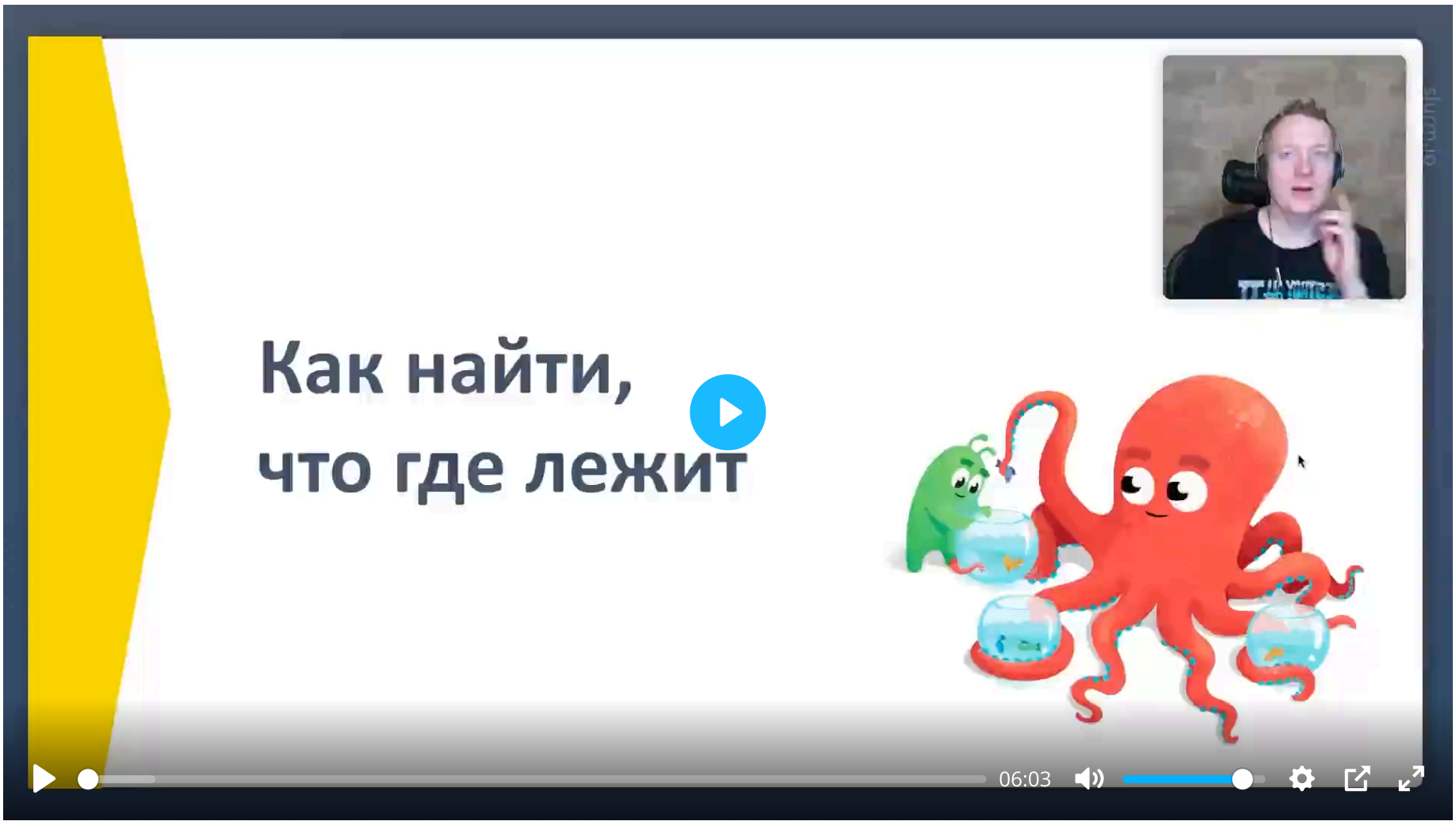
11.4 Частые проблемы при построении архитектуры и подходы к их решению



