Происхождение данных (Data lineage)

> Синонимы: Отслеживание данных, Родословная данных, Data provenance
> Разделы: Бизнес-задачи

Под термином *Data lineage* (англ. lineage — происхождение, родословная) понимают процесс отслеживания перемещения данных в организации — от момента их генерации до конечного потребителя. При этом учитываются все промежуточные этапы преобразования и обработки. Все точки перемещения данных регистрируются и документируются, формируя описание их жизненного цикла.

Это описание позволяет IT-специалистам и аналитикам наблюдать различные этапы перемещения и преобразования данных с целью оценки их качества, точности и согласованности. Кроме этого, в случае проблем можно отследить их до точки возникновения, что позволит выявить и устранить причины. Раннее обнаружение и устранение проблем позволяет минимизировать простои конвейера данных компании.

Процесс отслеживания реализуется путем последовательного считывания метаданных, которые формируются на каждом шаге продвижения информации в компании. В некоторых случаях в них могут добавляться специальные метки (теги), упрощающие процесс.

Внедрение Data lineage в практику управления данными в компании позволяет:

- дать IT-команде понимание того, как изменения на ранних этапах перемещения данных могут повлиять на последующие этапы, что позволяет выявлять и предупреждать проблемы;
- в случае возникновения проблем, продвигаясь назад по треку данных, быстрее обнаружить их первопричину и устранить ее;
- своевременно информировать потребителей данных о возможных ошибках и причинах их возникновения, что повышает доверие к данным.

Выделяют два уровня организации процесса Data lineage: таблицы (table-level lineage) и столбца (column-level lineage). Первый уровень является наиболее простым и часто используемым. Он показывает продвижение данных на уровне всей таблицы, но не позволяет отследить происхождение отдельных ее столбцов (например, если таблица была создана путем интеграции отдельных полей из разных источников). Отслеживание на уровне столбцов является более детальным и позволяет выявлять глубинные проблемы в данных, такие как пропуски и выбросы.

Одной из проблем внедрения процесса Data lineage является его высокая сложность, особенно при больших и разветвленных потоках данных. Поэтому его ручная реализация малоэффективна и необходимо использовать специализированные инструменты.