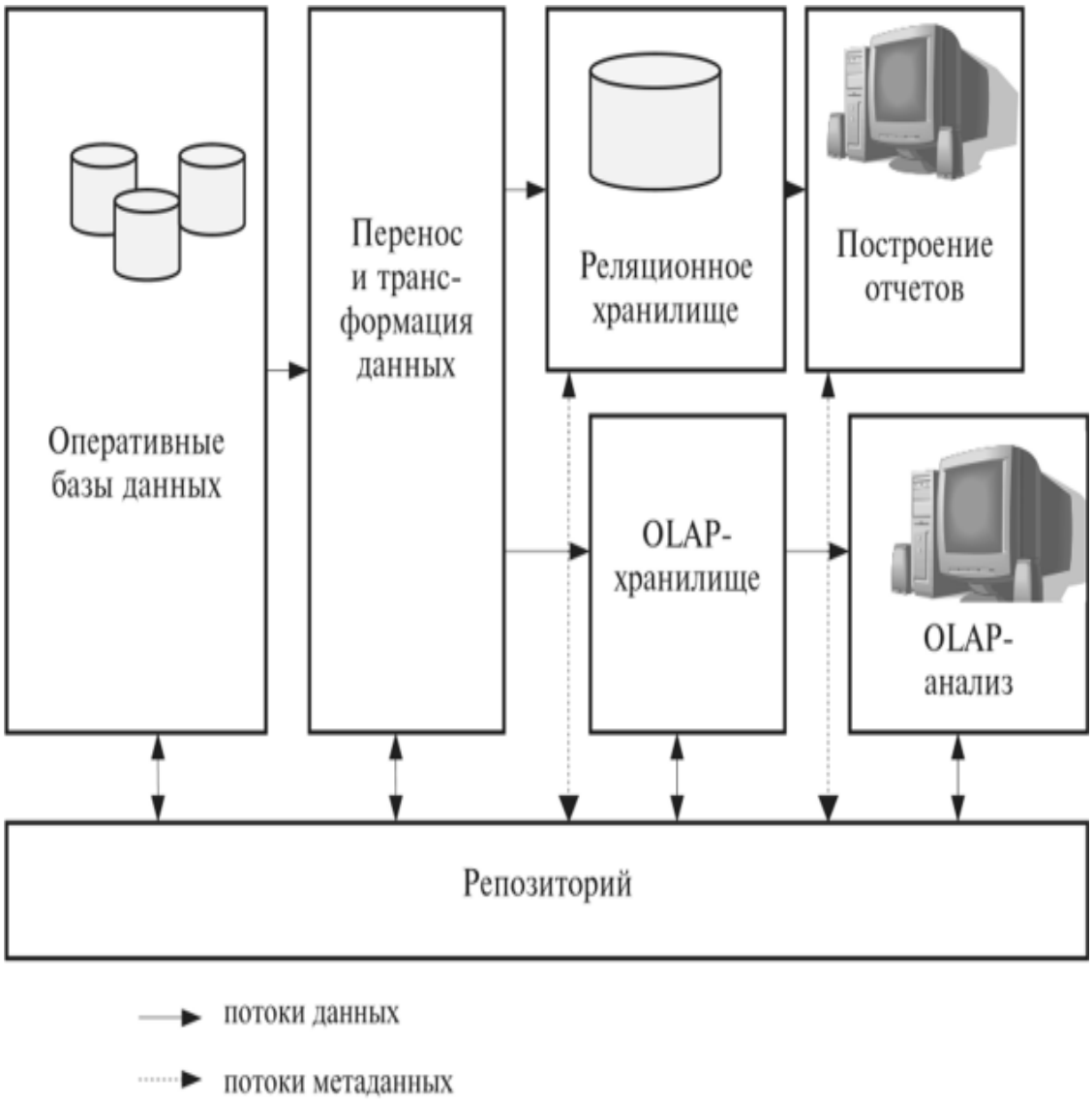

Рис. 2.2. Структура хранилища данных

Оперативные данные собираются из разных источников, очищаются, объединяются и складываются в реляционное ХД. Тогда эти данные уже доступны для анализа с помощью разных материалов построения отчетов. После этого данные подготавливаются для OLAP-анализа. Данные могут быть загружаться в специальную базу данных OLAP или остаться в реляционном ХД. Важным элементом данных являются метаданные, то есть информация о структуре, расположении и преобразования данных.

Эффективное взаимодействие разных компонентов обеспечивается за счет этих данных. Можно сказать, что предоставление сырья для анализа в одном месте и в простейшей, понятной структуре, называется задачей

хранилища. Существует еще одно основание, которое оправдывает возникновение отдельного хранилища. Текущая работа компании тормозится вследствие сложных аналитических запросов к оперативной информации, которые надолго блокируют таблицы и захватывают ресурсы сервера. Главными причинами, которые побуждают организацию внедрения ХД, являются:

1. Важность применения МД и технологии, которые ускоряют процесс решения запросов и подготовки отчетливости, но не применяемые для обработки транзакции
2. Образование среды, в которой даже сравнительно небольшие знания основ системы управления базой данных хватит для создания запросов и подготовки отчет
3. Важность решения аналитических запросов и генерации отчет на незадействованных главными информационными системами вычислительных средствах
4. Облегчение процесса подготовки отчет на основании информации из некоторых систем транзакции или наружных источников данных или данных, которые используются только для генерации отчет
5. Защита конечного пользователя от важности в любой степени вникать в состав и логику деятельности базы данных РС.
6. Образование выделяемого источника в тот момент, когда возможности ОС не соответствуют сроку хранения данных, которое требуется бизнесом. Важность иметь возможность подготовки отчет на определенные моменты времени в прошлом

Логическим следствием улучшения и усложнения информационно-логических структур, которые обрабатываются за счет компьютера называется перевод от данных к знаниям. Областью применения современных компьютеров, которая развивается активно в настоящее время, является создание БЗ и их использование в разных областях науки и технологии.

Закономерности предметных областей, которые получаются благодаря практической работы и профессиональной практики, позволяющие работающим поставить и выполнить задачи в этой области, называются знаниями.

Знания можно рассмотреть в виде стратегической информации, важная для создания цели и построения кинематической пути, а данные в виде оперативных знаний, которые применяются системой в динамическом процессе. Совокупность знаний, которая накапливаются человеком в определенной предметной области, выраженная благодаря некоторому языку представления знаний, называется базом знаний.

Для образования баз знаний, необходимо разработка соответствующих программных средств. Благодаря этим средствам обеспечивается загрузка, актуализация, поддержание в достоверном положении, увеличение баз знаний, создание, обработка и включение новых знаний, которые соответствуют текущему условию. База знаний составляет основу экспертной системы.

2.2 OLAP-системы

Благодаря применению OLAP-системы осуществляется автоматизирование стратегического уровня управления организацией. OLAP с английского переводится, как аналитическая обработка данных в реальном времени (Online Analytical Processing). Эта система также является мощной технологией обработки и обнаруживания данных. Системы, которые строятся на основании технологии OLAP, охватывают почти многие безграничные возможности. Например, составление отчет, выполнение сложнейших аналитически задач, построение прогнозов и сценарий, разработка множества методов планов.

Эти системы возникли в начале 90-ых годов, ставшим результатом формирования информационных систем поддержки утверждении решений. Различные, часто разрозненные, данные преобразуются в полезную информацию именно за счет систем OLAP. Эти системы могут составить данные соответственно некоторым наборам критерий. Но при этом критерии могут иметь нечеткие характеристики, это не обязательно. Свое использование эти системы нашли в разных вопросах стратегического управления учреждением. Ими являются следующие:

1. Управление эффектностью бизнеса
2. Стратегическое планирование
3. Бюджетирование
4. Прогноз развития
5. Подготовка финансовой отчетности
6. Анализ деятельности
7. Имитационное моделирование наружной и внутренней среды учреждения
8. Хранение данных и отчетности

Структура OLAP системы.

Основа деятельности OLAP системы представляет собой обработку многомерных массивов данных. Эти массивы устраиваются так, что любой элемент массива имеет много связей с другими элементами. Для этой системы необходимой является получение исходных данных из других систем или через наружный ввод, чтобы создать этот массив. Пользователь этой системы приобретает важные для себя данные в виде структуры в соответствии со своим интересом. Благодаря порядкам действий составляется структура OLAP системы. Эта структура сложная, поскольку состоит из нескольких элементов.

