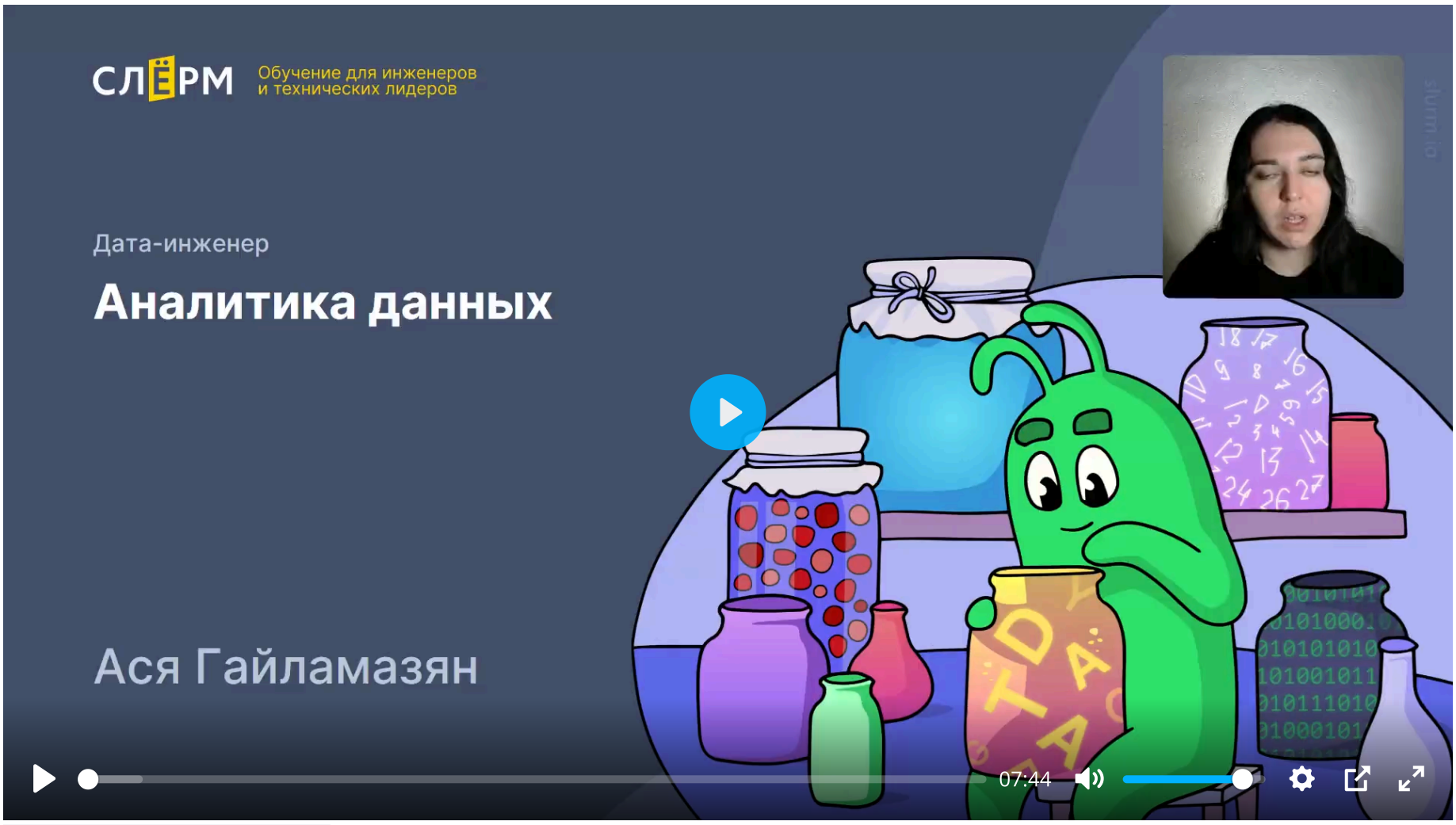


Презентация к уроку 12.2.1

Тексовая расшифровка видео:

АНАЛИТИКА ДАННЫХ

План:

- Изменчивые данные, согласованность и визуализация;
- Изменчивость данных;
- Эффективная обработка изменчивых данных;
- Примеры изменчивых данных;
- Каталоги данных;
- Data governance;
- Согласованность данных;
- Жизненный цикл данных.

