Происхождение данных в масштабах крупного предприятия

П.И. Заваленова

***Аннотация* — В настоящей статье рассматриваются методы происхождения данных, которые на текущий момент начинают применяться в рамках крупномасштабных предприятий. Установление родословной, контроль качества данных, оптимизация и интеллектуальное изменение потоков данных, рационализация сложных сред и эффективный надзор за крупномасштабными изменениями могут быть уязвимы для ошибок и неточностей, возникающих из-за отсутствия знания о том, как данные фактически проходят через экосистему от начала до конца. Линия данных помогает отслеживать движение информационных потоков в экосистеме всего предприятия. Прежде всего рассматриваются проблемы деятельности организации в рамках отсутствия какого-либо автоматизированного подхода к происхождению информации. В результате чего обосновывается необходимость соответствующего программного обеспечения, которое позволило бы наглядно продемонстрировать поток данных от системы-источника в систему-приемник.**

В статье рассматриваются ручные и автоматизированные подходы к передаче данных, после чего определяются их основные недостатки. В результате анализа предлагается наиболее приемлемый способ выявления происхождения данных – гибридный подход, который использует человеческий вклад, продвинутые алгоритмы и автоматизированные методы проверки и обнаружения потоков данных, которые в совокупности делают происхождение данных управляемым и эффективным процессом. Описываются преимущества предлагаемого метода и способы реализации программного обеспечения в рамках деятельности предприятий.

Полученные результаты представляются в виде концептуальных моделей происхождения данных. В работе продемонстрировано, из чего состоит современная система, позволяющая определять и наглядно представлять родословную информации.

***Ключевые слова*— управление данными, происхождение данных, хранилище данных, информация, цепочка данных.**

Статья получена 12 июня 2021

Статья представляет собой результат магистерской диссертации на тему «Создание концептуальной архитектуры системы формирования модели данных и цепочек происхождения данных для проектирования и сооружения АЭС».

Заваленова П. И. – магистрант, Национальный исследовательский ядерный университет «МИФИ» (e-mail: polina.zavalenova@gmail.com)

.

I. Введение

Малоэффективное использование данных, хранящихся в огромном количестве информационных систем на крупномасштабных предприятиях, может привести к большим временным, трудовым и денежным потерям. Описание данных, обрабатываемых в информационных системах, часто представляются в разных форматах. Эти описания могут не актуализироваться на протяжении эксплуатации информационных систем. Рано или поздно перед каждой организацией встает вопрос, как правильно систематизировать, обрабатывать, безопасно хранить и искать необходимую информацию.

Решение многих бизнес-задач – от соответствия нормативным требованиям, проверки качества данных, анализа влияния сведений на конечные отчеты, рационализации экосистем и других форм обновления данных – в решающей степени зависит от способности ответить на два фундаментальных вопроса о корпоративной информации: откуда появились данные? Какой путь был выполнен? Ответы на поставленные вопросы – это то, что предоставляет происхождение данных. Цепочка происхождения в этом свете является фундаментальной основой для удовлетворения множества насущных потребностей и юридических требований крупных предприятий с обширными экосистемами данных.

Линия передачи данных отслеживает движение информации в экосистеме компании. В крупных организациях эта экосистема может включать сотни приложений и систем хранения, а также десятки соединений, через которые происходит обмен данными между системами. При отсутствии первоисточника данных нет возможности для создания всеобъемлющего представления о перемещении информации, существует только множество «локальных» представлений владельцев приложений, которые в какой-либо степени знакомы с деталями входящих и исходящих каналов конкретной системы.

Простое агрегирование «локальных» представлений, выраженных, например, большим набором диаграмм потоков данных, которые ориентированы на приложения, не может обеспечить достаточную

поддержку для ключевых вариантов использования, таких как следующие:

• Согласованность: ключевым элементом к соответствию данных является объяснимость, то есть возможность отслеживать элементы данных до точек их происхождения. Цепочки зависимостей должны быть обнаружены и визуализированы для возможности идентификации критических сегментов потоков, оказывающих влияние на результат, который подлежит нормативным требованиям.

• Качество данных: обнаружение и визуализация того, как потоки данных объединяются, разделяются и преобразуются приложениями, является ключом к идентификации всех процессов, влияющих на качество информации или от которых зависит ее ухудшение.