Этими элементами являются следующие:

1. БАЗА ДАННЫХ. База данных представляет собой источник информации для деятельности OLAP системы. Его тип зависит от типа этой

системы и алгоритма работы сервера системы. Как известно применяются следующие виды

- реляционные базы данных
- многомерные базы данных
- хранилища данных и др.

2. OLAP СЕРВЕР. Этот сервер обеспечивает связь между БД и пользователями OLAP системы и управление многомерными структурами данных.

3. ПОЛЬЗОВАТЕЛЬСКИЕ ПРИЛОЖЕНИЯ. Эти приложения осуществляют управление интересами пользователей и формируют ответы обращения к БД.

Виды OLAP систем.

Согласно, методу хранения и обработки данных все эти системы разделяются на 3 главные вида. Ими являются следующие:

1. ROLAP.

Этот тип системы осуществляет работу с реляционной БД. Переход к данным происходит прямо в реляционную систему хранилища информации, в таблице, которой содержатся данные информации. В данном случае у пользователей имеется способность выполнять многомерный анализ нам известных в данной случае традиционных OLAP систем. Выполнение вышеуказанных целей осуществляется благодаря использования языка структурированных запросов и в том же числе запросов носящих специальный характер.

Преимущественной особенностью данного вида (рис. 2.3.) является характерность обрабатывания огромного количества информации. К этим же преимуществам можно отнести один из главных факторов обработки информации, это эффективное применение ее для текстовых, в том числе и информации носящих числовой характер.

Переходя к недостаткам можно определить ее низкую производительную способность. Также нужно упомянуть ограниченность его функциональности за счет использования языка структурированных запросов.

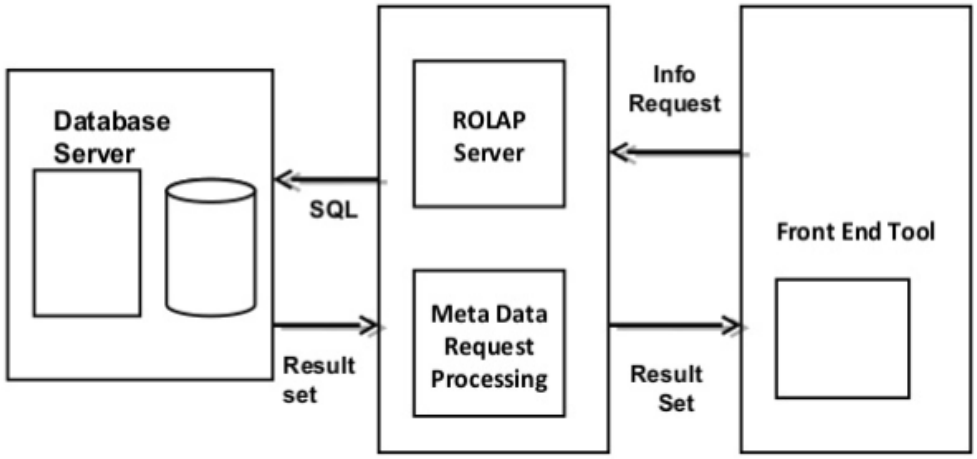

Рис. 2.3. Структура ROLAP

2. MOLAP.

В указанном случае будем насматривать многомерную систему. Данный тип (рис. 2.4.) является примером известных нам традиционных систем, а отличительной чертой данного типа является в том, что информация в том случае подготавливается предварительно, а также осуществляется ее оптимизация. По правилу в таких системах за основу берется отдельный сервер, в которой и происходит вышеуказанная подготовка для преждевременной обработки информации. В указанном случае информации хранится в многомерных массивах, а более точнее в OLAP кубах.

Благодаря простоты реорганизации информации в том же числе ее способностью структурировать информацию для разных типов запросов пользователя, данная система имеет статус эффективной системы. Также к преимуществам можно отнести выполнение сложных вычислений и скоростью выполнения запросов, что и происходит благодаря обработки предварительно вышеуказанных кубов.

Ограниченность объема информации, которая обрабатывается является ее главным недостатком, а также к тому можно добавить и случай создания кубов, информацию в случае чего следует копировать.

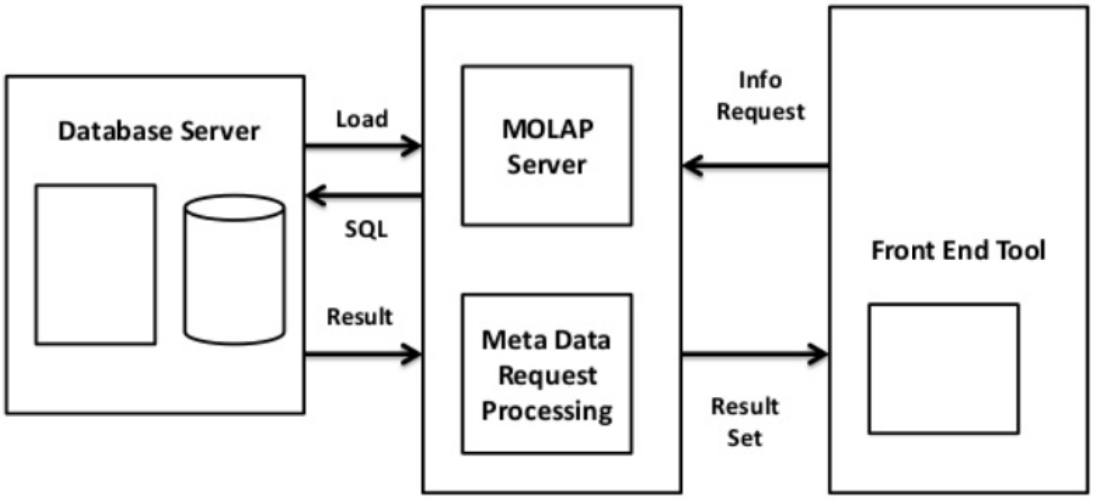

Рис. 2.4. Структура MOLAP

3. HOLAP.

Данная система является гибридной системой и состоит из объединения вышеуказанных систем ROLAP и MOLAP. В системах являющимися гибридными было решение соединить две данные системы: использовать многомерные хранилища информации и в том же числе способность управления хранилищами информации реляционного типа. Эти системы дают способность хранения огромного объема информации в реляционных таблицах, а обработанные информации хранятся в созданных заранее многомерных кубов.

Преимуществом данной системы является в расширение информации, ее скорости обработки и доступном в гибкости к ее источнику информаций.

Преимущества OLAP.

Одним из основных преимуществ OLAP является согласованность информации и расчетов. Независимо от того, сколько или как быстро данные

обрабатываются с помощью программного обеспечения или серверов OLAP, отчеты о результатах представляются в согласованной форме, поэтому аналитики и руководители всегда знают, где и что искать. Это особенно полезно при сравнении информации из предыдущих отчетов с информацией, содержащейся в новых и прогнозируемых будущих. Это позволяет избежать длительных дискуссий о том, кто имеет правильную информацию.

Сценарии “что, если” являются одними из самых популярных видов использования программного обеспечения OLAP и в высшей степени возможны благодаря многомерной обработке.

Еще одно преимущество многомерного представления данных заключается в том, что оно позволяет менеджеру извлекать данные из базы данных OLAP в широком или конкретном выражении. Другими словами, составление отчетов может быть таким же простым, как сравнение нескольких строк данных в одном столбце электронной таблицы, или настолько сложным, как просмотр всех аспектов массива данных.

Кроме того, многомерное представление может создать понимание отношений, ранее не реализованных.

OLAP создает единую платформу для всех информационных и бизнес-потребностей; планирование, бюджетирование, прогнозирование, отчетность и анализ.

И последнее, но не менее важное: кривая обучения использованию OLAP минимальна. Наиболее используемый интерфейс для анализа данных, хранящихся в технологии OLAP, - это хорошо известная и любимая электронная таблица.

2.3. Интеллектуальный анализ данных

По мере того, как все больше и больше данных собирается и хранится каждый день, были разработаны различные методы анализа данных, основанные на работах по распознаванию образов, статистике, искусственному интеллекту, машинному обучению, системе баз данных. Задача интеллектуального анализа данных (ИАД) включает в себя обнаружение знаний, прогнозирование, моделирование процессов, систем или построение систем, основанных на знаниях. Есть много достижений применения методов ИАД в различных областях, таких как маркетинг, медицина, финансы и сельское хозяйство. Инструменты и приложения ИАД дали положительные результаты и постоянно стимулировали исследование новых областей применения благодаря преимуществам, которые дает эта технология.