Изменчивые данные, согласованность и визуализация

Ранее мы говорили про понятия «V», а также о том, что эти понятия являются разными для Big Data и Business Intelligence. Тем не менее у них есть пересечения.

Мы поговорим об инструментах, которые отвечают за:

Veracity – достоверность данных:

- любой анализ данных бесполезен, если данные недостоверны;

  неточность данных может привести к неправильным решениям.

То есть это инструменты Data Quality, выявляющие аномалии данных, позволяющие дополнять пропущенные данные и находить ошибки.

Variability – изменчивость данных:

- значение одних и тех же данных может различаться в зависимости от контекста;
- алгоритмы должны быть в состоянии понять контекст и расшифровать точное значение слова в этом контексте.

Variability отличается от Variety.

Variety – это разные типы данных, а ***Variability*** – это самостоятельная изменчивость данных. С течением времени они меняются на источники, меняются форматы. Нужно готовиться к тому, что обработка данных тоже может меняться.

Visualization – визуализация данных:

- визуализация делает большие данные доступными для человеческого восприятия;
- визуализация больших объемов сложных данных понятнее для человека, чем электронные таблицы и отчеты.

То есть данные должны быть представлены в графическом/визуальном виде. Данные должны быть понятными и легко читаемыми.

Изменчивость данных

Изменчивые данные – это те данные, которые с течением времени могут каким-либо образом менять формат и содержание на источники из-за автоматических или пользовательских обновлений.

Принципы работы с изменчивыми данными:

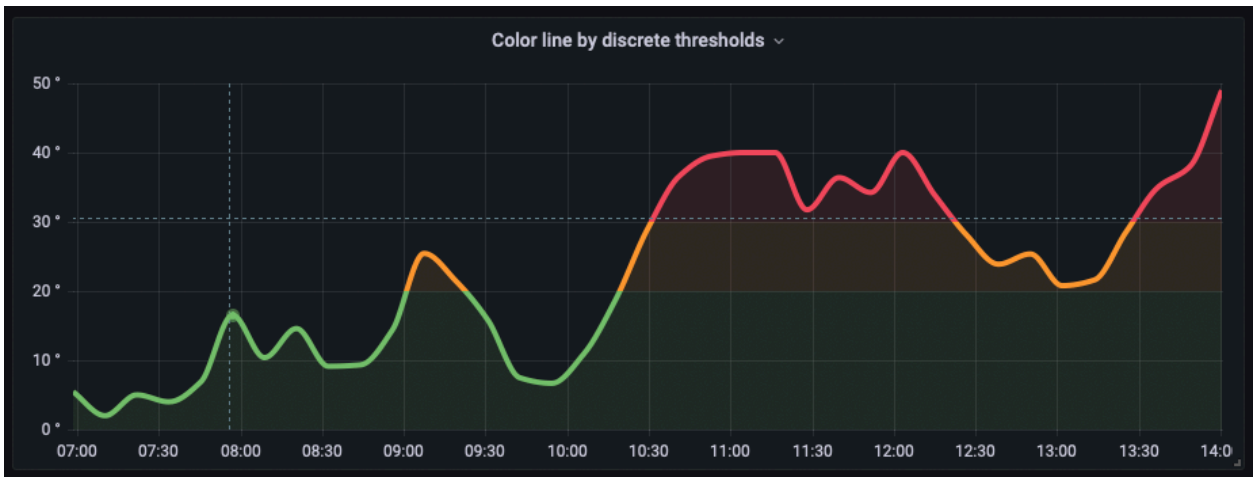
- Используем специальные инструменты потоковой обработки данных (например, Apache Kafka, Apache Flink).
- Учитываем при агрегации данных изменчивость и создаем более устойчивые системы.
- Настраиваем мониторинг (например, в Grafana) для того, чтобы мониторить и управлять качеством данных.
- Анализируем и прогнозируем изменения с помощью инструментов «Machine Learning»

