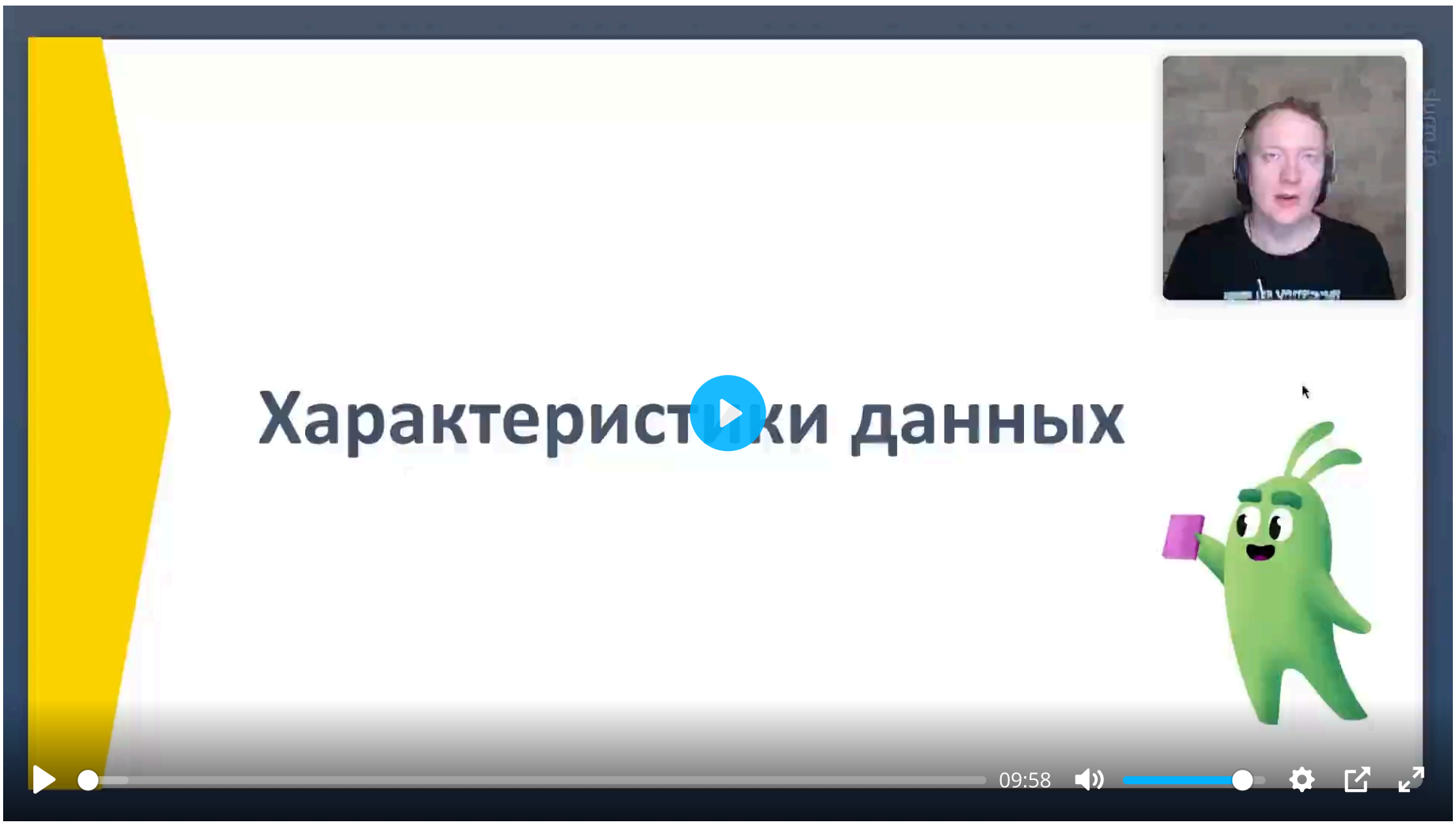
11.2 Традиционная архитектура хранилищ данных



Тексовая расшифровка видео:

ХАРАКТЕРИСТИКИ ДАННЫХ

План:

- Типы данных в организации;
- Много букв «V»;
- Master Data Management;
- DMBOK и DCAM;
- Выбор архитектуры зависит от компании.

Типы данных в организации

Еще на первых занятиях мы говорили о разных типах данных по структурированности:

