Содержание:

ВВЕДЕНИЕ

Интеллектуальный анализ данных — это обработка информации и выявление в ней моделей и тенденций, которые помогают принимать решения. Принципы интеллектуального анализа данных известны в течение многих лет, но с появлением больших данных они получили еще более широкое распространение.

Большие данные привели к взрывному росту популярности более широких методов интеллектуального анализа данных, отчасти потому, что информации стало гораздо больше, и она по самой своей природе и содержанию становится более разнообразной и обширной. При работе с большими наборами данных уже недостаточно относительно простой и прямолинейной статистики. Имея 30 или 40 миллионов подробных записей о покупках, недостаточно знать, что два миллиона из них сделаны в одном и том же месте. Чтобы лучше удовлетворить потребности покупателей, необходимо понять, принадлежат ли эти два миллиона к определенной возрастной группе, и знать их средний заработок.

Процесс анализа данных, поиска и построения модели часто является итеративным, так как нужно разыскать и выявить различные сведения, которые можно извлечь. Необходимо также понимать, как связать, преобразовать и объединить их с другими данными для получения результата. После обнаружения новых элементов и аспектов данных подход к выявлению источников и форматов данных с последующим сопоставлением этой информации с заданным результатом может измениться.

1.1 Интеллектуальный анализ данных как процесс

Для решения бизнес-задач требуется такой анализ данных, который позволяет построить модель для описания информации и в конечном итоге приводит к созданию результирующего отчета. Этот процесс иллюстрирует рисунок 1.