Быстро растущий объем данных в режиме реального времени, вызванный взрывом активности в сети, мультимедиа, электронной коммерции и других, способствовал спросу и предоставлению более сложных методов ИАД. Общая идея анализа больших объемов данных с богатым описанием является привлекательной и интуитивно понятной, но технически она значительно сложнее и труднее. Должны быть некоторые стратегии, которые должны быть реализованы для лучшего использования данных, собранных из таких больших и сложных источников данных.

Прежде чем обсуждать технические проблемы интеллектуального анализа данных и их последствия, кратко представим типичный процесс анализа данных и различные возможные задачи, и методы. Интегрированная система анализа данных, использующая методы машинного обучения, используется для анализа отраслевого приложения. Эмпирические результаты в базе данных о пересечении уровней демонстрируют, что методы машинного обучения могут применяться для

обнаружения скрытой информации из реальных наборов данных с высокой точностью и хорошей понятностью.

Пример процесса ИАД.

Интеллектуальный анализ данных - это процесс поиска полезных и интересных структур из данных, что помогает принимать решения. Типичный процесс ИАД начинается с выявления проблемы в зависимости от интереса аналитика данных. Затем идентифицируются все источники информации, и из накопленных данных для приложения ИАД генерируется подмножество данных. Для обеспечения качества набор данных предварительно обрабатывается путем удаления шума, обработки недостающей информации и преобразования в соответствующий формат. Метод ИАД или комбинация методов, соответствующих типу знаний, которые будут обнаружены, затем применяется к производному набору данных. Обнаруженные знания затем обрабатываются, оцениваются и интерпретируются, как правило, с использованием некоторых инструментов последующей обработки, таких как методы визуализации. Наконец информация предоставляется пользователю. Иногда этот процесс включает в себя поддержание результатов путем повторения всех этапов снова для удовлетворения пользователей или адаптации новой информации в будущем.

Обычно полученные знания представляют собой тип правил классификации, характерных правил, правил ассоциации, функциональных отношений, функциональных зависимостей, причинных правил, временных знаний или кластеров.

Различные анализ данных и методы.

В соответствии с целями и интересами конечного пользователя, такими как характеристика содержимого набора данных в целом или установление связей между подмножествами шаблонов в наборе данных,

анализ данных процесс может иметь три возможных задачи - прогнозное моделирование, кластеризацию и анализ связей.

Цель прогностического моделирования - делать прогнозы на основе основных характеристик данных. Цель состоит в том, чтобы построить модель для отображения элемента данных в один из нескольких предопределенных классов или в переменную прогнозирования с реальной стоимостью. Любой контролируемый алгоритм машинного обучения, который изучает модель по предыдущему или существующие данные, могут быть использованы для выполнения прогнозного моделирования. В модели приведены некоторые уже известные факты с правильными ответами, из которых модель учится делать точные прогнозы. К ним относятся: Нейронные сети, деревья решений, байесовские классификаторы, классификаторы K-ближайших соседей, рассуждения на основе случая, генетические алгоритмы, грубый набор и нечеткий набор - некоторые из подходов, используемых для отображения дискретных целевых переменных.

Методы регрессии, индукционные деревья, нейронные сети и радиальная базисная функция являются одними из подходов, используемых для отображения целевых переменных с непрерывным значением.

Целью кластеризации является идентификация элементов с похожими характеристиками и, таким образом, создание иерархии классов из существующего набора событий. Любой неконтролируемый алгоритм машинного обучения, для которого заранее определенный набор категорий данных не известен для набора входных данных, может использоваться для выполнения кластеризации. В модели приведены некоторые уже известные факты, из которых модель выводит категории данных с аналогичными характеристиками. Некоторые основные методы кластеризации - это алгоритмы разделения, иерархии, плотности и модели.

Анализ связей устанавливает внутренние отношения между элементами в данном наборе данных. Эта цель достигается путем обнаружения ассоциаций, последовательного обнаружения шаблонов и аналогичных задач обнаружения временных последовательностей.

Эти задачи раскрывают образцы и тренды, предсказывая соотношение элементов, которые в противном случае не очевидны. Методы анализа ссылок основаны на подсчете вхождений всех возможных комбинаций элементов. Некоторые из наиболее широко используемых алгоритмов - это априори и его вариация.

Технические проблемы в ИАД и их разветвления.

Существует много препятствий для применения методов ИАД к реальным проблемам, включая отсутствие эффективных и автоматических инструментов предварительной обработки, отсутствие инструментов, подходящих для больших, богатых и сложных наборов данных, отсутствие удобных и эффективных инструментов постобработки и отсутствие действительно интегрированная среда анализа данных. Ниже приводится обсуждение некоторых проблем, которые могут возникнуть в процессе анализа данных, и их предлагаемых решений.

1. Объем данных

С прогрессом в методах сбора данных, анализируемые данные обычно имеют большой объем. Набор данных может быть большим с точки зрения количества шаблонов, случаев, записей, кортежей или количества переменных, признаков, атрибутов, полей. Методы ИАД должны быть соответственно масштабируемы, например, если метод работает хорошо для задачи, включающей тысячи шаблонов, то он должен хорошо работать для одного с миллионами шаблонов, и если метод успешно применяется к задаче с участием десятков переменных, то это должно быть эффективно применено к задаче с сотнями переменных. Методы анализа данных должны удовлетворительно работать с таким большим объемом данных.

Перечисление всех шаблонов и переменных может быть дорогим и не обязательным. Несмотря на это, выбор репрезентативных шаблонов, которые отражают суть всего набора данных и их использование для анализа набора данных, может оказаться более эффективным подходом. Но тогда выбор такого подмножества данных становится проблемой. Более эффективный подход заключается в использовании итеративной и интерактивной техники, которая учитывает ответы в реальном времени и обратную связь при расчете. Интерактивный процесс вовлекает в процесс человеческого аналитика, поэтому в процесс может быть включена мгновенная обратная связь. Итерационный процесс сначала рассматривает выбранное количество атрибутов, выбранных пользователем для анализа или используя алгоритм выбора функции, а затем продолжает добавлять другие атрибуты для анализа, пока пользователь не будет удовлетворен. Новизна этого итеративного метода заключается в том, что он значительно сокращает пространство поиска (из-за меньшего количества задействованных атрибутов). Большинство существующих методов страдают от (очень большой) размерности пространства поиска.

Существует значительный прогресс в технологии агентов. Сегодня существуют агенты для поиска и обобщения соответствующей информации в сети или новостных лентах, или других реальных потоках данных с помощью пользовательского профиля поиска. Методы анализа данных могут использовать агентную технологию для решения проблемы больших наборов данных, заключив контракт с несколькими агентами. Каждый агент будет действовать независимо, например, выявляя, получая доступ и сохраняя соответствующие данные, предлагая цену за работу и предоставляя часть общего решения совместно с другими агентами.

2. Качество данных

Одним из основных источников трудностей для методов анализа данных является качество данных. Данные могут содержать шум,

неполную информацию и избыточные, и бесполезные данные. Шумные, поврежденные и неполные данные могут вводить в заблуждение при поиске и усложнять анализ. Однако качество данных повышается с использованием электронного обмена, поскольку имеется меньше места для шума из-за электронного хранения, а не ручной обработки.

Методы анализа данных должны обеспечивать адекватный механизм для нахождения точных результатов из зашумленных данных. Методы анализа данных должны облегчать как подбор соответствующих данных, так и обучение с неполными знаниями.

Методы предварительной обработки данных должны применяться в данной ситуации. Процедура обеспечения качества данных должна быть эффективной, в противном случае это может привести к неправильной обработке данных. Методы оценки полезности предварительно обработанных данных важны. Эксперт в области также должен быть включен в процесс, если это возможно. Обычно этап предварительной обработки данных ориентирован на приложения, и его трудно извлечь из предыдущих исследований. Некоторые исследования показывают, что можно разработать инструменты предварительной обработки данных, которые можно настраивать и использовать в различных приложениях.

Другое решение заключается в интеграции технологии базы данных, такой как хранилище данных, которая обеспечивает возможность хранения данных (хорошего качества). Хранилище объединяет данные из нескольких и разнородных операционных источников и обрабатывает такие проблемы, как несоответствие данных, пропущенные значения перед сохранением подробных данных.

3. Формат данных

В течение последнего десятилетия формат данных, подлежащих анализу, значительно различался. Существует много видов данных, доступных для анализа, таких как реляционные, объектно-

ориентированные, текстовые, временные, пространственные, комбинаторные, веб, XML, мультимедиа.

