Обработка данных на кластере с помощью hive

Данные для задания:
https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings/data

1. Скачать данные, положить в hdfs
2. Зарегистрировать external таблицу на данных в базе данных raw
3. Зарегистрировать таблицу ORC для данных в базе данных ods. Типы данных должны соответствовать данным (строки, числа и пр.), реализовать партиционирование по году
4. Написать и выполнить запрос для переноса данных из базы raw в базу ods
5. Зарегистрировать таблицу с метаданными о Platform в слое md. В таблице должны быть две колонки с Platform - строковое значение и назначенный вами числовой идентификатор
6. В слое ads собрать информацию по продажам в EU - название игры, идентификатор платформы из метаданных, объем продаж
7. Написать запрос и найти с помощью него за каждый год игру с максимальными продажами из слоя ads, вывести платформу с помощью зарегистрированного справочника

Усложненный вариант задания:
По собранным ранее данным из ВК подготовить схему хранения данных в хайв

1. Зарегистрировать external таблицу на данные
2. В orc формате в ODS слое разложить данные на таблицы “пользователь” и “посты пользователя”
3. Зарегистрировать таблицу для датасета
4. Написать запрос, собирающий датасет по данным для вашей модели. Описать признаки и проделанные преобразования

Присылать нужно весь код с комментариями, что он делает, или общим описанием процесса. Например “файл 1 выполняет регистрацию данных в raw, …”.
А так же пример выполнения запроса select * from … limit 10; из всех таблиц. Для первого варианта задания прислать результат выполнения запроса из пункта 7.