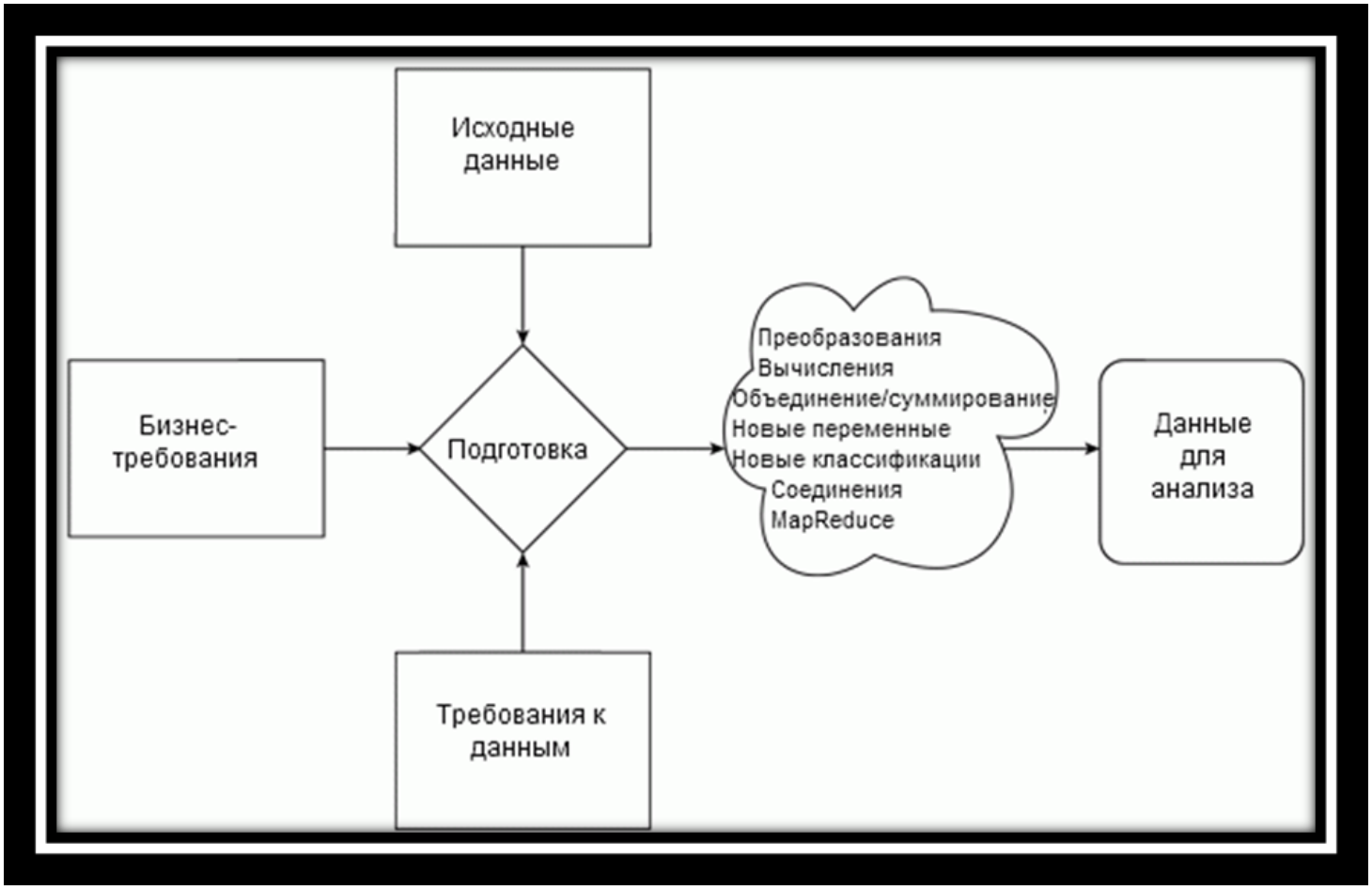

Рисунок 1. Бизнес - требования

Процесс анализа данных, поиска и построения модели часто является итеративным, так как нужно разыскать и выявить различные сведения, которые можно извлечь. Необходимо также понимать, как связать, преобразовать и объединить их с другими данными для получения результата. После обнаружения новых элементов и аспектов данных подход к выявлению источников и форматов данных с последующим сопоставлением этой информации с заданным результатом может измениться.

1.2 Инструменты интеллектуального анализа данных

Интеллектуальный анализ данных — это не только используемые инструменты или программное обеспечение баз данных. Интеллектуальный анализ данных можно выполнить с относительно скромными системами баз данных и простыми инструментами, включая создание своих собственных, или с использованием

готовых пакетов программного обеспечения. Сложный интеллектуальный анализ данных опирается на прошлый опыт и алгоритмы, определенные с помощью существующего программного обеспечения и пакетов, причем с различными методами ассоциируются разные специализированные инструменты.

Например, IBM SPSS®, позволяет строить эффективные прогностические модели по прошлым тенденциям и давать точные прогнозы. IBM InfoSphere® Warehouse обеспечивает в одном пакете поиск источников данных, предварительную обработку и интеллектуальный анализ, позволяя извлекать информацию из исходной базы прямо в итоговый отчет.

