CNN Digital Challenges: Data Scientist

We’re excited you’re thinking of joining CNN Digital! Our Team, Data intelligence, builds products and features that power strategic decisions about content and audiences. Hundreds of millions of people interact with CNN on a monthly basis, and we’re just getting started building top-notch features and capabilities to help the business understand these users and their relationship to products and content.

We’ve prepared this take-home project for candidates in order to help us get to know one another, and let you show off your data science skills. We expect you to take 4-8 hours on this project total.

Your final deliverable should be a GitHub repo with the following structure:

-- Readme.md (**Do not** mention CNN in this readme file)

-- Part 1 SQL (folder)

----Query 1 (file)

----Query 2 (file)

----Query 3 (file)

-- Part 2 ML (folder)

----Scripts/Notebooks (files)

----Models (folder)

------Final model/s pickled in this folder

----Data

------Requested Output

Please send the link to your GitHub repo to Bryce at [Bryce.Peake@warnermedia.com](mailto:Bryce.Peake@warnermedia.com), and he’ll distribute your code to the members of the interview panel.

# Part 1: SQL

**Deliverable**

You will submit 3 SQL queries based on the information below. These queries should include comments that explain your logic and or any assumptions you are making. Our goal is to evaluate your SQL abilities and explain your assumptions based on past experience.

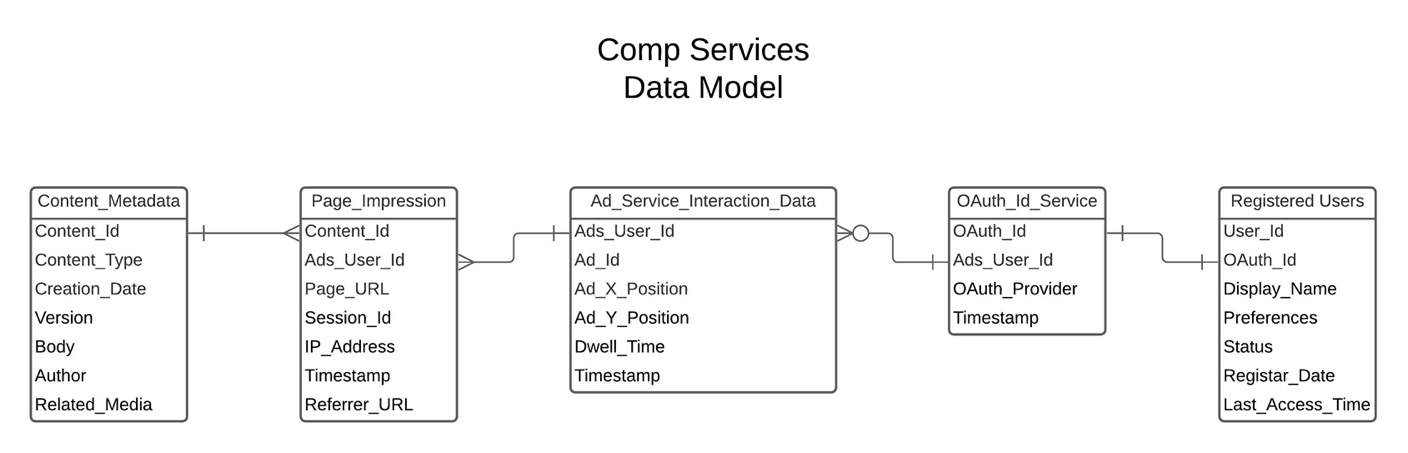
**Scenario**

As a Product Owner, I need insights that will help me to understand user behavior and how to prioritize product features. I would like to explore consumer usage data to determine the following:

* Top 5 most popular pieces of content consumed this week.
* Number of weekly active users for the latest full week (Monday – Sunday). WAU is calculated by counting registered users with > 60 seconds dwell time between Monday-Sunday.
* Top 5 pieces of content from each content type consumed this week by *only* active users (using the above definition).

Now assume the following:

* You have the below SQL database to work with, with datasets in a raw state
* You have the ability to write temp tables and views, but should focus on scalable SQL queries that balance speed and resource usage.



Content\_Metadata (10MM records, 100GB)

|  |  |
| --- | --- |
| **Column Name** | **Data Type** |
| Content\_id | Varchar |
| Content\_type | Varchar |
| Creation\_date | Datetime |
| Version | integer |
| Body | Array (paragraph strings) |
| Author | Varchar |
| Related\_media | Array (content\_ids) |

Page\_Impression (100B records, 15TB)

|  |  |
| --- | --- |
| **Column Name** | **Data Type** |
| Content\_id | Varchar |
| Ads\_user\_id | Varchar |
| Page\_URL | Varchar |
| Session\_id | Varchar |
| IP\_address | Varchar |
| Timestamp | Datetime |
| Referrer\_URL | varchar |

Ad\_service\_interaction\_data (150B records, 22TB)

|  |  |
| --- | --- |
| **Column Name** | **Data Type** |
| Ads\_user\_id | Varchar |
| Ad\_id | Varchar |
| Ad\_X\_Position | Integer |
| Ad\_Y\_Position | Integer |
| Dwell\_time | Float |
| Timestamp | datetime |

OAuth\_ID\_Service (60MM records, 15GB)

|  |  |
| --- | --- |
| **Column Name** | **Data Type** |
| OAuth\_Id | Varchar |
| Ads\_User\_Id | Varchar |
| OAuth\_provider | Integer |
| Timestamp | datetime |

Registered\_Users (80MM records, 17GB)

|  |  |
| --- | --- |
| **Column Name** | **Data Type** |
| User\_Id | Varchar |
| OAuth\_id | Varchar |
| Display\_Name | Integer |
| Preferences | Object |
| Status | Varchar |
| Register\_Date | date |
| Last\_access\_time | datetime |

# Part II

**Deliverables**

* **Notebook** showing your munging and modeling work (with markdown or inline comments)
* **Scores.csv** – a file with the header anon\_person\_id | target | prediction | probability |
* **TWO PowerPoint slides** explaining key insights and findings from the process that you would deliver to a business stakeholder.

**Scenario**

Pretend that the business is trying to identify prospects for a new podcasting app called “PodNN.” You’ve received a file that includes information on whether a sample of our users who have the mobile app listen to podcasts (target). The dataset includes some information about those users, including age, news and streaming subs, and income.

In R or Python, create a proof-of-concept for the business that identifies users who are the best prospects. Explain your approach to data exploration and model selection/evaluation in markdown and/or comments throughout the notebook. Our focus as a team is on reliable statistics and math, and your comments should illuminate this aspect of your work.

**NOTE:** This is a dummy dataset and is not actual user or customer information.

Data Dictionary

|  |  |  |
| --- | --- | --- |
| **ATTRIBUTE** | **DESCRIPTION OF VALUES** | **TYPE** |
| anon\_person\_id | Unique identifier per user; should be unique for each row | Numeric |
| target | Listens to podcast, does not | Boolean |
| state | Timestamp for each event | Categorical |
| presence\_of\_child\_ind | 1 = child residing in home, 0 = not | Boolean |
| streaming\_media\_subs | Y = subscribed to streaming services, N = not | Boolean |
| streaming\_entertainment\_subs | 1 = listens to streaming entertainment, 0 = not | Boolean |
| marital\_status | M = married, S = single, U = unknown | Categorical |
| gender | Male, Female | Categorical |
| income | Represents income bins | Categorical |
| number\_of\_news\_subs | Number of subscriptions to news services, non-digital and digital | Numeric |
| age | Person’s reported age | Numeric |