Тексовая расшифровка видео:

КАК НАЙТИ, ЧТО ГДЕ ЛЕЖИТ

План:

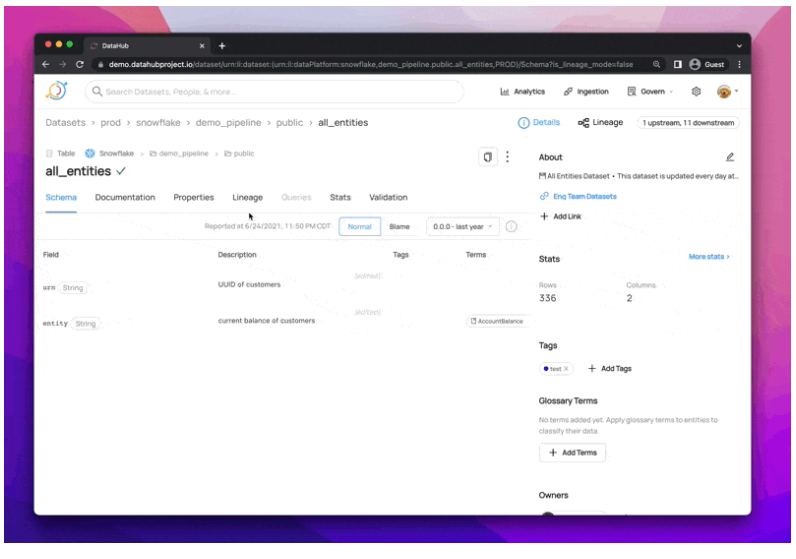
- Knowledge Sharing;
- Data Observability & Data Catalogs;
- Data Lineage;
- История экспериментов.

Knowledge Sharing

В компаниях часто есть Confluence или что-то еще Wiki-образное.

Каждая команда ведет документацию самостоятельно. Однако на вопрос «из какого датасета взять колонки?» чаще всего никто не может дать ответ.

Data Observability & Data Catalogs



Нам может помочь дата-каталог.

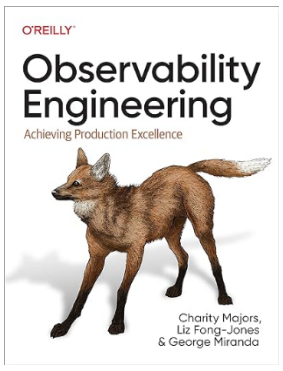
Дата-каталог – это дополнительная БД с интерфейсом для поиска, которая позволяет нам хранить информацию о том, что где лежит.

Мы можем доставать данные напрямую из хранилищ; можем подключить Greenplum и т.д., а можем модифицировать ETL-пайплайн (когда создаются датасеты, дополнительная метаинформация закидывалась в API дата-каталога и там сохранялась).

Более глобальный вариант – **Data Observability**.

Нам нужно представление данных в виде бизнес-структуры. Нам необходимо понимание того, насколько данные качественные. Вся эта концепция, когда мы можем проверить состояние данных, и есть Data Observability.

Подробнее об этом можно почитать в этой книге:



Reading on Data Observability

Data Lineage

Data Lineage – это история того, как данные передвигались в системах (по каким слоям проходили, в каких были стадиях и т.д.):

Разные дата-каталоги включают в себя инструменты визуализации Data Lineage по шагам.

История экспериментов

Когда мы работаем с данными, у нас есть отдел аналитики и дата-сайентисты. Одна из типичных проблем заключается в неорганизованности процесса. Новым сотрудникам требуется достаточно много времени, чтобы вникнуть во все это. Поэтому используется история экспериментов. Один из самых популярных инструментов – mlflow.

Mlflow – это БД, которая хранит историю того, как вела себя модель.

Там хранятся:

- Веса модели;
- Ссылка на датасет или сам датасет;
- Разные параметры и т.д.