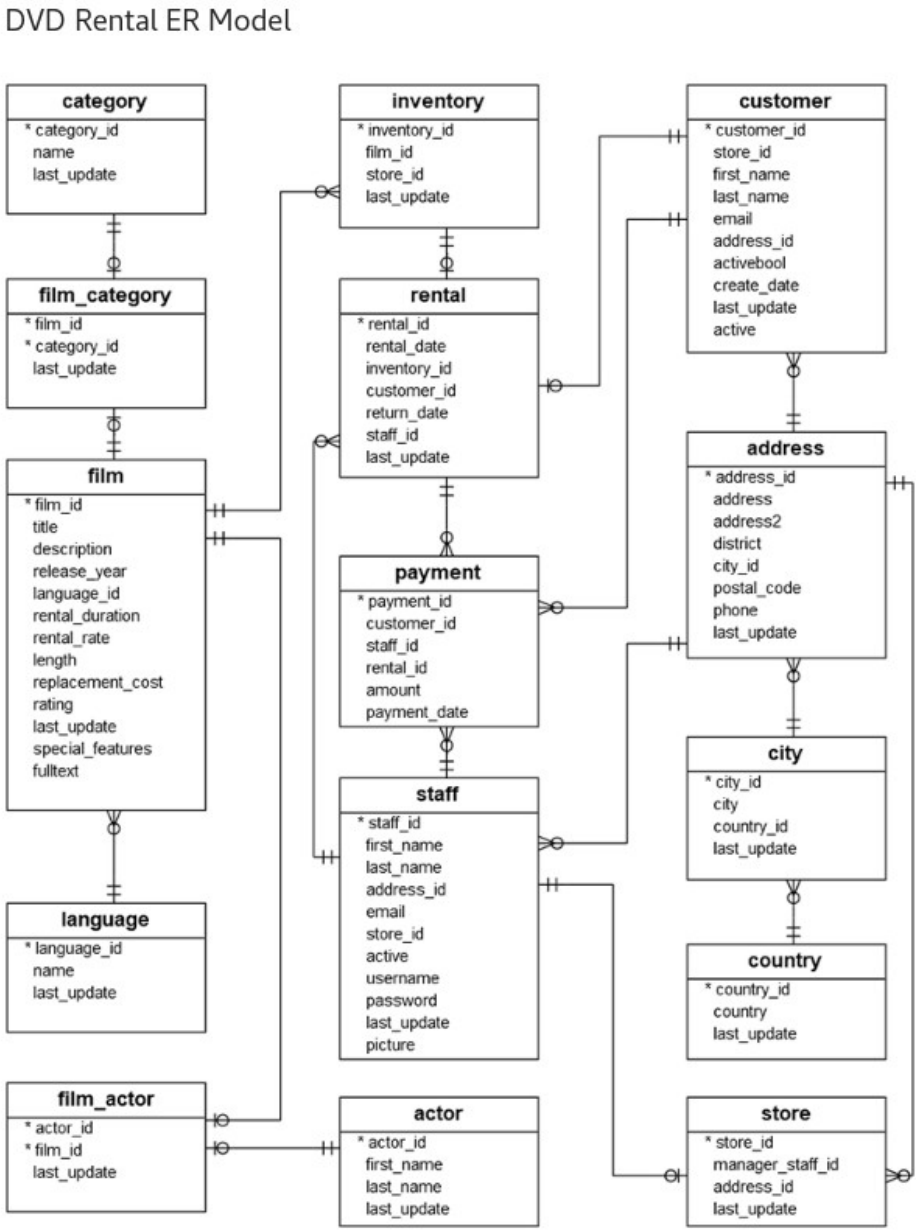
DATA ENGINEERING INTERVIEW

INSTRUCTION

1) THREE datasets are provide:

a) DVD rental which should be loaded to PostgreSQL database.

DATA SET EXPLORATION



The DVD database represents the business processes of a DVD rental store. The DVD rental database has many objects including:

- 15 tables
- 1 trigger
- 7 views
- 8 functions
- 1 domain
- 13 sequences

There are 15 tables in the DVD Rental database:

- actor – stores actors data including first name and last name.
- film – stores film data such as title, release year, length, rating, etc.
- film_actor – stores the relationships between films and actors.
- category – stores film's categories data.
- film_category- stores the relationships between films and categories.
- store – contains the store data including manager staff and address.
- inventory – stores inventory data.