Этот тип данных требует дополнительных шагов перед применением к традиционным моделям и алгоритмам ИАД, источник которых в основном ограничен структурированными или текстовыми, или числовыми данными. Этот дополнительный шаг включает преобразование расширенного формата данных в формат, подходящий для традиционных методов ИАД.

Например, данные, собранные из продвинутых приложений, таких как источники электронного бизнеса с веб-интерфейсом, являются полу структурированными и иерархическими, то есть данные не имеют заранее установленной абсолютной схемы, а извлеченная структура может быть нерегулярной или неполной.

Языки запросов могут использоваться для получения структурной информации из полу структурированных данных. На основе этой структурной информации генерируются данные, соответствующие традиционным методам ИАД. Языки веб-запросов, которые комбинируют выражения пути с синтаксисом в стиле SQL, являются хорошим выбором для извлечения структурной информации.

Формат данных может быть XML, JSON, так как предполагается, что через несколько лет они также будут наиболее широко используемым языком для представления документов. Предполагая, что метаданные хранятся в XML, интеграция двух разнородных источников данных становится намного более прозрачной, имена полей могут быть сопоставлены легче, а семантические конфликты могут быть описаны явно. В результате могут быть определены типы ввода и вывода данных из изученных моделей и детальная форма моделей. Более того, многие языки запросов, такие как XML-QL, XSL и XML-GL, разработаны специально для запросов XML и получения структурированной информации из этих

документов. Тем не менее, существуют серьезные проблемы, которые необходимо решить, например, как использовать извлеченную информацию об обобщенной структуре DTD при анализе данных, как использовать метаданные, хранящиеся в XML, при анализе данных, как заполнить недостающую информацию, если есть несоответствие в атрибутах.

Иногда большая часть данных организации находится не в простых числах и тексте, а в других средствах массовой информации, таких как изображения или аудио. Технология предварительной индексации и поиска изображений, звуковых файлов и видео должна использоваться для предварительной обработки данных этого типа. Эти технологии находятся в стадии разработки, но незрелые.

4. Адаптивность данных

Системы анализа данных должны быть адаптированы для работы с данными в реальном времени, в которые для анализа включаются новые данные транзакций, а также для включения новых моделей и алгоритмов анализа данных.

Системы анализа данных должны использовать преимущества вновь полученной информации, которая ранее отсутствовала на момент извлечения знаний, и сочетать ее с существующей моделью данных. Например, каждый раз, когда вводится новый продукт, компания должна изучить новый набор лучших практик.

В новых приложениях, таких как мультимедиа, XML и т.д., разрабатываются новые алгоритмы (или обновляются существующие алгоритмы) для работы с такими данными. Существующие системы анализа данных, которые включают методы для анализа простых пронумерованных данных, также должны быть достаточно гибкими, чтобы включать методы для анализа данных расширенного типа.

Решением может быть динамическое изменение анализируемой информации при изменении набора данных или включение обратной связи с пользователем для изменения действий, выполняемых системой. Агенты пользовательского интерфейса могут быть использованы для максимизации производительности взаимодействия текущих пользователей с системой путем адаптации поведения.

5. Представление знаний

Информация, полученная из производной модели данных, должна быть понятной, интерпретируемой для пользователей и, в конечном итоге, полезной при принятии решений. Взаимодействие между выходными данными процесса моделирования данных и инструментами представления должно быть прозрачным. Например, результаты анализа на основе нейронных сетей должны быть представлены пользователям в понятном формате, таком как символические правила, а не только в математических уравнениях.

Должны быть предприняты некоторые усилия для обеспечения стандартных интерфейсов прикладного программирования (API) для поддержки взаимодействия извлеченной базы знаний. Сообщество по анализу данных также должно быть связано с XML и связанными с ним протоколами и стандартами для стандартного представления.

6. Оценка знаний

Когда метод ИАД применяется к набору данных, он обычно дает несколько наборов результатов, особенно если в анализе используются методы перекрестной проверки и голосования. Возникает вопрос - какой из них сообщить или использовать при принятии решений.

Полученные знания должны оцениваться не только на основе точности. Должны быть предприняты некоторые меры, чтобы оценить, насколько эффективна, полезна, интересна и понятна или понятна

анализируемая информация. Поскольку меры по оценке эффективности зависят от задачи обучения, область применения должна играть главную роль в определении критериев оценки.

Выбор метода ИАД.

Приемлемость подходящего метода для набора данных должна основываться на определенных факторах, таких как характеристики набора данных, сильные и слабые стороны каждого рассматриваемого метода, наличие ресурсов, полезность заключенного результата и некоторые предыдущие, но недавние сравнительные исследования эффективность различных методов.

В последнее время, по мнению ряда исследователей, комбинация методов анализа данных (классификаторов) часто может давать более точные результаты, чем какой-либо отдельный метод (классификатор) в отдельности. Интеграция классификаторов может быть выполнена статически, как разделение проблемы на подзадачи, или динамически, с учетом переменных нового шаблона во время анализа.

Вывод

Анализ данных, как правило, представляет собой итеративный и интерактивный процесс, включающий постановку задачи, обеспечение качества данных, построение модели, интерпретацию и последующую обработку результатов.

Многие поставщики программного обеспечения и публикации прогнозируют, что все работники умственного труда станут аналитиками данных в будущем. Но все же сложные инструменты, такие как нейронные сети, деревья решений и визуализация данных, широко доступные для наивных пользователей, могут быть ошибкой. Системы анализа данных должны разрабатываться как интегрированные системы с учетом потребностей конечных пользователей, не являющихся технологами, -

скрывая детали базовых методов и обеспечивая эффективные и удобные для пользователя интерфейсы. Основное внимание следует уделять процессу в целом, а не отдельным компонентам в процессе анализа данных. Нам нужны системы анализа данных, которые включают в себя задачи предварительной обработки, задачи множественного обнаружения и задачи последующей обработки в их среде.

Для успеха методы анализа данных должны сочетаться с технологией управления данными для систематического сбора данных для систематического начала процесса, эффективной, но простой технологией пользовательского интерфейса, обеспечивающей представление анализируемых знаний. Например, файл данных, содержащий результаты запроса из хранилища данных, вводится в инструменты анализа данных. Результаты анализа данных идут через некоторые инструменты постобработки, такие как визуализация или обработчик правил, для значимых интерпретаций. Эти измененные результаты могут быть доступны через широкую интернет группу пользователей с помощью технологий клиент-сервер.

ГЛАВА 3. ЗАДАЧИ ПРЕДПРИЯТИЯ, РЕШАЕМЫЕ С ПОМОЩЬЮ ИНФОРМАЦИОННЫХ ХРАНИЛИЩ

3.1. Оперативная практическая аналитика

В современном мире данные являются активом предприятий, поскольку они используются для создания знаний и принятия решений. Это является основным сырьем 21 века.

Данные, доступные в различных формах, создаются машинами, которые оснащены электронными устройствами, такими как датчики движения, камеры, микрофоны, GPS, акселерометры, банкоматы, транзакции, мобильные телефоны, кредитные карты и другие электронные устройства. Данные также производится из социальных сетей, электронной почты, коллекции, клики по интернету, мобильные устройства для считывания штрих-кодов, датчики RFID и многие другие источники.

В сегодняшней быстро меняющейся, тесно взаимосвязанной глобальной бизнес-среде, предприятия полагаются на сложные системы, построенные из взаимосвязанных электронных датчиков и гаджетов, построенные вокруг промышленного интернета. Системы производят данные с большим объемом, скоростью и разнообразием. Генерируемые потоки данных разнообразны и несут много систем и бизнес информации, которые имеют решающее значение для успеха в бизнесе. Получение видимости в этом потоке и критические корпоративные операции, которые они поддерживают, могут означать разницу между успехом и неудачей.

Взаимосвязанные сложные ИТ-системы стали основными движущими силами систем промышленного производства, играющими на экономике во всем мире. Системы построены на машинах, системах хранения и расходных материалах, которые соответствуют определенному стандарту и связаны как

кибер-физические системы. С промышленным интернетом, также известным как Industry 4.0, системы начали революционизировать методы работы. Системы отвечают за оцифровку внутренних, горизонтальных и вертикальных цепочек добавленной стоимости и непосредственно отвечают за предоставление продуктов и услуг предприятиям. Это новое поколение систем 2-го поколения дало нам возможность управлять гранулированными потоками данных, открывая тем самым новый мир возможностей, ведущих к лучшему пониманию.

Потоки данных.