• Рационализация и рефакторинг: обновление архитектур данных, устранение неэффективности и преобразование унаследованных приложений по своей сути являются сложными процессами, требующими изменения схемы потоков данных – нарушение одних существующих потоков и создание других. Прозрачность потока данных является ключевым фактором как на этапе планирования, так и на стадии реализации этих процессов.

• Эволюция архитектуры данных: среды хранения информации претерпевают крупные изменения, обусловленные циклами разработки приложений, инициативами по обеспечению соответствия, слияниями и т.д. Оценка осуществимости, планирование, выполнение и контроль проекта зависят от детального знания потоков данных.

Все вышеперечисленное определяет актуальность создания средств по автоматизации процесса описания и обновления модели данных в информационных системах и формирования цепочек происхождения данных. В настоящей работе представлен подход к происхождению информации в рамках крупномасштабных предприятий.

II. Область проблемы происхождения данных

Определение происхождения данных очень простое, но предлагает широкий спектр возможных интерпретаций. Согласно [1], происхождение данных – это поток данных от первоисточника до пункта назначения, дающий ответы на следующие вопросы:

• Как перемещаются данные между таблицами, представлениями или отчетами?

• Когда данные были загружены, обновлены или рассчитаны в конкретном столбце, таблице, представлении или отчете?

• Какие компоненты (отчеты, запросы и структуры) будут затронуты при изменении других компонентов?

• Какие данные, структура или отчет кем и когда используются?

• Какова стоимость внесения изменений?

• Что сломается, когда будут внесены изменения?

При передаче данных движение информации прослеживается через среду данных от точки ее происхождения до точки конечного потребления в экосистеме. Эта трассировка построена в терминах бизнес-концепций таким образом, чтобы можно было понять и визуализировать происхождение отдельных элементов данных конкретных логических типов. Следовательно, он выходит за рамки более локального анализа совокупности процессов управления хранилищами данных или потоков сообщений (перемещение информации), а также простого сравнения значений исходного и целевого полей в целях проверки целостности (согласованности информации).

Формирование цепочки происхождения данных, контроль качества, оптимизация и интеллектуальное изменение потоков данных, рационализация сложных сред и эффективный надзор за крупномасштабными изменениями – уязвимы для ошибок и неточностей, возникающих из-за отсутствия сведений о том, как информация фактически продвигается внутри экосистемы от первоисточника до конечной точки. Следовательно, первое измерение проблемы: определение фактического существования и направленности всех потоков данных между узлами или приложениями через экосистему.

Однако для того, чтобы всеобъемлющее отображение потока данных было применяемым для бизнес-целей, оно должно быть логично оформлено. Недостаточно понять и визуализировать поток данных в физических терминах или просто знать, какие столбцы таблиц представлены в записях, передаваемых из одной системы в другую. Необходимо обнаружить и разобрать цепочки зависимостей данных, сформулированных в терминах основных бизнес-концепций – ключевых классов сущностей (клиент, продукт, учетная запись пользователя, подписка и т. д.), составляющих бизнес-данные для конкретного предприятия. Иными словами, необходимо отслеживать и визуализировать не просто движение данных, но и поток информации, относящейся к конкретной основной бизнес-концепции. Поставленная задача невыполнима без привязки столбцов к концепциям. Таким образом, точная классификация данных на уровне объектов и атрибутов является вторым аспектом проблемы.

Какими бы всеобъемлющими не были высокоуровневые представления о движении информации, недостаток глубины делает эти понятия практически бесполезными для принятия подробных решений и объяснения на основе фактов, что является третьим аспектом проблемы. Представление о потоке информации, обеспечиваемом линией передачи данных, должно быть не только всеобъемлющим, но и достаточно детальным – вплоть до уровней объектов и атрибутов.

Очевидно, что происхождение данных необходимо вычислять объективно на логическом и физическом уровнях, на правильном уровне детализации, прежде чем его можно будет эффективно использовать для решения описанных выше вариантов использования. Учитывая эти три аспекта проблемы, можно выделить

основные подходы, прояснить их недостатки, а также отметить оптимальный путь к решению поставленных задач.

На сегодняшний день ключевой подход к происхождению данных основан на ручном процессе: взаимодействие с владельцами приложений, сопоставление результирующих инвентаризаций входящих и исходящих потоков данных для каждого приложения, а затем ручное создание всеобъемлющей картины потока данных внутри предприятия. Однако ручной подход к передаче данных неадекватен по ряду причин, рассмотренных далее. При этом полностью автоматизированный подход к формированию цепочек происхождения данных является не только не соответствующим, но и трудноразрешимый из-за внутренней сложности проблемы.

