Задание 2:

Представим, что вы работаете в крупной сети гипермаркетов и вам предстоит с 0 строить платформу для Аналитики и Data Science.
Для каждого из приведенных примеров решить, что лучше использовать для хранения данных: классическую RDBMS, Hadoop HDFS или простой FTP-сервер.
Каждый ответ пояснить 1-3 предложениями.

1. Куда лучше загружать чеки с информацией о купленных товарах? Данные предполагается использовать на каждодневной основе для аналитики продаж, а также для генерирования фичей для будущих рекомендательных систем.

Лучше складывать в Hadoop HDFS (нереляц.БД) - не надо заморачиваться со структурой БД в случае изменения чека (новые поля, свойства и тд.)

2. Куда лучше загружать xlsx-файлы с отчётами, которые будет делать команда аналитиков?

НА FTP-сервер: поскольку уже готовый xlsx-файл, то его можно быстро «втыкать» на страницы аналитики и нет смысла хранить его в других системах.

3. Куда загружать презентации и фотоотчеты с прошедших корпоративных мероприятий?

В Hadoop (на дальнюю полку гаража) для будущих проектов. Поскольку они не требуются ежедневно – смысла платить за них в RDBMS нет

4. Бизнес решил копировать и хранить данные, которые генерирует платформа Google Analytics, на своей стороне. Куда лучше сохранять эти многочисленные тяжелые логи?

В Hadoop. Данные тяжелые, но нужны редко, для исследований. Поэтому смысла хранить в RDBMS нет. Про FTP-сервер и подавно.

5. Куда лучше сохранять записи с камер, учитывая, что в планах лежит разработка ряда нейросетей по противодействию кражам и постепенного перехода к концепции Unmanned store.

В Hadoop. Данные не реляционные, большого объема. По другому невыгодно хранить.

6. Куда лучше выгружать логи всех наших IT-систем, для анализа, который будет проводиться раз в несколько месяцев.

В Hadoop. Объем данных растет быстро, используется редко.

7. Куда лучше сохранять хешированные данные наших клиентов, для обмена с партнёрами (кобренд-активности)? FTP-

В RDBMS. Прирост данных не лавинный. Данные структурированы, могут храниться в RDBMS