Потоковые данные отличаются от других видов данных из-за различных операционных атрибутов, которые к ним привязаны. Он часто слабо структурирован по сравнению с другими наборами данных. Например, количество потоков данных электронной почты, генерируемых в организационном контексте, достаточно велико и колеблется во времени. Использование биометрических устройств в организации позволяет лучше понять сродство и мобильность организации. Датчики трафика с коммутацией данных позволяют лучше понять сетевой трафик. Данные всегда доступны, и новые данные всегда генерируются. Время простоя для первичной системы сбора означает, что данные навсегда потеряны. Аналитика генерируется с подходящим сочетанием потоков данных, и, следовательно, сложность довольно высока. Будущие заводы и системы будут иметь четко определенные интерфейсы соединителей данных для превосходной видимости данных. Новые технологии обработки данных позволяют гибко заменять машины по всей цепочке создания стоимости. Industry 4.0 подчеркивает идею последовательной оцифровки и объединения всех производственных единиц в экономике. Информация об обработке событий, новой технологии, которая помогает получать практические ситуационные знания из крупномасштабных потоков событий в режиме реального времени, является интересной областью для новых приложений.

Физический и цифровой мир встречаются с очень высокой скоростью. Большая часть физического процесса оснащается датчиками, телематикой и устройствами, которые управляют постоянно растущими данными. Основной проблемой, с которой сталкивается современный мир, является затопление цифровых данных. Сбор, хранение и анализ данных с промышленных датчиков, сетевых журналов и другого оборудования, подключенного к Industry 4.0, стал еще более возможным благодаря появлению технологий больших данных. Hadoop - это ведущая платформа с открытым исходным кодом для масштабируемых, надежных и распределенных вычислений. Это меняет не только технологию, но и экономику хранения и хранения данных. Благодаря таким решениям, как Hadoop Distributed File System (HDFS), Hive, HBase, Pig, Oozie, Zoo-Keeper, Flume и т.д., Hadoop стала недорогой отраслевой стандартной экосистемой для безопасного анализа больших объемов данных из различных корпоративных источников. Простой обзор экосистемы (рис 3.1.):

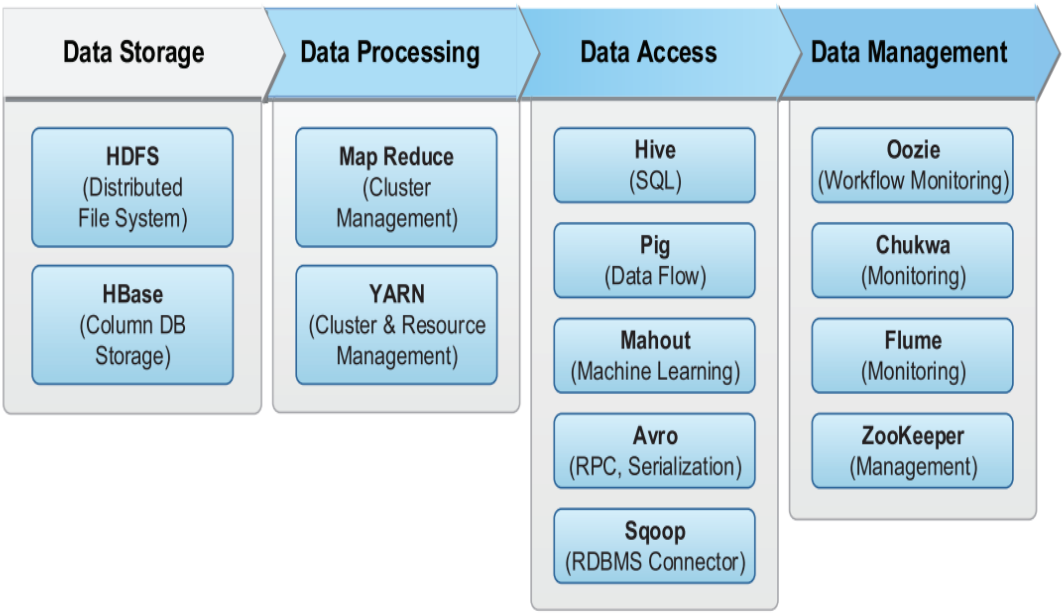

Рис. 3.1. Пример экосистемы

Аналитика.

Принятие решения или сбор информации для принятия решения не является простой и прямой задачей. Для получения информации или получения знаний с помощью анализа требуется подходящая комбинация потоков агрегированных данных в качестве сырья и аналитических инструментов. Как правило, аналитика относится к анализу данных с использованием анализа Парето, трендов, сезонности, регрессии, корреляции, контрольных диаграмм и других статистических методов. Это не всегда ряд чисел, отмеченных на индикаторе. Сегодня это комбинация потоков данных, которая повышает ценность организации. Когда аналитика интегрируется в бизнес и процессы принятия решений, понимание автоматически распространяется на тысячи работников умственного труда и тысячи решений, принимаемых каждый день людьми или компьютерами.

Анализ больших данных требует взаимосвязанного набора решений от сбора данных до принятия решений для повторного анализа. Раннее использование аналитики в организации было сосредоточено на разработке панелей данных, которые отображают информацию в графической форме. Такой подход позволяет руководителям легко выявлять тенденции и предпринимать действия, и это был важный шаг к более быстрому превращению информации в идеи. Согласно Википедии, «аналитика - это процесс разработки оптимальных или реалистичных рекомендаций по принятию решений на основе выводов, полученных путем применения статистических моделей и анализа, на основе существующих или смоделированных будущих данных». Когда аналитика применяется для повседневных операций, можно перейти к оперативной аналитике. Операционные транзакции имеют решения, и каждое действие определяется решением. Целью любого аналитического решения является предоставление организации действенных идей для принятия более разумных решений и улучшения бизнес-результатов.

Оперативная аналитика.

Big Data и прогнозируемая аналитика используются бизнесом и отраслями для изменения и улучшения процессов и операций в основном. Industry 4.0 соединяет различные датчики уровня производства в цепочке создания стоимости. Оперативная аналитика позволяет предприятиям анализировать и анализировать сгенерированные машиной данные датчиков или производственные данные, чтобы в реальном времени получать информацию о работе предприятия. Потоковая передача данных позволяет проводить анализ в реальном времени. Вместо того, чтобы запрашивать статические данные, потоки данных в реальном времени постоянно оцениваются статическими вопросами. Анализ данных в режиме реального времени дает многочисленные преимущества.

Оперативная аналитика автоматизирует аналитику, помогая конечным пользователям (или системам) во время самого процесса принятия решений, что приводит к оперативному интеллекту. Решения, будь то тактические или стратегические, имеют решающее значение для успеха каждой организации.

Основой оперативной аналитики, является повышенная доступность и комплексное использование соответствующих данных путем объединения всех продуктов, ресурсов и предприятий, участвующих в цепочке создания стоимости. Он включает в себя возможность генерировать дополнительную ценность из имеющихся данных и максимизировать выгоды для клиентов. Это требует фундаментальной трансформации процессов, портфеля продуктов и услуг, а также существующих бизнес-моделей. Детальное представление о системах и устройствах с аналитикой и аналитическими данными помогает предприятиям повысить эффективность своей работы и сократить расходы.

IT Оперативная аналитика (ITOA) также известен как расширенная оперативная аналитика или IT аналитика даных. Он включает в себя технологии, которые в основном используются для обнаружения сложных шаблонов в больших объемах «шумных» данных доступности и

производительности ИТ-системы. Существуют различные аналитические информационные панели:

1. Операционные информационные панели: помогают непосредственным работникам и руководителям отслеживать основные операционные процессы.

2. Тактические информационные панели: позволяют менеджерам, аналитикам отслеживать, анализировать деятельность, процессы и проекты департамента.

3. Стратегические информационные панели: позволяет руководителям и сотрудникам отслеживать прогресс в достижении своих стратегических целей.

Оперативно аналитическое приложение - это многослойное приложение, построенное на инфраструктуре бизнес-аналитики и интеграции данных, позволяющее предприятиям более эффективно измерять, отслеживать и контролировать эффективность бизнеса. Они сосредоточены на функциях мониторинга больше, чем на функциях анализа или управления. Тактические панели инструментов подчеркивают аналитические функции больше, чем функции мониторинга или управления. Аналитическая функциональность позволяет пользователям исследовать коренные причины проблем, проблем или тенденций. Стратегические панели инструментов подчеркивают функции управления больше, чем функции мониторинга или анализа. Операционная аналитика применяется к организациям всех размеров в разных отраслях по всей длине цепочки поставок. Понимание поведения клиентов, повышение их опыта и прибыльности одинаково важно независимо от размера клиентской базы. В Оперативной аналитике может выполняться множество комбинаций транзакционных данных, таких как агрегации продуктов или агрегации клиентов, которые географически близки. Описательные модели используются для определения прогноза спроса, затрат на производство и распределение и взаимосвязей затрат, будущих затрат на сырье и их

доступности, а также ряда других параметров и взаимосвязей, требуемых базой данных оперативных решений.