В настоящий момент для крупномасштабных предприятий предлагается гибридный подход, использующий человеческий вклад, продвинутые алгоритмы и автоматизированные методы проверки, обнаружения потоков информации, в совокупности обеспечивающие управляемость и эффективность происхождения данных.

III. ПОДХОДЫ К ПРОИСХОЖДЕНИЮ ДАННЫХ

Полностью ручные подходы к передаче данных подвержены ряду серьезных недостатков.

• Возникновение ошибок. Исходные данные для полностью ручного подхода в основном состоят из сопоставленных ответов владельцев приложений на запросы информации о потоках, которые поступают в их приложения и исходят из них. Многие владельцы приложений хорошо знакомы с деталями собственных документов, хранящихся в информационных системах, однако ответы на поставленные вопросы все еще могут содержать неточности, непреднамеренные упущения, невольное использование неполной или устаревшей информации и догадок.

• Классификация данных не масштабируется с человеческой точки зрения. Сотрудники могут определить систематизацию бизнес-концепций и указать, какие принципы могут быть задействованы в потоке данных в конкретное приложение и из него. Однако стержнем логического описания потока данных является классификация информации, которая связывает отдельные атрибуты с концепциями. Классификация данных должна быть автоматизирована, поскольку количество столбцов базы данных по всем приложениям может исчисляться миллионами или десятками миллионов для крупного предприятия.

• Ограничение высокоуровневой информацией. Сопоставление ответов от владельцев приложений может обеспечить высокоуровневое описание потока данных, но не позволяет детализировать уровнь объектов и атрибутов, необходимых для строгого отслеживания данных с точки зрения бизнес-концепций.

• Отсутствие объективной проверки. В заключении, утверждения людей о происхождении и потоке данных – это просто утверждения людей. Мысли, подкрепленные доказательствами, вычисленными на основе необходимых данных, с большей вероятностью удовлетворят регулирующие органы и гораздо более безопасны в качестве основы для принятия решений, чем заверения на бумаге, не имеющие объективной проверки.

Изложенные недостатков ручного процесса выявления родословной данных ставит новый вопрос: можно ли полностью автоматизировать передачу информации?

Как бы привлекательно это ни звучало, но полностью автоматизированный подход к передаче данных оказывается непрактичным. Человеческий интеллект превосходит высокоуровневое описание и анализ. Без участия человека полностью автоматизированный подход к происхождению данных должен осуществляться методом исчерпания: попытка обнаружить поток данных и зависимости путем проверки наличия всех возможных связей между приложениями в экосистеме.

Помимо универсального доступа, который потребуется для выполнения этого метода использования, сложность исчерпывающего поиска сегментов потока между тысячами приложений становится недопустимым с вычислительной точки зрения. На рис. 1 изображена диаграмма, включающая в общей сложности семь информационных систем и двенадцати сегментов потока.

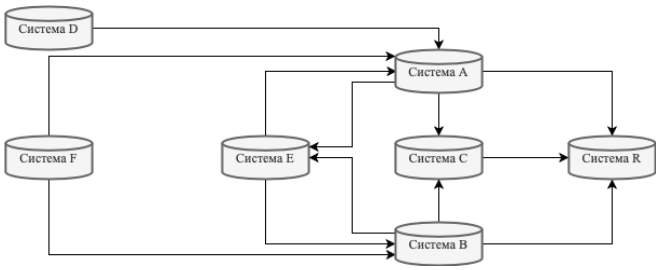

Рис. 1. Диаграмма потока данных для среды, состоящей из семи систем.

Следует обратить внимание, что каждый сегмент потока является направленным [2]. Например, сегмент, соединяющий систему D и систему A, является исходящим потоком данных из D и входящим в A. Системы E и B соединены двумя сегментами потока: одним потоком из E в B, а другой течет из B в E.

Чисто автоматизированный подход к обнаружению потока данных в экосистеме, состоящей из семи приложений, должен будет проверять наличие соединения в любом направлении для всех различных пар систем. Другими словами, он должен будет проверить все возможные сегменты потока, чтобы обнаружить 12 фактических сегментов. Применяя аналогичные вычисления, как в приведенном выше примере, экосистема данных с тысячами приложений может содержать миллион возможных сегментов потока. Таким образом, сложность потокового вычисления растет пропорционально квадрату количества задействованных систем.