Рисунок 2. IBM SPSS®

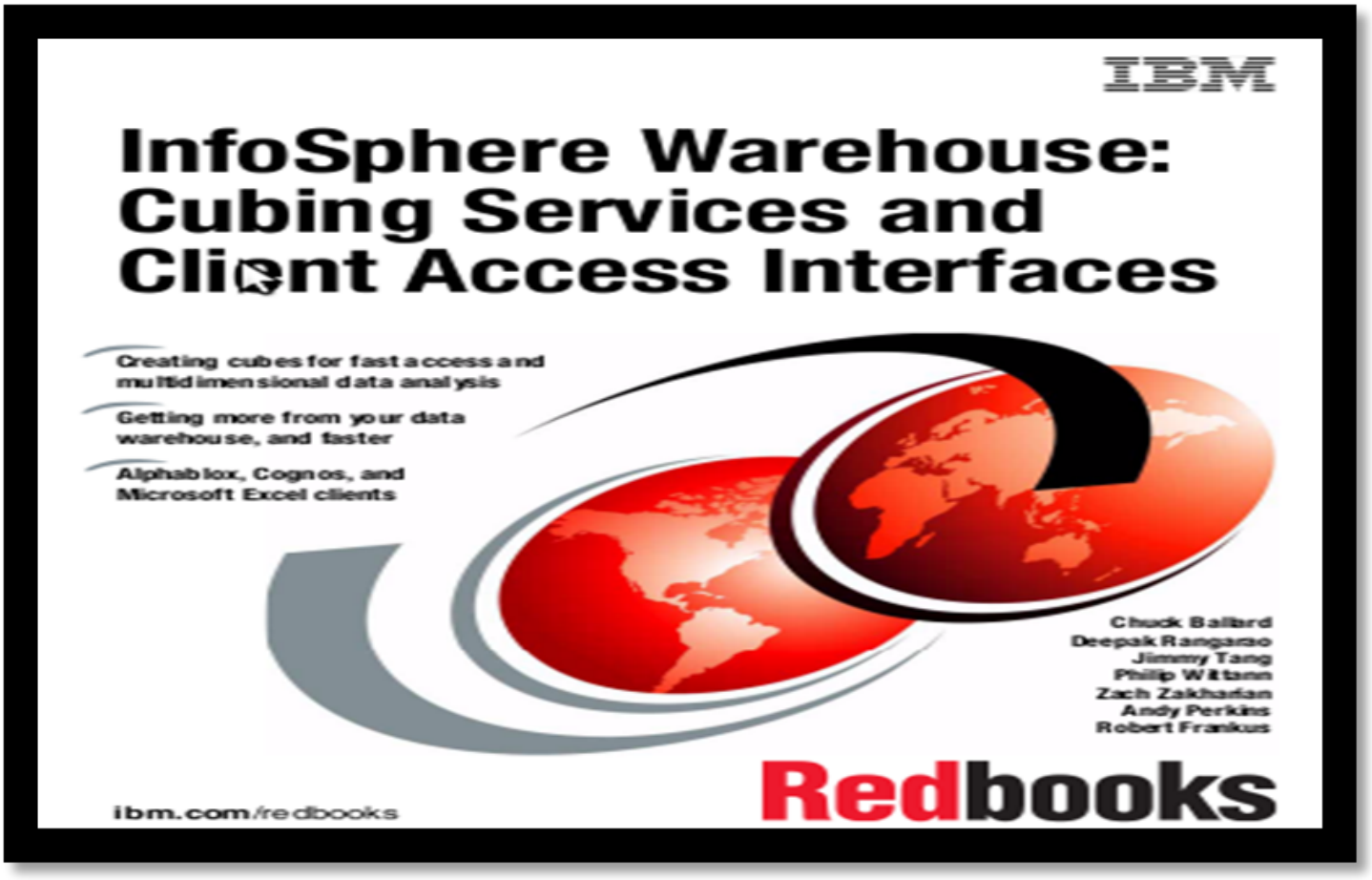

Рисунок 3. IBM InfoSphere® Warehouse

В последнее время стала возможна работа с очень большими наборами данных и кластерная/крупномасштабная обработка данных, что позволяет делать еще более сложные обобщения результатов интеллектуального анализа данных по группам и сопоставлениям данных. Сегодня доступен совершенно новый спектр инструментов и систем, включая комбинированные системы хранения и обработки данных.

Можно анализировать самые разные наборы данных, включая традиционные базы данных SQL, необработанные текстовые данные, наборы "ключ/значение" и документальные базы. Кластерные базы данных, такие как Hadoop, Cassandra, CouchDB и Couchbase Server, хранят и предоставляют доступ к данным такими способами, которые не соответствуют традиционной табличной структуре.

В частности, более гибкий формат хранения базы документов придает обработке информации новую направленность и усложняет ее. Базы данных SQL строго регламентируют структуру и жестко придерживаются схемы, что упрощает запросы к ним и анализ данных с известными форматом и структурой.