Оперативная практическая аналитика использует технологии больших данных для последних приложений, чтобы анализировать машинные данные и получать информацию, которая дает лучшие бизнес-результаты. Данные, сгенерированные машинами, собранные ИТ-системами, содержат ценную информацию. ИТ-операционная аналитика автоматизирует процесс сбора и организации данных для поиска шаблонов, которые помогают идентифицировать бизнес-результаты и повысить производительность системы. На текущем рынке существует несколько программных решений для оперативной аналитики, которые помогают легко преобразовать отключенные данные (которые находятся в разрозненных системах) в действенную информацию.

3.2. Решение проблем анализа прогнозирования и планирования

Существует широко распространенный мнения о том, что большие данные могут помочь в улучшении прогнозов при условии, что есть возможность делать анализ и обнаружить скрытые закономерности, и согласно то, что прогнозы можно усовершенствовать благодаря принятию решений на основе информаций.

Большие данные наиболее востребованы для построения прогностических моделей в мире, где прогнозы продолжают оставаться важной статистической проблемой. Затем подходит вопрос, в чем заключается проблема прогнозирования? Самое простое объяснение состоит в том, что обычные инструменты прогнозирования не могут делать обработку, скорость и сложность, присущие большим данным. Это связано с отсутствием

структуры в этих наборах данных и размера. В результате обычные методы редко предпочитаются для решения больших данных.

Возможности получения прибыли от прогнозирования с помощью больших данных разнообразны. В настоящее время расширяются исследования использования больших данных для получения точных прогнозов погоды, и первоначальные результаты свидетельствуют о том, что большие данные значительно улучшат прогнозы погоды.

На самом деле, прогнозирование погоды было одним из главных преимуществ больших данных, но прогнозы все еще являются неточными после недельного периода. Авиационная отрасль является еще одной областью, где прогнозирование больших данных имеет решающее значение.

В этом разделе внимание сосредоточено в основном на проблемах, которые необходимо преодолеть при прогнозировании с помощью больших данных. Необходимо отметить, что наличие больших данных само по себе не является концом проблем. Хорошим примером является наличие огромного количества данных о землетрясениях, но отсутствие надежной модели, способной точно прогнозировать землетрясения.

Существует большая угроза ложных открытий из больших данных. Это связано с тем, что, хотя получение прогнозов с использованием соответствующего метода является основной проблемой, это не совсем так. Учитывая огромное количество данных, которые необходимо обработать и спрогнозировать, с большими данными становится все сложнее различать случайность и статистически значимые результаты. Таким образом, повышается вероятность сообщения о случайности как статистически значимого результата и введения в заблуждение заинтересованных сторон, заинтересованных в прогнозе.

Исследователи в области экономики были основными разработчиками больших данных для прогнозирования различных экономических переменных. Использовав динамическую факторную модель, основанную на

методологии, для прогнозирования большого набора данных, включающего испанские диффузионные индексы, которые описывают как исчерпывающее описание испанской экономики. Модели данные являются расширением факторных моделей и часто используются для прогнозирования с помощью больших данных.

Применение оценок максимального правдоподобия факторных моделей для прогнозирования больших данных было оценено в ходе имитационного исследования, где авторы считают этот подход является эффективным и действенным. Сезонная модель AR использовалась, чтобы показать, как большие данные из поисковой системы Google могут использоваться для прогнозирования экономических показателей.

Большие данные, относящиеся к различным обменным курсам, используются для прогноза курса валют, британского фунта и японской иены, предлагаемая модель с коррекцией ошибок с улучшенным коэффициентом опережает модель VAR с коэффициентом дополнением прогнозирование трех основных двусторонних обменных курсов.

Далее сгруппируем приложения Data Mining и статистических методов для прогнозирования с помощью больших данных в области экономики по темам, основанным на использовании больших данных для прогнозирования экономических переменных.

Большие данные будут расти еще больше в предстоящие годы, и если организации не будут склонны и готовы принять вызовы, развивать и применять обязательные навыки, они окажутся в тяжелом положении. В этом обзоре, который сфокусирован на прогнозировании с помощью больших данных, первоначально было определено несколько проблем и обрисовали потенциал, который большие данные могут предложить и генерировать прибыльные результаты при условии, что было посвящено достаточно времени и усилий для преодоления выявленных проблем. После этого следует

отметить ряд ключевых проблем, которые в настоящее время мешают и мешают точности и эффективности прогнозов больших данных.

С точки зрения применения статистических методов и методов интеллектуального анализа данных для прогнозирования с использованием больших данных, основанных на прошлой литературе, очевидно, что факторные модели являются наиболее распространенным и популярным инструментом, используемым в настоящее время для прогнозирования больших данных, в то время как нейронные сети и байесовские модели являются двумя другими популярными выбор. В обзоре также указывается, что область экономики является наиболее популярной областью с точки зрения использования больших данных для прогнозирования переменных, представляющих интерес, при этом тематикой ВВП и денежно-кредитной политики уделяется основное внимание. По данным опубликованных исследований, области народонаселения и энергетики являются вторыми и третьими по популярности. Очевидно, что по-прежнему существуют широкие возможности для исследований в области прогнозирования с использованием больших данных и что такая работа может дать более совершенные методы, которые могут повысить точность прогнозирования.

Этап планирования содержит в себе решение задач выбора стратегий реализаций и метода ее создания, анализа проблем, в случае чего разрабатывается хранилище данных, также информации для решения проблем предприятия, использование той или иной архитектуры хранилищ, определение стоимости, планируется разработка, включая набор метаданных.

Цель этого этапа определить задачи на ХД, выбрать способ решения этих проблем, определить программный и технологический объект и узнать, как, в какие сроки и за какие деньги этот объект реализуется.

Выбор стратегии реализации определяет подход, который используется при создании ХД.

Как правило, используются следующие подходы: сверху вниз, снизу вверх, середина и комбинированный подход, который в последнее время становится все более популярным. Подход «сверху вниз» выбран для вновь созданного ХД, т.е. Если «с нуля», все решения относительно технологического внедрения системы принимаются.

Восходящий подход используется, когда уже существует определенная вычислительная среда и объекты, из которых можно создать новый объект.

Подход «из центра» включает в себя эволюционное, поэтапное создание объекта, когда впервые разрабатывается так называемое ядро объекта, которое на следующих этапах создается с новыми функциями. Комбинированный подход использует комбинацию вышеуказанных подходов.

Стратегия реализации определяет, какая концептуальная архитектура используется и как создается в зависимости от порядка реализации выбранной концептуальной архитектуры.

Выбор методологии создания хранилищ данных образно определяет язык проекта, на котором говорят члены команды проекта, как создается техническая документация и какие принципы разработки используются. Это может быть метод структурного анализа и проектирования, спиральный метод, методы, основанные на использовании UML и т. д.

Методология разработки основана на использовании концепции трех схем, типичных для базы данных, и включает методы анализа спецификаций, а также методы концептуального, логического и физического проектирования.

Метод структурного анализа и проектирования является хорошо разработанным методом и основан на использовании стандартов и процедур IDEF. Метод создания по спирали реализует концепцию эволюционного подхода к созданию системы. Использование методов объектно-ориентированного анализа для реляционных баз данных требует преобразования результирующей схемы объекта из базы данных несущей в

реляционную схему. Анализируя задачи хранилищ данных предполагается идентификация данных предприятия, информацию о которой будет содержана в базе.

Определение тематической ориентации является одной из важнейших задач на этом этапе. Внутри каждого фрагмента предметной области определяется:

- Количество и тип источников данных (организационные единицы);
- Количество выбранных источников данных;
- Данные хранятся на ХД;
- Цель использования данных;
- реализовано ли хранилище данных на существующей аппаратной и программной платформе или на платформе, аналогичной существующей.

3.3. Оценка эффективности внедрения хранилищ данных.

Существует множество вариантов реализации внедрения платформы хранилища данных. Несмотря на то, что их оценка не должна быть сложным процессом, принятие соответствующих мер поможет гарантировать, что средства инвестируются в лучшие технологии, возможные для конкретных потребностей для организации.