Лучшей стратегией передачи данных является использование сильных сторон человеческого и машинного интеллектов в гибридном подходе.

IV. ГИБРИДНЫЙ ПОДХОД ДЛЯ КРУПНОМАСШТАБНЫХ ПРЕДПРИЯТИЙ

В предложенном гибридном подходе человеческий интеллект выполняет две важные функции.

1) Сужение проблемы потокового поиска. Человеческий интеллект, как было сказано ранее, превосходит высокоуровневое описание и анализ. Несмотря на подверженность ошибкам и отсутствие детализации, огромное преимущество человеческого интеллекта здесь заключается в сужении всевозможных потоков между системами, улучшая проблему масштабирования, рассмотренную выше.

Хотя количество возможных потоков пропорционально квадрату количества систем, на самом деле существует управляемое количество проверяемых потоков из каждой системы. Опросы и интервью с владельцами приложений служат, в первую очередь для того, чтобы собрать воедино общую архитектуру потока данных, что позволяет исключить большую часть бесконечных возможностей.

Собранные вместе высокоуровневые человеческие описания потока данных служат в гибридном подходе в качестве набора гипотез потока, которые затем подвергаются машинной проверке.

2) Логическое описание данных предприятия с точки зрения бизнес-концепций. Вторая важная функция человеческого интеллекта в этом контексте – определить набор основных бизнес-концепций и указать их присутствие в ландшафте данных.

Почему это важно? Бизнес-концепции являются ключами для интерпретации логического значения сегментов потока и их бизнес-контекста (например, как конкретный сегмент может быть квалифицирован по географическому признаку или направлению деятельности). Они позволяют понять, как логическое значение и контекст меняются по мере того, как приложения объединяют, разделяют и трансформируют данные от начала до конца в своем путешествии по экосистеме. Бизнес-концепции определяют набор объектов и типов атрибутов, моделируемых программным обеспечением, которое выполняет автоматическую классификацию данных.

Принимая во внимание гипотезы о потоках данных человека и логическое описание бизнес-концепций предприятия, машина теперь вносит свой вклад в гибридный подход. Это имеет ряд аспектов.

1) Автоматическая масштабируемая классификация данных (связывание столбцов с бизнес-концепциями). В то время, как сотрудник может просматривать наборы столбцов и определять связи с логическими концепциями, ручная классификация миллионов или десятков миллионов столбцов явно неосуществима. Однако решения по классификации нескольких классов можно смоделировать с помощью алгоритмов машинного обучения. Как уже было отмечено, логическая классификация данных в соответствии с бизнес-концепциями является стержнем осмысленного и действенного, а также всеобъемлющего взгляда на движение информации.

2) Проверка сегментов потока данных. Человеческий вклад сократил вселенную возможностей потока до управляемого набора. Первой важной функцией машины является проверка того, что эти гипотетические сегменты потока между приложениями и системами действительно существуют (наряду с гипотетической направленностью потока), и что семантика машинно-классифицированных данных, передаваемых в этих сегментах, соответствует человеческому описанию. Последовательная выборка записей в сочетании с автоматической классификацией данных может обнаруживать объекты в таблицах, которые попадают под определенные бизнес-концепции, тем самым проверяя точки потока для этой концепции через экосистему. Сам поток данных можно проверить с помощью ряда машинных методов.

При анализе меток времени выбираются записи в источнике и месте назначения, столбцы которых сопоставлены с определенным семантическим доменом или бизнес-концепцией (в соответствии с классификацией данных) и содержат информацию о метке времени, такую, как «Дата / время последнего обновления». Затем программа ищет пересечения этих записей в двух системах и упорядочивает их в соответствии с информацией о временных метках. Шаблоны последовательности, показывающие аналогичные временные интервалы, указывают как на существование, так и на направление потока между источником и местом назначения.

Однако информация о временных метках часто отсутствует и может быть ненадежной (иногда «метки времени» просто копируются из исходной системы в целевую). В таких случаях для проверки потока данных можно использовать стробоскопический анализ.