Рисунок 4. Кластерные базы данных

1.3 Ассоциация

Ассоциация (или отношение), вероятно, наиболее известный, знакомый и простой метод интеллектуального анализа данных. Для выявления моделей делается простое сопоставление двух или более элементов, часто одного и того же типа. Например, отслеживая привычки покупки, можно заметить, что вместе с клубникой обычно покупают сливки.

Создать инструменты интеллектуального анализа данных на базе ассоциаций или отношений нетрудно. Например, в InfoSphere Warehouse есть мастер, который выдает конфигурации информационных потоков для создания ассоциаций, исследуя источник входной информации, базис принятия решений и выходную

информацию.

Рисунок 5. Информационный поток, используемый при подходе ассоциации

1.4 Классификация

Классификацию можно использовать для получения представления о типе покупателей, товаров или объектов, описывая несколько атрибутов для идентификации определенного класса. Например, автомобили легко классифицировать по типу (седан, внедорожник, кабриолет), определив различные атрибуты (количество мест, форма кузова, ведущие колеса). Изучая новый автомобиль, можно отнести его к определенному классу, сравнивая атрибуты с известным определением. Те же принципы можно применить и к покупателям, например, классифицируя их по возрасту и социальной группе.

Кроме того, классификацию можно использовать в качестве входных данных для других методов. Например, для определения классификации можно применять

деревья принятия решений. Кластеризация позволяет использовать общие атрибуты различных классификаций в целях выявления кластеров.

1.5 Прогнозирование

Прогнозирование — это широкая тема, которая простирается от предсказания отказов компонентов оборудования до выявления мошенничества и даже прогнозирования прибыли компании. В сочетании с другими методами интеллектуального анализа данных прогнозирование предполагает анализ тенденций, классификацию, сопоставление с моделью и отношения. Анализируя прошлые события или экземпляры, можно предсказывать будущее.

Например, используя данные по авторизации кредитных карт, можно объединить анализ дерева решений прошлых транзакций человека с классификацией и сопоставлением с историческими моделями в целях выявления мошеннических транзакций. Если покупка авиабилетов в США совпадает с транзакциями в США, то вполне вероятно, что эти транзакции подлинны.

1.6 Последовательные модели

Последовательные модели, которые часто используются для анализа долгосрочных данных, — полезный метод выявления тенденций, или регулярных повторений подобных событий. Например, по данным о покупателях можно определить, что в разное время года они покупают определенные наборы продуктов. По этой информации приложение прогнозирования покупательской корзины, основываясь на частоте и истории покупок, может автоматически предположить, что в корзину будут добавлены те или иные продукты.

1.7 Деревья решений

Дерево решений, связанное с большинством других методов (главным образом, классификации и прогнозирования), можно использовать либо в рамках критериев отбора, либо для поддержки выбора определенных данных в рамках общей структуры. Дерево решений начинают с простого вопроса, который имеет два ответа (иногда больше). Каждый ответ приводит к следующему вопросу, помогая

классифицировать и идентифицировать данные или делать прогнозы.