В том же числе существует несколько типов платформ хранилищ данных, а также различные варианты развертывания и различные варианты использования хранилищ данных. Как только решается вопрос инвестирования в платформу хранилища данных, следующим шагом будет создание процесса оценки доступных продуктов, а затем поиск того, который наилучшим образом соответствует требованиям, которые являются благополучными и эффективными для предприятия.

Для этого сначала следует определить важные функции платформы, которые обеспечивают эффективную разработку хранилища данных. Затем появляется возможность точно определить вариант развертывания, который наилучшим образом соответствует требованиям организации.

Теперь перейдем к основному моменту, в данном случае о особенности эффективной разработки хранилища данных.

Изучая следующие функции платформы хранилища данных, надо помнить, что требования к реализации и использованию будут определять наиболее важные аспекты. Не каждый проект хранилища данных требует всех обсуждаемых функций. Оценивая конкретные продукты, можно использовать следующее, чтобы углубиться в конкретные функции, поддерживаемые каждым поставщиком. С учетом этого предостережения хранилище данных предоставляет следующие ключевые возможности:

Он предлагает последовательное представление данных. Для эффективной поддержки приложений бизнес-аналитики (BI) для анализа и составления отчетов о прошлых бизнес-операциях платформа хранилища данных должна быть способна извлекать данные из нескольких исходных систем и делать их похожими на единый пул информации. Данные, необходимые для использования BI, извлекаются из операционных систем и обычно преобразуются для обеспечения их согласованности, а затем загружаются в хранилище данных для анализа.

Это позволяет организации моделировать и создавать проекты баз данных для хранилищ данных. Общим требованием к хранилищу данных является частичная денормализация схемы базы данных для оптимизации запросов и аналитической производительности. В отличие от этого, онлайновые системы обработки транзакций обычно используют полностью нормализованные схемы, чтобы гарантировать согласованность и целостность данных.

На практике это означает, что хранилище данных обычно спроектировано на основе многомерной модели с центральным интересным фактом и несколькими измерениями, по которым этот факт анализируется.

Например, предположим, что мы заинтересованы в анализе продаж компании - есть много аспектов, которые могут формировать, как эта информация анализируется, таких как продукт, территория, магазин и время. В простом случае мы можем запросить общий объем продаж (факт) по продукту (измерению) за январь 2020 года (измерение) на юго-восточной территории (измерение).

Чтобы включить такой анализ, хранилища данных используют многомерные модели, известные как схема «звезда» и «снежинка». Для схемы типа «звезда» несколько таблиц измерений связаны с одной таблицей фактов с использованием отношения «один ко многим». Схема «снежинка» аналогична схеме «звезда», но измерения можно хранить в нескольких нормализованных таблицах вместо одной таблицы измерений. При рассмотрении платформ хранилищ данных обязательно следует учитывать их встроенную поддержку проектирования баз данных со схемами типа «звезда» и «снежинка» и оптимизации запросов.

Он поддерживает функции OLAP, позволяющие хранилищу данных обрабатывать BI-запросы. Примеры интерактивных функций аналитической обработки включают в себя возможность детализации, свертывания, поворота и ранжирования данных. Преимущество функций OLAP заключается в том, что они позволяют разработчикам и конечным пользователям кодировать менее сложные запросы. Кроме того, функции OLAP обычно превосходят более сложные запросы, выполняющие те же задачи. Если возможности запросов и SQL на платформе хранилища данных не поддерживают встроенные функции OLAP, вам, вероятно, потребуется приобрести дополнительные инструменты запросов, которые предлагают такие функции.

Это обеспечивает ключевую производительность и оптимизацию запросов. Будучи аналитической платформой, хранилище данных предъявляет различные требования для оптимизации запросов из операционной или транзакционной системы управления базами данных (СУБД). Функции, полезные для максимизации производительности хранилища данных, включают поддержку оптимизации соединения по звездам, растровые индексы и карты зон.

Способность оптимизировать запрос типа «звезда», в котором таблица фактов объединена с рядом различных таблиц измерений, является важной функцией платформы хранилища данных. Но каждая платформа реализует соединения звезд по-разному.

Например, хотя растровые индексы полезны для оптимизации объединений, их поддержка варьируется от продукта к продукту. Некоторые платформы допускают явное создание индекса растрового изображения, в то время как другие генерируют растровое изображение как часть процесса оптимизации типа «звезда».

Поддержка самых больших хранилищ данных требует гибкого подхода с индивидуальным оборудованием и программным обеспечением.

Еще одна функция, связанная с производительностью, - это поддержка карт зон.

Зона - это набор непрерывных блоков данных или страниц на диске. Карта зон - это структура базы данных, в которой хранится информация о данных, хранящихся в табличных зонах. Используя карту зоны, запросы можно оптимизировать, обрезав блоки данных, которые не могут помочь ответить на запрос, поэтому к ним нет доступа.

Функциональность в памяти. Используя память вместо диска для хранения и обработки данных, можно улучшить производительность.

Варианты включают использование СУБД в памяти или использование платформы хранилища данных, которая предоставляет функции в памяти.

Возможности перемещения данных.

Хранилище данных отделено от операционных систем баз данных, которые выполняют ежедневные бизнес-транзакции. Таким образом, данные должны регулярно перемещаться из одной среды в другую. Существует несколько методов и технологий для перемещения данных, в том числе:

Простая загрузка и выгрузка утилит.

Функциональность ETL для извлечения, преобразования и загрузки данных.

Технология репликации, которая собирает измененные данные из исходных баз данных и отправляет только изменения в целевое хранилище данных.

Все эти технологии перемещения данных могут быть получены отдельно от платформы хранилища данных. Действительно, если есть глубокие требования к сложным преобразованиям или высокоскоростной репликации, вероятно, лучшим вариантом будет дополнительный инструмент, поскольку он обычно предоставляет больше возможностей и более высокую функциональность. Конечно, многие платформы хранилищ данных имеют встроенные возможности перемещения данных, которые могут удовлетворить потребности передачи данных.

При расширении или расширении существующего хранилища данных обычно лучше всего использовать текущую платформу, а не усложнять ее путем преобразования в другое. Использование одной или двух новых функций, таких как улучшенная оптимизация или новые функции OLAP, может помочь избежать преждевременного отказа от существующей платформы для блестящего нового устройства хранилища данных или облачной службы. Однако переход на новую платформу хранилища данных

может иметь смысл, если бизнес-потребности организации и требования BI значительно изменились со времени развертывания хранилища данных.

Для организаций, которые хотят переложить поддержку, а также развертывание, DWaaS является лучшим вариантом, потому что архитектура хранилища данных поддерживается поставщиком в облаке. Кроме того, если организация выполняет большую часть транзакций обработки данных в облаке, то DWaaS может быть лучшим вариантом. Хранение данных, которые генерируются и хранятся в облаке, для хранения данных - логичный подход.

Подход гибридной транзакции / аналитической обработки (HTAP) становится все более популярным, поскольку единую платформу можно использовать для нескольких целей. HTAP может минимизировать кривую обучения, уменьшая количество различных технологий, которые необходимо освоить. Конечно, некоторое количество обучения все еще необходимо, так как этот подход обычно требует дополнительных новых технологий и опций. Крупные и средние организации, стремящиеся сократить количество поддерживаемых технологий и повысить гибкость, должны рассмотреть подход HTAP.

Крупнейшие на сегодняшний день хранилища данных содержат более 10 петабайт необработанных данных. Поддержка самых больших хранилищ данных требует гибкого подхода с индивидуальным оборудованием и программным обеспечением. Обычно это означает комбинацию программного обеспечения СУБД и СУБД, работающего на самых быстрых серверах, устройствах хранения и сетевых устройствах.

Если необходимо интегрировать аналитику больших данных с требованиями BI в хранилище данных, следует рассмотреть подходы, которые предоставляют хранилища данных Polyglot. Термин полиглот заимствован из движения базы данных NoSQL, которое поддерживает постоянство полиглота, что означает, что данные хранятся в СУБД наиболее подходящего типа для предполагаемого использования.