Эффективная обработка изменчивых данных

Основные подходы обработки изменчивых данных:

- Ведите непрерывный мониторинг.
- Установите стандарты качества данных (авто-тесты).
- Используйте статистические методы для анализа и понимания изменчивости данных.
- Рассмотрите возможность сглаживания или агрегирования данных.
- Установите надежные процессы обеспечения качества для проверки входящих данных (авто-тесты/ручные тесты).
- Используйте машинное обучение.
- Храните подробную документацию об источниках данных, методах сбора и любой контекстной информации.
- Рассмотрите возможность использования методов, устойчивых к изменениям или тенденциям в данных.

Рассмотрим график:

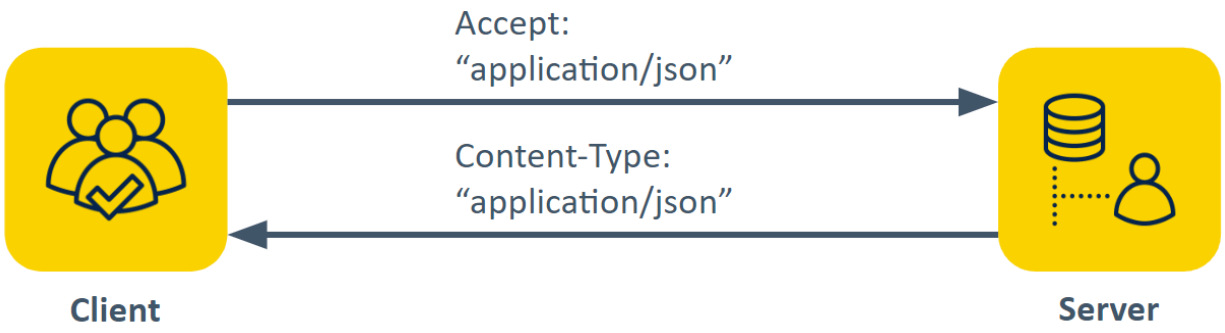


Это типичный график в Grafana. Пунктирные полоски на нем – это некоторые пороговые значения, после достижения которых срабатывают Alert'ы.

Alert'ы – это предупреждения о сработавших пороговых значениях.

Примеры изменчивых данных

Примером изменчивых данных является API сайта, которое после каждого обновления сайта отправляет разный контент. Помимо API, примерами могут быть данные датчиков умного дома или данные в транспортном потоке, так как эти данные тоже подвержены изменениям.