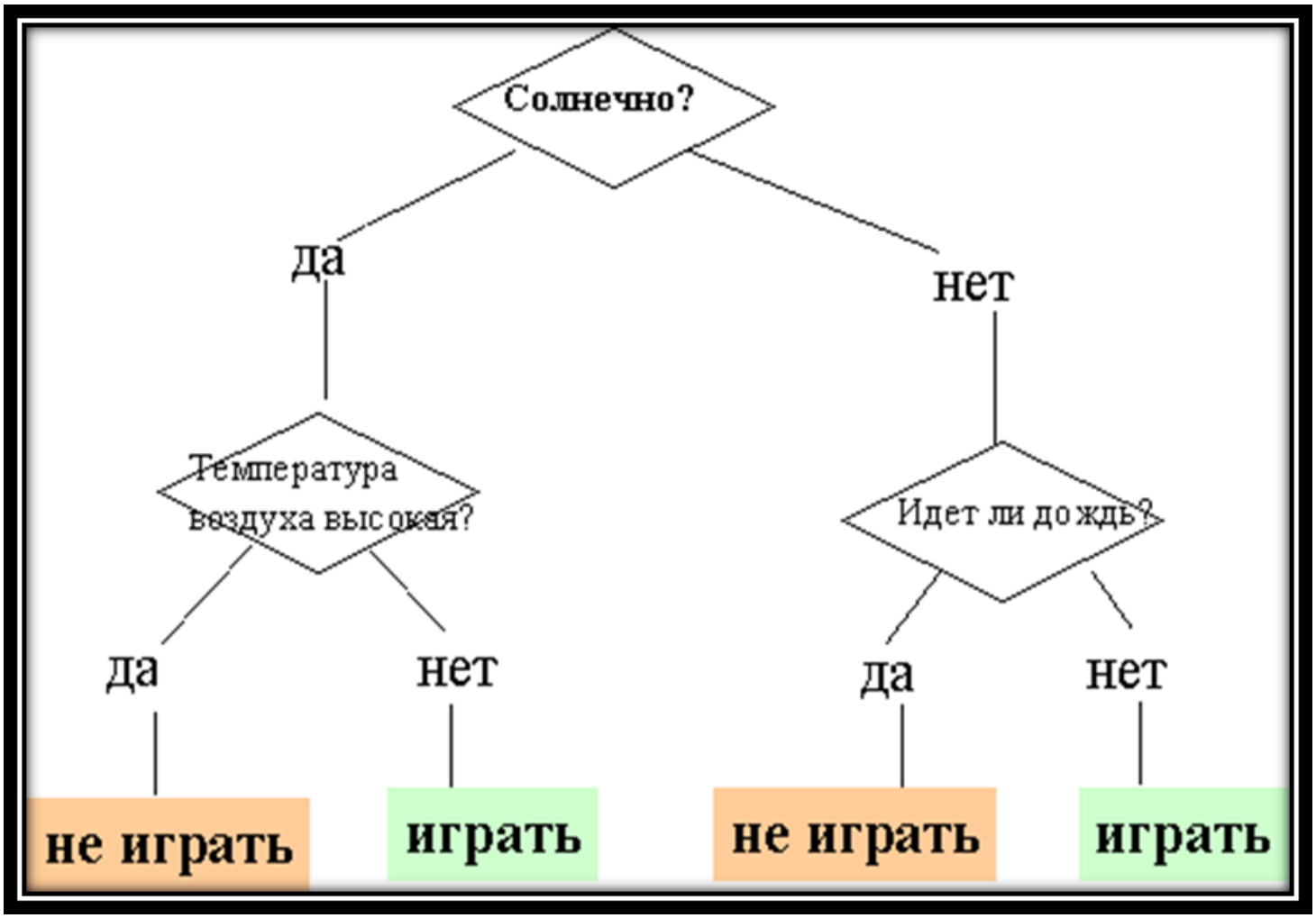

Рисунок 6. Дерево решений

Деревья решений часто используются с системами классификации информации о свойствах и с системами прогнозирования, где различные прогнозы могут основываться на прошлом историческом опыте, который помогает построить структуру дерева решений и получить результат.

1.8 Обработка с запоминанием

При всех основных методах часто имеет смысл записывать и впоследствии изучать полученную информацию. Для некоторых методов это совершенно очевидно. Например, при построении последовательных моделей и обучении в целях прогнозирования анализируются исторические данные из разных источников и экземпляров информации.

В других случаях этот процесс может быть более ярко выраженным. Деревья решений редко строятся один раз и никогда не забываются. При выявлении новой информации, событий и точек данных может понадобиться построение дополнительных ветвей или даже совершенно новых деревьев.