В среде хранилищ данных, которая позволяет управлять традиционными данными BI и получать к ним доступ наряду с новыми типами больших данных, подход с использованием полиглота включает несколько типов платформ данных. Они варьируются от реляционных и аналитических баз данных до СУБД NoSQL и новых платформ, таких как Spark и Hadoop. Несмотря на то, что это добавляет сложности, оно также предоставляет пользователям хранилища данных возможность объединить историческую BI с более перспективной прогностической аналитикой и интеллектуальным анализом данных

ЗАКЛЮЧЕНИЕ

Определенные термины базы данных, хранилища данных и бизнес-аналитики показывают, как каждый из них имеет отношение к другому. В то время как база данных представляет собой набор информации, хранилище данных - это все базы данных, объединенные для удовлетворения потребностей своей организации.

Предприятия - это все меняющиеся и растущие организации, и системы, которые их поддерживают, должны быть соответствующим образом усовершенствованы. Организации привыкли полагаться на дорогие серверы, которые часто оказывались единственной точкой отказа для бизнеса. Данные - это жизненно важный организационный ресурс, которым нужно управлять, как и другими важными активами. Предприятия хотят эффективно использовать информацию, чтобы получить конкурентное преимущество для принятия более эффективных решений, которые улучшают и оптимизируют бизнес-процессы, делая прогноз рынка более точным.

Хранилище данных - ведущая и самая надежная технология, используемая сегодня компаниями для планирования, прогнозирования и управления, например, для планирование ресурсов, финансовое прогнозирование и контроль и т. д. После эволюции концепции хранилища данных в начале 90-х годов считалось, что эта технология будет развиваться очень быстрыми темпами, но, к сожалению, не реальность. Много было сделано в этой области в отношении дизайна и развитие хранилищ данных и многое еще предстоит сделать, но одна область, которая требует особого внимания со стороны исследовательского сообщества, это данные обслуживание склада.

Основная причина неудач проекта хранилища данных – плохая услуга поддержания. Без надлежащего технического обслуживания желаемые результаты почти невозможно получить из хранилища данных. В отличие от

операционных систем хранилища данных нуждаются в гораздо большем обслуживании и команде поддержки квалифицированные специалисты необходимы для решения возникающих проблем после его развертывания, включая извлечение данных, загрузку данных, сеть управление, обучение и связь, управление запросами и некоторые другие связанные задачи.

Для выполнения всех этих функций и процессов требуется квалифицированная команда штатных квалифицированных специалистов, которые могут эффективно и постоянно заботиться обо всем хранилище данных вопросы технического обслуживания своевременно.

В то время как гибридная архитектура создает платформу управления данными, которая облегчает некоторые проблемы управления данными ИТ-отдела. Вопрос всегда будет: есть ли прямая выгода для бизнеса? Но в то же время гибридная архитектура создает очень аполитичный стабильный слой, из которого построить хранилище размерных данных в соответствии с требованиями бизнеса.

ЛИТЕРАТУРА

На русском языке

1. С. Я. Архипенков, Д. В. Голубев, О. Б. Максименко “Хранилища данных”, 2010.
2. Майкл Армстронг, Дарлен Армстронг-Смит “Разработка специальных запросов и анализ данных”, 2008.
3. Валентина Д.К. “Автоматизированное рабочее место экономиста”, 2009.
4. Т. А. Гаврилова, В. Ф. Хорошевский “Базы знаний интеллектуальных систем”, 2012.
5. В. В. Корнеев, А. Ф. Гареев, С. В. Васютин, В. В. Райх “Базы данных. Интеллектуальная обработка информации”, 2013.
6. Омельченко, В.П., Демидова, А.А.“Информатика”, 2013.

На английском языке

7. Bill I. “Building the data warehouse”, 2014
8. Lawrence Corr “Agile Data Warehouse Design”, 2016.
9. Vincent Rainardi “Building a Data Warehouse: With Examples in SQL Server”, 2007.
10. Rolan B. “Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho”, 2009.
11. Marianne B. “Modern ERP: Select, Implement, and Use Today's Advanced Business Systems”, 2015

Интернет ресурсы

12. http://www.eiminstitute.org/resource-portals/data-warehousing/data-warehouse-goals-and-objectives-part-1/
13. https://www.herzing.edu/blog/what-data-warehousing-and-why-it-important
14. https://www.researchgate.net/publication/228752147_Efficiency_and_effectiveness_of_data_warehousing_a_case_study
15. https://tdan.com/how-to-do-a-data-warehouse-assessment-and-why/4873
16. https://dl.acm.org/doi/abs/10.1145/3018009.3018056
17. https://revolution.allbest.ru/marketing/00485972_0.html
18. https://www.coursehero.com/file/p5mfbvn/CONCLUSION-Databases-data-warehouse-and-business-intelligence-terms-once/
19. http://dwprojectmanagement.blogspot.com/2009/03/conclusion.html
20. https://studfile.net/preview/5240037/page:53/
21. https://shodhganga.inflibnet.ac.in/bitstream/10603/49/10/chapter%207_%20s%20s%20%20reddy.pdf
22. https://www.tandfonline.com/doi/abs/10.1080/17509653.2015.1113394?src=recsys&journalCode=tmse20
23. https://www.prj-exp.ru/dwh/dwh_roi
24. https://www.element61.be/en/topics/dwh-data-warehousing-modeling
25. https://www.slideshare.net/virathin/data-warehousing-fundamentals-for-it-professionals
26. https://epdf.pub/data-warehousing-fundamentals-for-it-professionals.html
27. https://anuradhasrinivas.files.wordpress.com/2013/03/data-warehousing-fundamentals-by-paulraj-ponniah.pdf
28. https://www.javatpoint.com/data-warehouse-design

29. https://www.xplenty.com/blog/the-ultimate-guide-to-data-warehouse-design/

30. https://www.1keydata.com/datawarehousing/processes.html

31. https://www.guru99.com/data-warehouse-architecture.html

32. https://www.itprotoday.com/sql-server/7-steps-data-warehousing

33. https://www.researchgate.net/publication/27473559_A_Data_Warehouse_Architecture_for_Clinical_Data_Warehousing

РЕЗЮМЕ

Научно-исследовательская работа говорит о важности хранения данных, исследований, управления предприятием. Изучаются принципы построения информационных технологий для решения проблем на предприятиях, в том числе организации автоматизированных рабочих мест. Были затронуты вопросы, связанные с получением качественных результатов с использованием базы данных при принятии решений по информационной поддержке и маркетингу в корпоративных системах.

Также были рассмотрены проблемы компьютерного моделирования технологий для решения маркетинговых задач, прогнозирования области конкуренции в маркетинговой системе предприятий. Подробно обсуждались ситуационный анализ на предприятиях, а также поиск подходящего программного пакета для решения программных задач хранения данных. Обсуждались многомерные модели, оценка эффективности хранилища данных на основе OLAP и принципы создания информационно-аналитической базы в этой системе.

XÜLASƏ

Tədqiqat işi məlumat anbarının, tədqiqatın, müəssisə rəhbərliyinin əhəmiyyətindən danışır. Müəssisələrdə problemlərin həlli üçün, o cümlədən avtomatlaşdırılmış iş yerinin təşkili üçün informasiya texnologiyalarının qurulması prinsipləri öyrənilir. Müəssisə sistemlərində məlumat dəstəyi və marketinq qərarları qəbul edərkən verilənlər bazasından istifadə edərək yüksək keyfiyyətli nəticələr əldə etməklə bağlı məsələlər qaldırıldı.

Marketinq problemlərinin həlli üçün texnologiyanın kompüter modelləşdirilməsi, müəssisələrin marketinq sistemində rəqabət sahəsinin proqnozlaşdırılması problemləri də nəzərdən keçirildi. Müəssisələrdə situasiya təhlili, habelə məlumat anbar proqramları problemlərinin həlli üçün uyğun bir proqram paketinin axtarışı, axtarışı barədə ətraflı müzakirə edildi. Çoxölçülü modellər, OLAP əsasında məlumat anbarının səmərəliliyinin qiymətləndirilməsi və bu sistemdə informasiya-analitik bazanın yaradılması prinsipləri müzakirə edildi.

SUMMARY

The research work speaks about the importance of data warehouse, research, enterprise management. The principles of building information technologies for solving problems at enterprises, including the organization of an automated workplace, are studied. Issues related to information support in enterprise systems and obtaining high-quality results using a database when making marketing decisions were raised.

The problems of computer modeling of technology for solving marketing problems, forecasting the competitive sphere in the marketing system of enterprises were also considered. It was discussed in detail about the search, the selection of a suitable software package for solving the problems of situational analysis of enterprises, as well as data warehouse programs. The principles of evaluating the effectiveness of the data warehouse based on multidimensional models, OLAP and the creation of an information-analytical base in this system were discussed.