- Структурированные данные

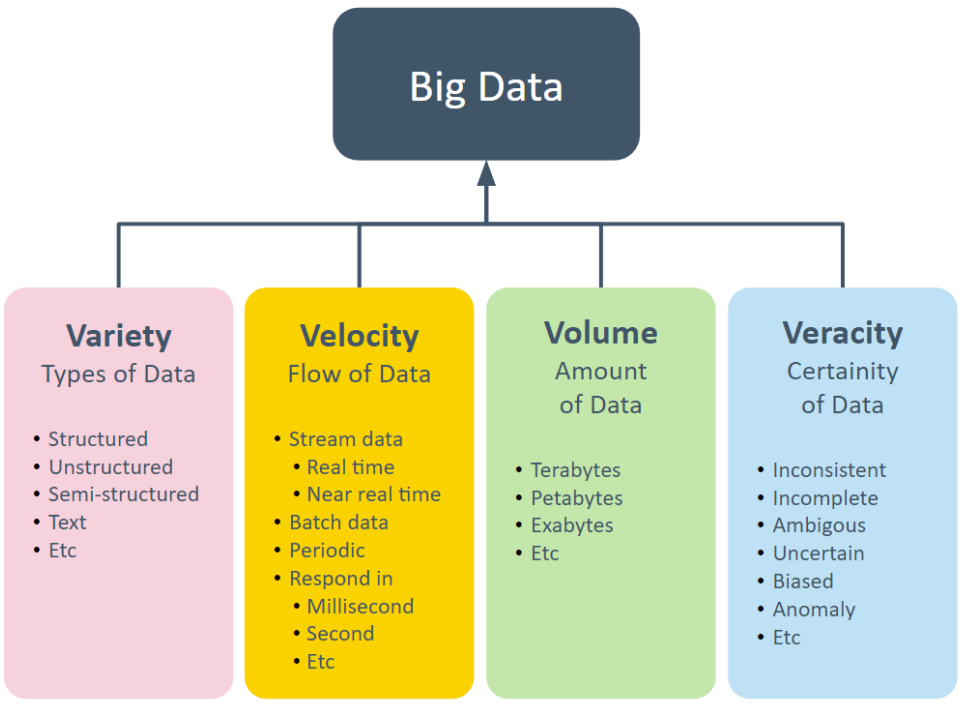
- Полуструктурированные данные

- Неструктурированные данные

Чем больше компания, тем вероятнее столкнуться с «адской» смесью всех трех типов, где одна часть структурированная, а другая – неструктурированная. Наша цель – подсветить правильным образом определенные аспекты этих данных.

Много букв «V»

В западной литературе неспроста работу с большими данными описывают словами на букву «V»:



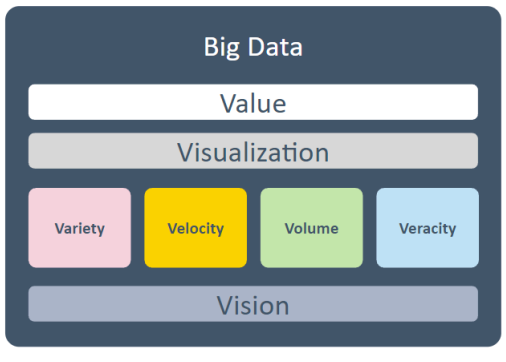
Variety – это про степень структурированности, а также про разные форматы, которые у вас есть на входе.

Velocity, где может быть streaming, batching.

Volume – это про общий объем (сколько вам нужно сохранить в процессе).

Veracity – это про проверку качества данных (насколько они хорошо подходят к задачам, которые мы с их помощью пытаемся решить). Это один из самых сложных аспектов.

В некоторые источники добавляют еще больше слов на букву «V»:



Когда мы начинаем моделировать какую-либо архитектуру, нужно понять, с какими данными мы имеем дело и что хотим из них получить.

Master Data Management

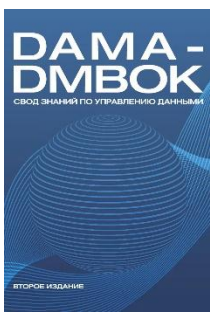
Сам процесс, с точки зрения бизнеса и дата-стратегии, глазами бизнеса выглядит следующим образом:

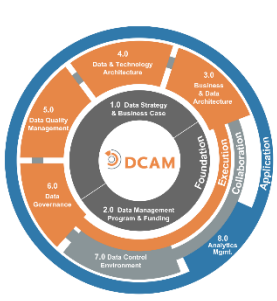
- Бизнес имеет **доменную область**, в которой он оперирует;
- В ней есть определенные **бизнес-объекты**, представление о которых необходимо иметь, чтобы **принимать решения**;
- Подобными объектами могут быть **Клиент, Дилер, Локация, Маркетинговая Кампания**;
- Готовую коллекцию объектов называют «**Golden Record**», типичный пример – **360-градусный обзор клиентов;**
- **MDM** – название **процессов по формированию** этих объектов **на основе собираемых данных.**

DMBOK и DCA

Термины, о которых мы говорим, существуют уже достаточно давно. Существуют корпоративные книги, где описаны стандарты того, как подобные вещи делаются, называются, как они между собой связаны.

Есть несколько конкурирующих стандартов – DMBOK (Data Management Body of Knowledge) и DCA:





DCA – альтернативный стандарт, где располагаются те же термины, однако у них есть различия в трактовке.

Если вы идете архитектурить серьезное дата-решение в какую-либо большую компанию, где уже знакомы с такими стандартами, то следует уточнить по какому именно стандарту там строят.

Выбор архитектуры зависит от компании

Важно понимать, что **представляет из себя компания, где мы это выстраиваем**, а также важно понимать, **какими ресурсами компания обладает.** Если компания маленькая, то процессы будут отличаться от процессов большой компании. Это же касается стека: в разных командах и компаниях может быть legacy-стек, может быть что-то собранное на Hadoop и т.д., а может быть «свежак», например, развернутая новая версия Teradata.

Важно понимать, **что компания считает своим продуктом.** Возможно, компания производит конечное решение. У такой компании может быть набор собственных подходов к тому, как происходят релизы, как происходит выпуск новой версии решения, которые абсолютно не подходят и не масштабируются на дата-задаче, когда нужно быстро исследовать гипотезы, быстро поднимать дашборды и т.д. **Важно уметь балансировать** в случае, если компания продуктовая или R&D.

Важно понимать корпоративную культуру – как именно происходит предложение и внедрение новых технологий в компании/в команде, а также насколько сложно привнести новые решения. Например, если компания западная, то при их внедрении часто происходит целая «эпопея»: нужно заключить бизнес-контракт с поставщиками этого решения, подписать с обеих сторон и т.д. Если же это небольшой стартап, то можно поставить себе комьюнити-версию.

Как вам урок? 😠 😐 🔥

Изучил, далее >