Михальченко Вера DSU-30

Итоговая работа по курсу Big Data

1. Практика Google Sheets
2. Практика Python
3. Теоретическая часть

Основные бизнес-отчёты

1. Отчет по основным бизнес-метрикам:
   - Выручка по транзакционной и подписной моделям потребления
   - Средняя выручка на посетителя
   - Средний чек
   - Среднее число покупок на пользователя
   - Конверсия в покупку
   - Конверсия в просмотр по подписке
   - Конверсия в пробный период

Основные бизнес-отчёты

2. Отчет по пользователям:

- Портрет пользователя
- География
- Предпочтения
- Графики зависимостей (устройство, время дня и т.п)
- Кластеризация пользователей (какие группы пользователей и в каком соотношении)

Основные бизнес-отчёты

3. Отчет по контенту:

- Динамика количества поступления нового контента
- Ключевые поставщики контента
- Соотношение показателей просмотров (по типам контента, поставщикам и др.)
- Классификация контента по разным основаниям (пользователям, поставщикам, географии, устройствам и пр.)
- Эффективность работы рекомендательной системы

Основные бизнес-отчёты

4. Отчет по оценке эффективности аналитики и DS:

- Сравнение ключевых показателей для бизнеса по результатам деятельности команды аналитики и DS (по времени и проектам)
- Траты и “выгода”, связанные с работой команд
- Точки роста (необходимые меры и траты для еще более эффективной работы)

Основные имеющиеся данные, источники их поступления и процесс заливки

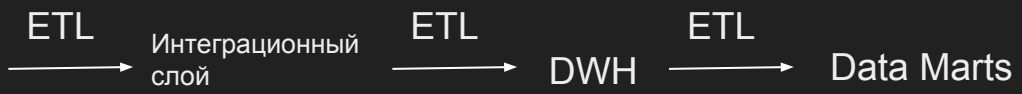
1. Данные:

   Внутренние:

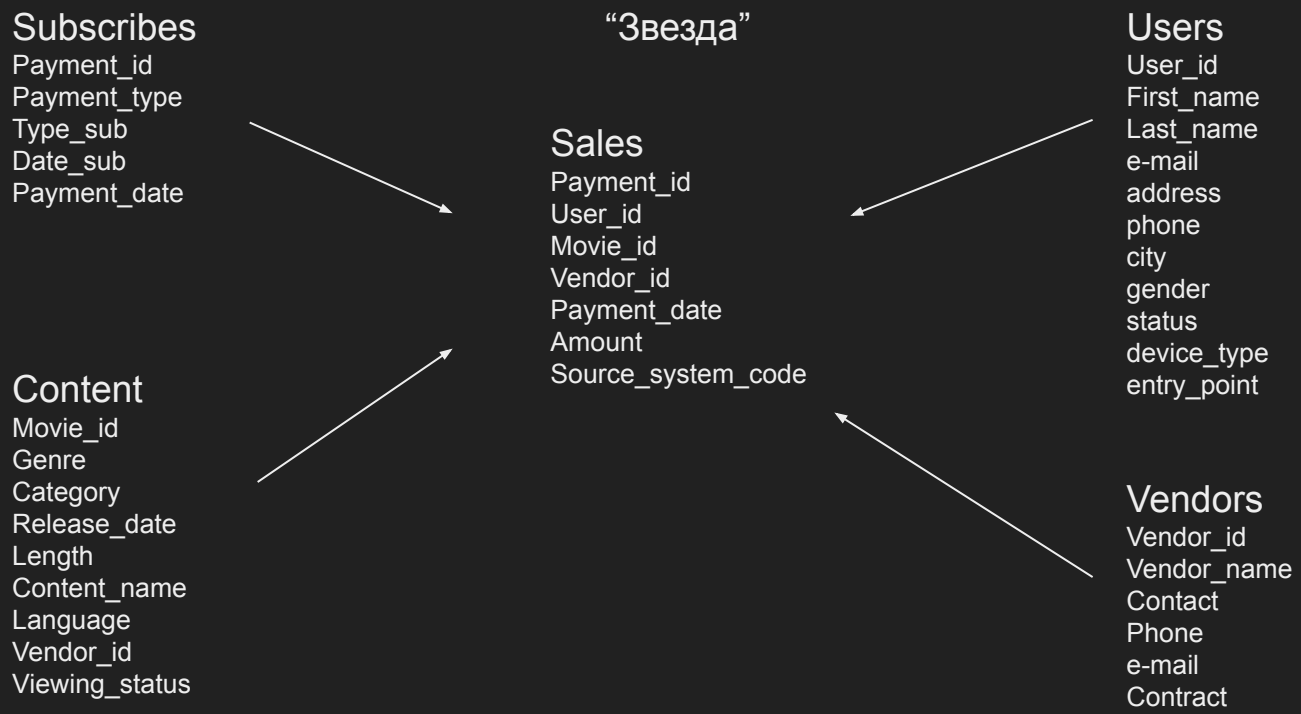
   - Маркетинг
   - Продажи
   - CRM
   - ERP

   Внешние:

   - ERP
   - Данные других компаний, открытые, платные статистические данные



Основные сущности в хранилище данных



Основные проверки на качество данных

Осуществляются дата инженерами и аналитиками

Качественные показатели:

- Корректность
- Согласованность
- Полнота
- Своевременность
- Метаданные

Для последующего анализа:

- Доступность источников данных
- Определение ключевых атрибутов
- Размер источников данных, подмножество данных
- Значения Null
- Дубликаты
- Единый формат (телефоны, e-mail, имена, даты, адреса и т.п)
- Пустые значения
- Количество уникальных значений

Data - проект по улучшению показателей бизнеса

Разработка более эффективной рекомендательной системы для онлайн кинотеатра, основанной на гибридной модели, объединяющей 2 подхода: коллаборативную фильтрацию и контентную модель.

Crisp DM

1) Business Understanding (аналитики)

   Сбор справочной информации (текущая рекомендательная система)

   Оценка ситуации (анализ трат и выгод внедрения новой системы, риски)

   Определение целей (увеличение конверсии в подписки, покупок и аренды контента в сравнении с контрольной группой (AB - тестирование))

   Критерии успеха. Финансовая выгода > затрат на внедрение и поддержание новой системы

   Создание плана

   Определение тестовой и контрольной группы

Crisp DM

2) Data Understanding (аналитики)

Сбор, описание, исследование и изучение качества данных

3) Data Preparation (дата сайентисты, дата инженеры)

Выбор, очистка, расширение и сохранение данных

4) Modeling (дата сайентисты)

Выбор и проверка модели

5) Evaluation (аналитики)

Анализ результатов(сравнение показателей тестовой и контрольной групп, проверка ошибок, наличие инсайтов, точки роста)

Crisp DM

6) Deployment (дата сайентисты, разработчики)

Планирование внедрения модели в случае успешных результатов тестирования

Планирование мониторинга и технического обслуживания

Итоговый обзор проекта