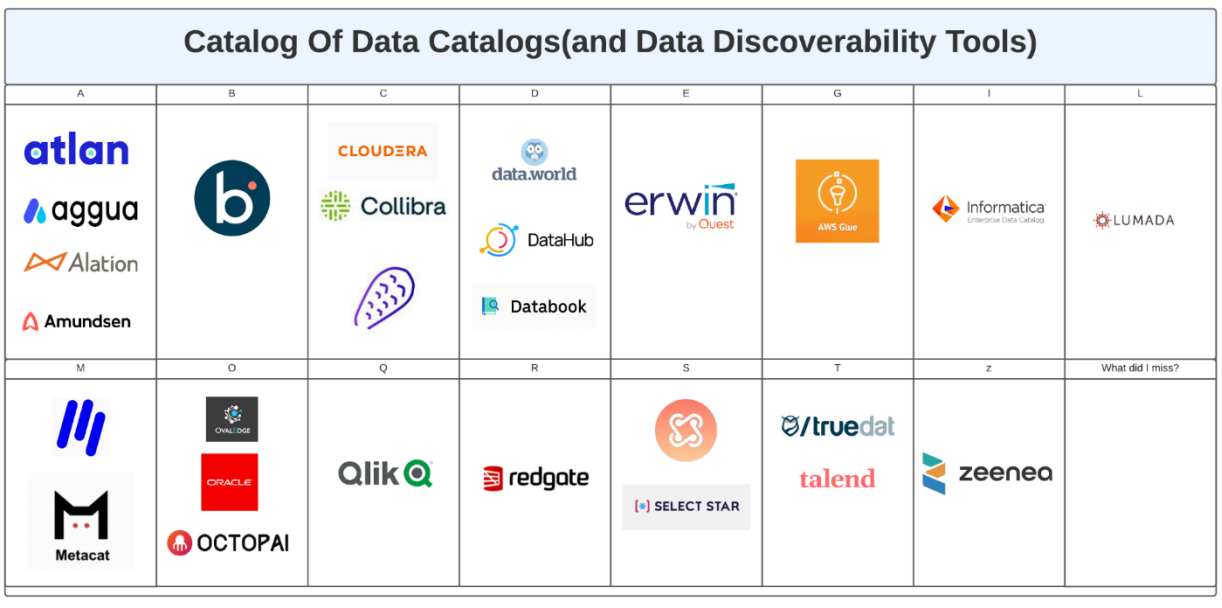


Каталоги данных

Каталоги данных – это структуры, которые хранят в себе информацию о данных в организации.

В каталоге данных можно:

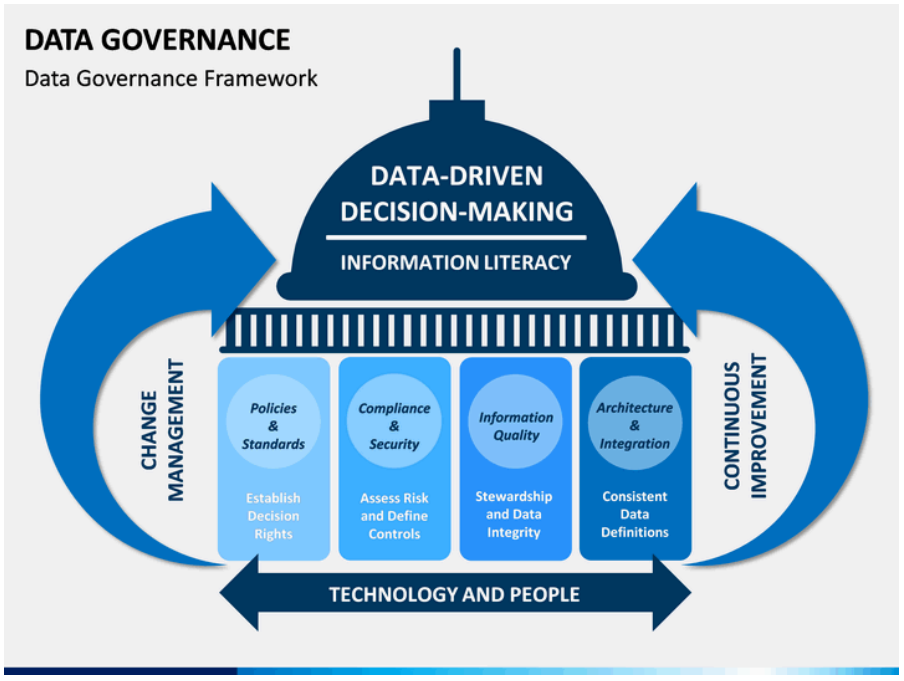
- найти информацию о данных;
- найти краткую сводку;
- найти структуру данных;
- проследить *lineage* (путь от источника до целевой таблицы);
- посмотреть профайлинг данных (краткая сводка о таблице);
- посмотреть владельцев данных и роли доступа.



Data governance

Data governance – это фреймворк управления данными. Это набор характеристик, правил, политик, которых

- **Data governance** – это набор документов.
- **Data management** – это прикладные практические действия, выполняющиеся для соблюдения Data governance.