Некоторые из этих процессов можно автоматизировать. Например, построение прогностической модели для выявления мошенничества с кредитными картами сводится к определению вероятностей, которые можно использовать для текущей транзакции, с последующим обновлением этой модели при добавлении новых (подтвержденных) транзакций. Затем эта информация регистрируется, так что в следующий раз решение можно будет принять быстрее.

1.9 Специальный формат анализа данных

При работе с InfoSphere Warehouse создание поведенческо-демографической модели в целях анализа данных о покупателях для понимания моделей их поведения предусматривает использование исходных данных SQL, основанных на информации о транзакциях, и известных параметров покупателей с организацией этой информации в заранее определенную табличную структуру. Затем InfoSphere Warehouse может использовать эту информацию для интеллектуального анализа данных методом кластеризации и классификации с целью получения нужного результата.

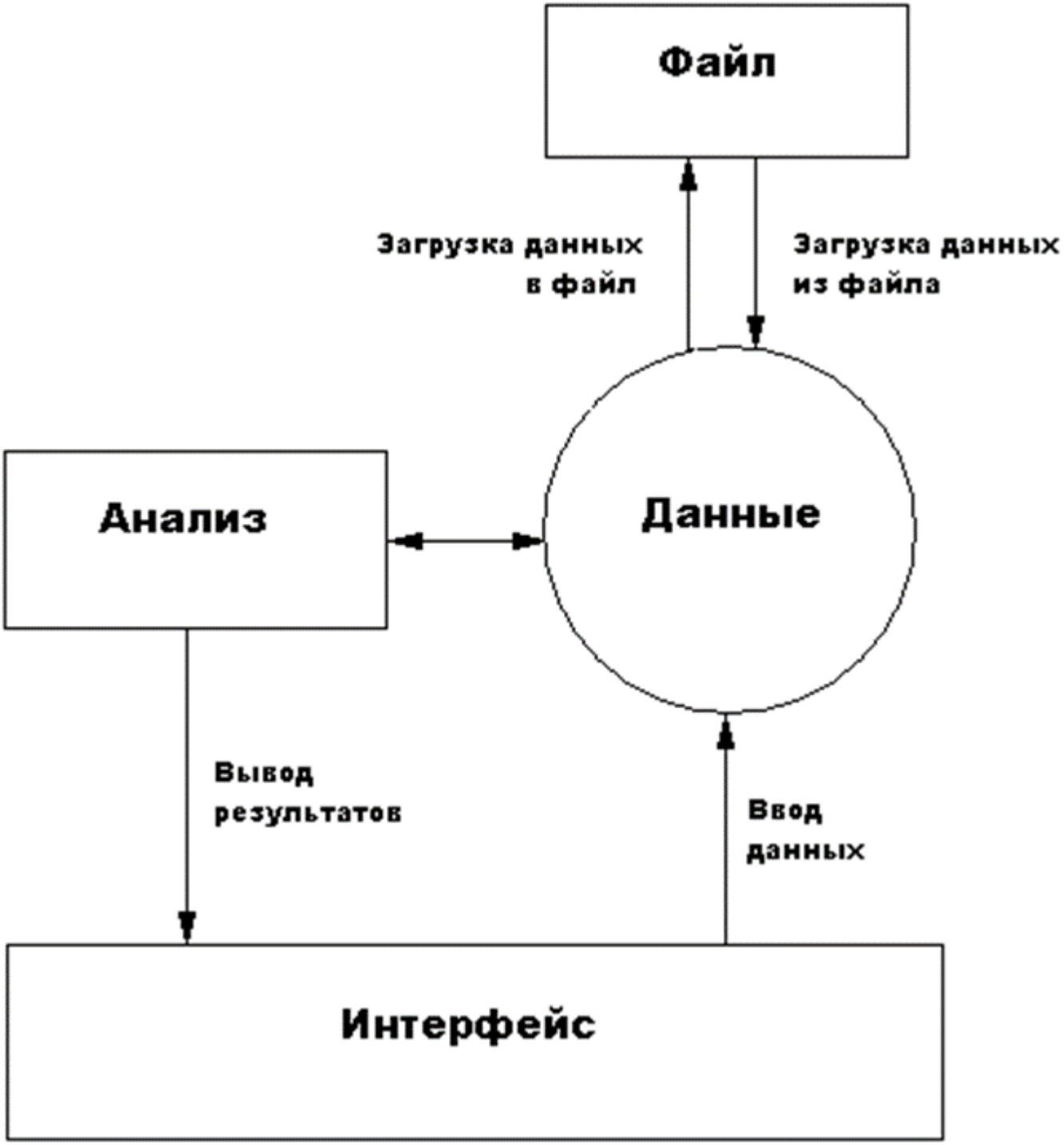

Рисунок 7. Специальный формат анализа данных

Например, по данным о продажах можно выявить тенденции продаж конкретных товаров. Исходные данные о продажах отдельных товаров можно преобразовать в информацию о транзакциях, в которой идентификаторы покупателей сопоставляются с данными транзакций и кодами товаров. Используя эту информацию, легко выявить последовательности и отношения для отдельных

товаров и отдельных покупателей с течением времени. Это позволяет InfoSphere Warehouse вычислять последовательную информацию, определяя, например, когда покупатель, скорее всего, снова приобретет тот же товар.

Из исходных данных можно создавать новые точки анализа данных. Например, можно развернуть (или доработать) информацию о товаре путем сопоставления или классификации отдельных товаров в более широких группах, а затем проанализировать данные для этих групп, вместо отдельных покупателей.