- rental – stores rental data.
- payment – stores customer's payments.
- staff – stores staff data.
- customer – stores customer data.
- address – stores address data for staff and customers
- city – stores city names.
- country – stores country names.

b) Big-mart sales which is a csv file.

C) Resource Monitoring Data-set(csv files)

DETAILED DESCRIPTIONS OF DATA FILES

==========

Brief descriptions of the data.

physical_cores -> Physical Core: is an independent CPU instance on a multi core-processor.

logical_cores -> Logical Core: intern refers to the ability of each core doing 2 or more tasks simultaneously. This is achieved by enabling hyper-threading on the cores. Each single physical core can be divided in to multiple logical core by enabling hyper-threading on them.

max_cpu_frequency -> Maximum cpu Clock speeds measured in megahertz

min_cpu_frequency -> Minimum cpu Clock speeds measured in megahertz

current_cpu_frequency -> Current cpu Clock speeds measured in megahertz

total_ram -> Total ram measured in GB, MB, KB or B depending on ram available in test machine

total_available_ram -> Total available ram measured in GB, MB, KB or B depending on ram available in test machine

total_swap -> Total swap space measured in GB, MB, KB or B depending on ram available in test machine

free_swap -> Total free swap measured in GB, MB, KB or B depending on ram available in test machine

start_time -> Time in seconds seconds since the epoch (Assumed to be January 1, 1970, 00:00:00 (UTC)) at the start of the task type

stop_time -> Time in seconds seconds since the epoch (Assumed to be January 1, 1970, 00:00:00 (UTC)) at the end of the task type

time_spend -> Time spend on the task it the difference between stop_time and start_time in second

task_count -> Total Individual task of type X performed in a give time

task_type -> Type of task that they system is working on

2) Submission can be in the following format:

a) Json file.

b) Notebook.

c) python file.

SECTION ONE

1.a)Using data-set provided (DVD rental) write python code to output the following object and save it to a json file.

```
{
    "customer_name":"customer_name",
    "address":"address",
    "email":"email",
    payment:[
        "customer_id":"customer_id"
        "staff_id":"staff_id",
        "rental_id":"rental_id"
    ],
    film_section:[
        "title":"title",
        "description":"description",
        "rental_duration":"rental_duration"
    ],
    store_section:[
        "store_id":"staff_id",
        "manager_staff_id":"manager_staff_id"
    ]
}
```

1.b) Mock an endpoint for the above object.

2.Using pandas analyze how many customer are from Egypt, Kuwait, India and return their Name as First Name and Second Name.

3.Create a list of Customers with their payments.

SECTION TWO

1.Using the big-mart data-set provided and Pandas perform Exploratory Data analysis considering the below:

a) Uni-variant analysis.

b) Multi-variant analysis.

c) Bi-variant analysis.

SECTION THREE

1.Using python code write a function that will convert the below dictionary to a csv file.


```
{
 "successful": [
  "30000922",
  "30000910"

 ],
 "id_undetermined": [
  "30000911",
  "30000913"
 ],
 "unsuccessful": [
  {
   "id_number": "30000921",
   "errors": {
    "fee_status": [
```

```
      "\"KNOWN\" is not a valid choice."
     ]
    }
   },
   {
    "id_number": "29000921",
    "errors": {
     "fee_status": [
      "\"KNOWN\" is not a valid choice."
     ],
     "fee_receipt": {
      "0": {
       "registration_section": {
        "9": {
         "sectionNo": [
          "This field may not be null."
         ]
        }
       }
      }
     }
    }
   }
```


SECTION FOUR

Using Resource Monitoring Data provided create recommender engine considering the below:

a) Resource, Task Type and Task Count can predict the time that will be spend with some key resource requirements?**Hint:**Use data-set named *"test_resource_data_q1"*

b) Task Type, Task Count and Time Spend can predict the resources required to perform the task as specified. **Hint:**Use data-set named *"test_resource_data_q2"*

Note:

It will be an added advantage if the model can be provided with an interface for Easy Use.

Submission:

1) Well commented python Model Creation code

2) Prediction for Question 4 (a) and Question 4 (b) test data

3) Explanation of the model and results finding