

УДК 004

АВТОМАТИЗАЦИЯ DATA LINEAGE И КАТАЛОГИЗАЦИЯ ДАННЫХ В РАСПРЕДЕЛЕННЫХ СИСТЕМАХ

Илларионов С.А.

архитектор данных III уровня

*Санкт-Петербургский Государственный Университет Гражданской
Авиации (СПбГУГА),*

Fidelity National Financial

Санкт-Петербург, Россия

АННОТАЦИЯ

В статье рассмотрены подходы к автоматизации Data Lineage и каталогизации данных в распределенных системах. Показано, что ручное ведение метаданных и прослеживание происхождения данных не масштабируются при росте числа источников и команд, что приводит к снижению доверия к данным и увеличению затрат на их поиск и проверку. На основе анализа современных практик выделены ключевые требования к системе автоматизации: фиксация источников и версий алгоритмов преобразования, интеграция с оркестраторами и потоковыми платформами, поддержка справочных данных и обеспечение воспроизводимости. Рассмотрены типовые архитектурные компоненты, включая каталог данных, хранилище метаданных и сборщики lineage. Сформулированы условия, при которых автоматизация lineage и каталогизация становятся обязательными элементами платформы управления данными.

Ключевые слова: Data Lineage, каталогизация данных, распределенные системы, управление метаданными, Data Catalog, происхождение данных, качество данных, Data Mesh

AUTOMATION OF DATA LINEAGE AND CATALOGING OF DATA IN DISTRIBUTED SYSTEMS

Illarionov S.A.

Data Architect III

Saint Petersburg State University of Civil Aviation,

Fidelity National Financial

Saint Petersburg, Russia

ABSTRACT

The article discusses approaches to Data Lineage automation and data cataloging in distributed systems. It is shown that manual metadata management and tracing the origin of data do not scale with an increasing number of sources and teams, which leads to a decrease in confidence in the data and an increase in the cost of searching and verifying it. Based on the analysis of modern practices, the key requirements for the automation system are highlighted: fixing not only sources, but also versions of conversion algorithms, integration with orchestrators and streaming platforms, support for reference data and ensuring reproducibility. Typical architectural components are considered, including a data catalog, a metadata repository, and lineage collectors. The conditions under which lineage automation and cataloging become mandatory elements of the data management platform are formulated.

Keywords: Data Lineage, Data cataloging, Distributed systems, Metadata management, Data Catalog, Data origin, Data quality, Data Mesh

О.О. Леонова и Д.С. Пупков, анализируя управление качеством данных в государственных системах, отмечают, что данные обладают не только жизненным циклом, но и происхождением (lineage), а ведение дата-каталогов служит единым источником истины для организации [1]. В распределенной среде это свойство становится не вспомогательной инженерной задачей, а необходимым механизмом, позволяющим установить, откуда появилось

значение, как оно трансформировалось и кто за него отвечает. С.В. Кирьяков, рассматривая принципы Data Mesh, добавляет, что в децентрализованной модели единый каталог становится источником информации о доменах и данных, а система Data Quality выполняет роль внешнего аудитора [2]. Data Mesh требует, чтобы каждая доменная команда публиковала свои данные как продукт, что подразумевает не только описание наборов данных, но и автоматическое ведение метаданных, включая происхождение. В противном случае децентрализованное управление рискует превратиться в набор разрозненных хранилищ с несогласованными определениями. Особое значение приобретает связь lineage со справочными данными. А.С. Головкова и Н.В. Колос подчеркивают, что нормативно-справочные данные должны быть эталонными для всех программных компонентов: если один и тот же контрагент или номенклатура в разных системах имеет разные обозначения, даже корректно построенный технический lineage не даст ответа на вопрос о причине расхождений [3]. Поэтому каталогизация должна охватывать схемы таблиц и пайплайнов и справочные данные, которые часто выступают связующим звеном между доменами. Автоматизация lineage должна учитывать версионирование алгоритмов. Д.Н. Сотников, проектируя систему формирования статистических показателей, показывает, что результат вычислений неотделим от контекста расчета и версии методологии [4]. Для Data Lineage это означает, что фиксация только источников и таблиц недостаточна: необходимо сохранять версии кода, параметры обработки и логику трансформаций. Только в этом случае можно восстановить причину ошибки или расхождения и обеспечить воспроизводимость вычислений.