Стробоскопический анализ использует дифференциальную выборку записей через равные промежутки времени для наблюдения за потоком и определения его направленности. Программное обеспечение выполняет последовательную выборку записей как источника, так и места назначения примерно в одно и то же время, используя критерии выбора столбцов, аналогичные критериям анализа временных меток. Затем он повторяет эту выборку с выбранным интервалом «стробирования» в течение некоторого количества итераций. Подобно стробоскопу, этот вид анализа делает снимки данных в движении, чтобы обнаружить закономерности обновления записей или появления новых записей в целевой системе, которые соответствуют данным в исходной системе.

3) Обнаружение сегментов потока данных. Те же методы, которые используются для проверки человеческих гипотез относительно потока данных, можно использовать для обнаружения

недокументированных сегментов потока данных. Вместо того, чтобы использовать метод исчерпания, как это будет пытаться реализовать полностью автоматизированный подход, человеческие гипотезы высокого уровня, касающиеся потока данных, служат здесь в качестве руководства для дополнительных машинно-сгенерированных гипотез, которые могут обнаруживать незарегистрированные сегменты потока или предоставлять доказательства, исправляющие ошибки в человеческих данных.

4) Детализированная визуализация происхождения данных. Машины превосходны в детализации, сопоставлении и компиляции больших объемов подробной информации, недоступной человеку. Регуляторная отчетность, управление данными и решения по развитию экосистемы — все это выигрывает от возможности визуализировать поток данных в экосистеме логически (в рамках бизнес-концепций) и с возрастающими уровнями детализации, вплоть до уровней объектов и атрибутов. Для обеспечения такой прозрачности данных машины привлекаются для выполнения тяжелой работы, руководствуясь концептуальными моделями данных и высокоуровневыми описаниями потока, предоставляемыми людьми.

V. ЗАКЛЮЧЕНИЕ

Гибридный подход к происхождению данных, использующий сильные стороны человеческого интеллекта, классификацию данных и автоматическое обнаружения потоков, больше всего подходит для предоставления всеобъемлющего, значимого и действенного представления о движении данных в больших экосистемах.

Всестороннее и логическое понимание потоков данных с точки зрения бизнес-концепций важно для каждого содержательного обсуждения происхождения данных, для идентификации и отслеживания данных, подлежащих управлению, из какой бы экосистемы они не происходили. Это обеспечивает понимание последствий перенаправления потока информации для оценки эффективности и неэффективности архитектуры данных в масштабе предприятия, а также для точной оценки предложений по рационализации и оптимизации среды данных. Человеческий интеллект определяет семантику, а классификация информации использует машинное обучение для автоматизации связывания миллионов столбцов данных с бизнес-концепциями.

Человеческий интеллект также сужает всевозможные сегменты потока таким образом, что автоматическая проверка и обнаружение потоков становятся возможными. Детализация на уровне объектов и атрибутов, доступная в гибридном подходе, незаменима для любых строгих демонстраций происхождения данных и принятия обоснованных решений в управлении и развитии архитектуры данных в рамках крупномасштабных предприятий.

БИБЛИОГРАФИЯ

[1] Kalle Tomingas Margus Kliimask. The Art of Data Lineage». 2013.
[2] Robert Ikeda, Jennifer Widom. Data lineage: A survey. Technical report, Stanford InfoLab. 2009. URL: http://ilpubs.stanford.edu:8090/918/1/lin_final.pdf.

Enterprise-wide data origins

P.I. Zavalenova

Abstract — The article represents data origin methods that are applied in large-scale enterprises nowadays. The data lineage helps to track the movement of information flows in the ecosystem of the enterprise. First of all, the problems of the organization's activities are considered in the absence of any automated approach to the origin of information. As a result, the necessity of appropriate software is substantiated, which would make it possible to visually demonstrate the flow of data from the source system to the receiving system.

The article shows manual and automated approaches to data transfer, after which their main disadvantages are identified. As a result of the analysis, the most acceptable way of identifying the origin of data is proposed - a hybrid approach that uses human input, advanced algorithms and automated methods for checking and detecting data flows, which make the origin of data a manageable and efficient process.

The results obtained are presented in the form of conceptual models of the origin of the data. The work demonstrates what a modern system consists of, which makes it possible to define and visually represent the pedigree of information.

Keywords: Data Governance, Data Lineage, Data store, information, Data Provenance.

REFERENCES

[1] Kalle Tomingas Margus Kliimask. The Art of Data Lineage». 2013.
[2] Robert Ikeda, Jennifer Widom. Data lineage: A survey. Technical report, Stanford InfoLab. 2009. URL: http://ilpubs.stanford.edu:8090/918/1/lin_final.pdf.