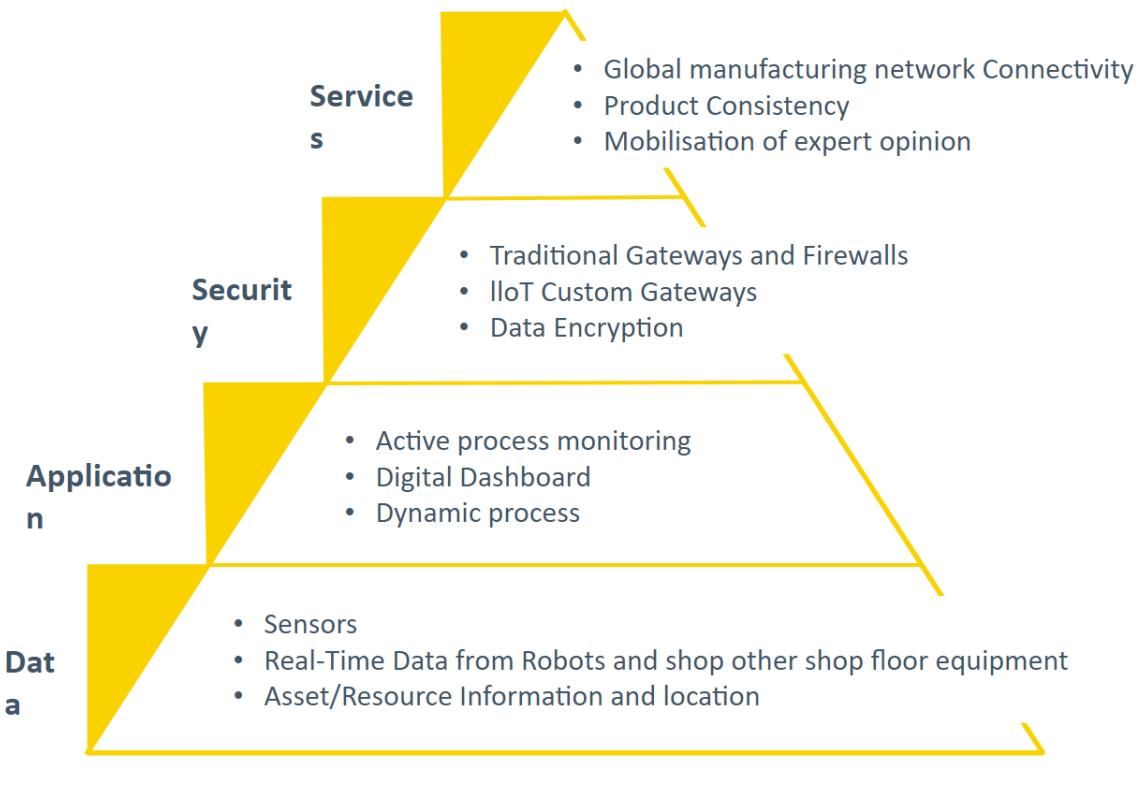


Согласованность данных

Поддержка консистентности данных (согласованности) – это важный аспект Data governance в управлении данными, который гарантирует, что данные в системе остаются непротиворечащими друг другу.

Для этого используются механизмы, например, поддержка механизма транзакций.

Это очень похоже на CAP-теорему:



- «S» здесь означает «консистентность». Здесь используются те же правила: например, используются транзакции для проведения операции. В случае каких-либо ошибок все операции откатываются назад.
- Контроль версии, чтобы восстановить к предыдущему состоянию систему в случае неполадок/необходимости.
- Мониторинг и аудит данных. Отслеживаются изменения и идентифицируются потенциальные проблемы.
- Управление доступом к данным. Это разделенный доступ по ролям во избежание несанкционированного доступа к критичным данным.

На изображении показаны разные слои данных:

- **Базовый (начальный) слой** – сами данные;
- **Приложения**, которые мониторят и отображают на дашборде.

  Безопасность, где данные шифруются и отправляются каким-либо образом, используя безопасные

- **Сервисы**, где производятся какие-либо действия над данными.

Жизненный цикл данных

- На этапе «**Collection**» данные собираются из разных API или генерируются.
- На этапе «**Storage**» данные сохраняются в хранилище. Они могут быть архивированными или передаваться другим системам.
- На этапе «**Maintenance**» происходит обработка данных. Данные могут агрегироваться.
- На этапе «**Usage**» происходит анализ и использование данных для принятия решений, выявления тенденций и т.д.
- На этапе «**Cleaning**» данные могут быть либо отправлены в долгосрочные архивы, либо удалены. Позже данные снова поднимаются из архивов или собираются из источников.

Как вам урок? 😠 😐 🔥

Изучил, далее >