А.В. Сидорин и С.А. Федосин, анализируя архитектуру систем агрегации данных, отмечают, что для обеспечения качества и доверия к аналитике применяются «инструменты data observability и data quality (Great Expectations, Soda), а также метаданные и каталоги (Amundsen, DataHub), которые помогают отслеживать происхождение данных (data lineage) и упростить работу

аналитиков» [5]. В распределенных системах, где данные проходят через множество источников и преобразований, ручное ведение такого каталога невозможно: оно не масштабируется и быстро устаревает. Поэтому автоматизация lineage становится обязательной частью платформы данных. В противном случае организация сталкивается с ситуацией, когда невозможно ни восстановить историю трансформаций, ни понять, почему в разных хранилищах одни и те же показатели расходятся. Автоматизация Data Lineage в распределенных системах требует иного подхода, чем в монолитных хранилищах. Во-первых, lineage должен фиксировать не только перемещение данных между таблицами, но и версии кода трансформаций. Как отмечалось ранее, без этого невозможно установить, какая именно логика обработки была применена к конкретному набору данных [4]. На практике это означает, что сборщики lineage должны интегрироваться не только с системами хранения, но и с инструментами оркестрации (Airflow, Dagster), ETL/ELT-движками (Spark, dbt), а также потоковыми платформами (Kafka, Flink).

Во-вторых, каталог данных перестает быть статическим справочником и превращается в живой реестр, автоматически пополняемый из пайплайнов. В распределенной среде доменные команды публикуют свои данные как продукты, и каталог становится источником информации о том, какие наборы данных существуют, кто за них отвечает, как они связаны и какого они качества [2]. Это означает, что каталог должен поддерживать не только технические метаданные (схемы, владельцы, расположение), но и бизнес-описания, глоссарии и показатели качества.

В-третьих, важным аспектом является интеграция lineage со справочными данными. Как показано в работе А.С. Головковой и Н.В. Колос, даже корректно построенный технический lineage не решает проблему расхождений, если одни и те же объекты учета по-разному называются и кодируются в разных системах [3]. Поэтому каталогизация должна охватывать нормативно-справочную информацию и связывать ее с техническими артефактами. Типовая

архитектура автоматизированной системы lineage и каталогизации включает три контура:

- сбор метаданных: адаптеры к источникам данных, оркестраторам и вычислительным движкам;
- хранение и обработка: графовое или реляционное хранилище lineage, поисковый индекс, API для доступа;
- представление: пользовательский интерфейс каталога, поиск, визуализация графов зависимостей, автоматические уведомления об изменениях. В индустрии эти функции частично реализуются в таких инструментах, как Apache Atlas, DataHub, Amundsen, OpenLineage и Marquez. Выбор конкретного инструмента вторичен по сравнению с принципом: lineage должен собираться автоматически из пайплайнов, а не вноситься вручную. Связь с качеством данных также принципиальна. Автоматически собранный lineage позволяет быстро определить, какие датасеты и отчеты затронуты изменением схемы или сбоем в одном из источников. Это сокращает время расследования инцидентов и снижает риски использования устаревших данных. Кроме того, lineage дает возможность проверять соответствие реальных пайплайнов утвержденным политикам обработки, что важно для аудита и соответствия нормативным требованиям [1].

Проведенный анализ показал, что в распределенных системах ручное управление lineage и каталогами данных неэффективно и не обеспечивает требуемого уровня доверия к данным. Автоматизация этих процессов требует интеграции с пайплайнами обработки, учета версий алгоритмов, поддержки справочных данных и связи с системами качества. Ключевым условием успеха является отказ от восприятия каталога как статического документа и переход к модели, в которой lineage собирается автоматически и непрерывно.

Библиографический список

1. Леонова, О. О., Пупков, Д. С. Управление качеством данных в условиях цифровой трансформации государства в сфере интеллектуальной собственности // Вестник ФИПС. – 2026. – Т. 5, № 1(15). – С. 62-72.
2. Кирьяков, С. В. Data Mesh и микросервисная архитектура: децентрализованное управление данными // Научная статья года 2024. – Пенза: Наука и Просвещение (ИП Гуляев Г.Ю.), 2024. – С. 8-10. – EDN LХКНАU.
3. Головкова, А. С., Колос, Н. В. Архитектура хранения и нормализация нормативно-справочных данных на этапе создания единого цифрового контура на предприятии // Экономика. Информатика. – 2023. – Т. 50, № 1. – С. 173-182. – DOI 10.52575/2687-0932-2023-50-1-173-182.
4. Сотников, Д. Н. Архитектура информационной системы формирования статистических показателей в распределенных информационных средах // Современные студенческие исследования. – Пенза: МЦНС «Наука и Просвещение» (ИП Гуляев Г.Ю.), 2026. – С. 26-31.
5. Сидорин, А. В. Агрегация и анализ данных из различных источников / А. В. Сидорин, С. А. Федосин // Научный спектр. – 2026. – № 1. – С. 5-14. – EDN QGBKUY.