1.10 Документальные базы данных и MapReduce

Обработка с помощью функции MapReduce многих современных документальных и NoSQL баз данных, таких как Hadoop, нацелена на очень большие наборы данных и информацию, которая не всегда соответствует табличному формату. При работе с программным обеспечением интеллектуального анализа данных эта система может принести пользу — и вызвать проблемы.

Основная проблема данных на основе документов — это неструктурированный формат, который может потребовать дополнительной обработки. Много различных записей могут содержать аналогичные данные. Сбор и согласование этой информации в целях упрощения ее обработки зависит от этапов подготовки и применения MapReduce.

В системе, основанной на MapReduce, на этапе преобразования исходные данные нормализуются — приводятся к стандартной форме. Этот шаг может быть относительно простым (определение ключевых полей или точек данных) или сложным (анализ и обработка информации для создания выборки данных). В процессе преобразования данные приводятся к стандартизированному формату, который можно использовать в качестве базы.

Сокращение — это суммирование или количественная оценка информации с последующим выводом этой информации в стандартизованную структуру, основанную на итогах, суммах, статистике или других результатах анализа, выбранных для вывода данных.

Запросы к этим данным часто бывают сложными — даже при использовании специализированных инструментов. Идеальный подход к интеллектуальному анализу данных заключается в использовании этапа MapReduce в рамках

подготовки данных.

Например, при выполнении интеллектуального анализа данных методом ассоциации или кластеризации на первом этапе лучше всего построить подходящую статистическую модель, которую впоследствии можно будет применять для выявления и извлечения необходимой информации.

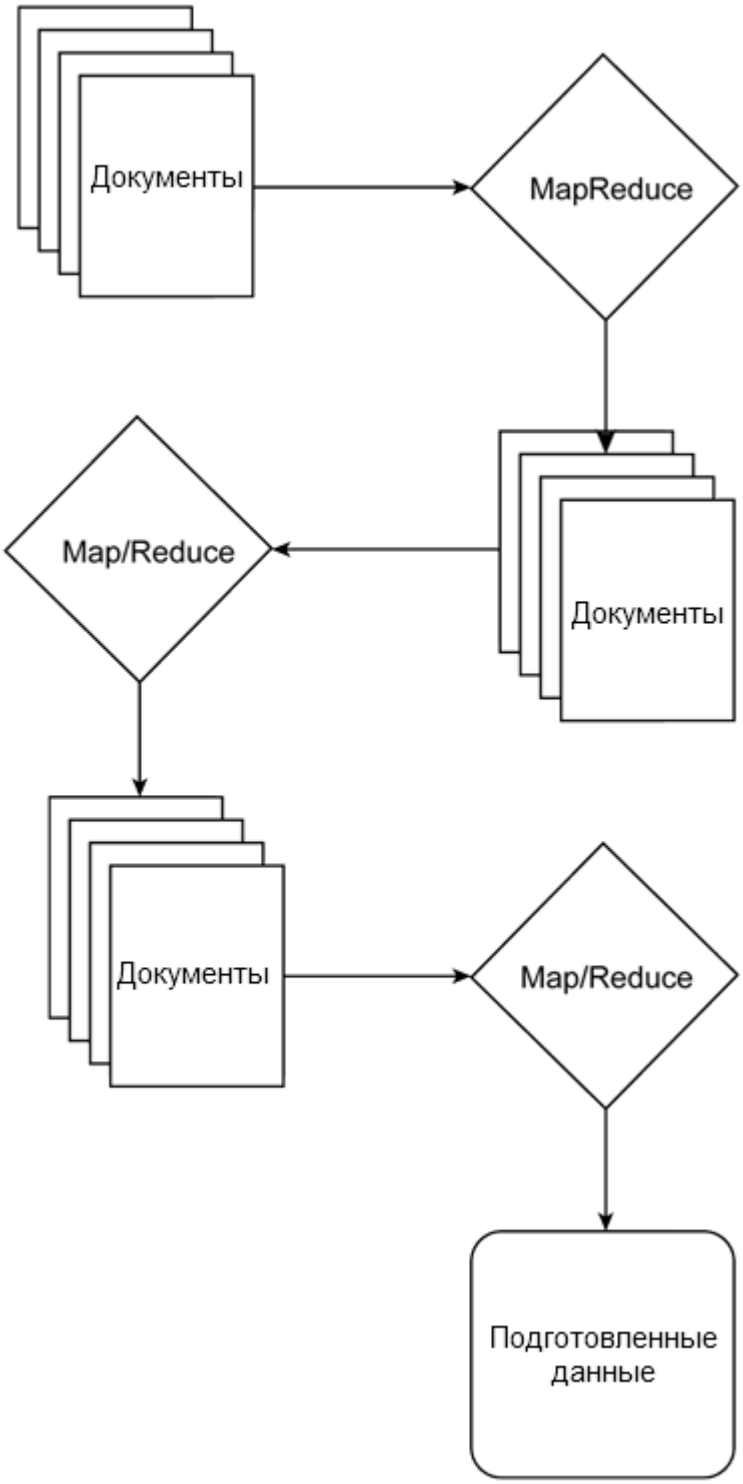

Рисунок 8. Структура MapReduce

Независимо от исходных данных, многие инструменты могут использовать неструктурированные файлы, CSV или другие источники данных. Например, InfoSphere Warehouse в дополнение к прямой связи с хранилищем данных DB2 может анализировать неструктурированные файлы.

ЗАКЛЮЧЕНИЕ

Интеллектуальный анализ данных — это не только выполнение некоторых сложных запросов к данным, хранящимся в базе данных. Независимо от того, используете ли вы SQL, базы данных на основе документов, такие как Hadoop, или простые неструктурированные файлы, необходимо работать с данными, форматировать или реструктурировать их. Требуется определить формат информации, на котором будет основываться ваш метод и анализ. Затем, когда информация находится в нужном формате, можно применять различные методы (по отдельности или в совокупности), не зависящие от требуемой базовой структуры данных или набора данных.