**Prepared By:** Team Aditya Nikunj

**Abstract:**

The theme of this assignment is based on Domain Fortune 500 Companies, Among the fortune 500 companies, the companies ranking, and revenue is based on the user’s ranking and company’s investments. CEO’s of the individual company play a pivotal role in generating the company’s reputation. Now-a-days the social media presence has been significant in determining the company’s ranking and revenue. In this assignment, we are inferencing the impact of social media presence for the company called the ‘Nerd Analytics’. As a part of database team, we will analyze the producer consumer model for the company’s ranking and revenue. Also, through our data base model we emphasize on providing statisticians team and machine learning experts database model that is cleaned and data gathered from social media. For Fortune 500 Companies we will model the revenue and ranking and analyze the social media impact on it. As a part of model, the Producer are the CEO, Consumers are users and companies are Fortune 500 companies.

The Conceptual Model compromises of entities that have the following:

1. **Producer**

* As a part of Producer Consumer model, the producer is the company’s CEO who is the primarily responsible for generating the revenue. Also, the ranking of the company affects a lot whenever the CEO’s social media presence is either garnered or hampered. We are establishing the model for producers from CEO’s tweets using Tweepy API.

1. **Consumer**

* For Consumer we had used Reddit API, though reddit API we had gathered the user’s data such as the upvotes, user names and their comments on the description. The ranking of the company will be substantially impacted based on the consumer’s (users) views and descriptions. Here, the reddit submission id attribute is uniquely identified as each reddit will have unique submission id and the same can be used as primary key.

1. **Companies**

* For the company data, the source is CSV file which is extracted using the Data Scrapping of website of Fortune 500 company. The data scrapping has been done from the beautiful soup and the detailed information related to company details such as company ceo, revenue, address and ranking related details are extracted.

The amalgamation from different sources has been done to create a conceptual database model.

**Part 1**

1. **Design Requirements**
2. **A Domain**

* Our Domain is Company

1. **Conceptual models (entities) for a tweet/post, a Social Media user, a person and a company**

* Our Conceptual Model includes Tweeter Posts of CEO’s and Companies along with reddit posts of several users

1. **Conceptual Models (entities) that represent consumers, producers and companies in your chosen model**

* In our chosen model following are the representations:

|  |  |
| --- | --- |
| Producers | Twitter Post of CEO’s |
| Consumers | Reddit Users Post’s |
| Companies | Fortune 500 Companies |

1. **Conceptual models (entities) for at least two things specific to the domain (e.g. a game, a film, a song, etc.)**

* **In our conceptual model the two things specific to the domain are:**

|  |
| --- |
| Company Ranking |
| Company Revenue |

1. **Relationships that connect the entities**

* We have established the relationship among several entities using the Primary Keys and Foreign keys

1. **Appropriate attributes and keys**

* All the attribute and keys are appropriately defined and designed for the conceptual model

1. **ER Diagrams that illustrate the entire conceptual model**

* Yes, our ER Diagram illustrates all schema for entire conceptual model

1. **The ER diagrams can use standard ER symbols or UML**

* The diagram illustrates the ER Symbols and relationships.

1. **Questions You Must Answer about Conceptual Model**
2. **What are the ranges, data types and format of all the attributes in your entities?**

* To describe the range, data type and format of all the attributes in your entities we can use the describe function for individual table.
* **Desc Company\_Details\_Table;**

|  |  |
| --- | --- |
| **Attribute Name** | **Type** |
| Name (Primary Key) | Varchar |
| Id | Integer |
| Assets | Integer |
| CEO | Varchar |
| HQ\_Address | Varchar |
| Ranking | Integer |
| Website | Varchar |
| Job Growth | BIT |
| Employees | Number |
| Industry | Varchar |
| Revenue | Varchar |
| Sector | varchar |

* Table: Company\_Details\_Table
* **Desc Company\_CEO\_Table;**

|  |  |
| --- | --- |
| **Attribute Name** | **Type** |
| CEO\_Name (Primary Key) | Varchar |
| Company\_Name (Foreign Key) | Varchar |
| CEO\_Title | Varchar |
| CEO\_Founder | Bit |
| CEO\_Women | Bit |

* Table: Company\_CEO\_Table
* **Desc Company\_Address\_Table;**

|  |  |
| --- | --- |
| **Attribute Name** | **Type** |
| CEO\_HQ\_Address (Primary Key) | Varchar |
| Company\_Name (Foreign Key) | Varchar |
| Company\_HQ\_City | Varchar |
| Company\_HQ\_State | Varchar |

* Table: Company\_Address\_Table
* **Desc Company\_Revenue\_Table;**

|  |  |
| --- | --- |
| **Attribute Name** | **Type** |
| CEO\_Revenue (Primary Key) | Integer |
| Company\_Name (Foreign Key) | Varchar |
| Company\_CEO (Foreign Key) | Varchar |
| Company\_Profit | Integer |
| Company\_Profitable | BIT |

* Table: Company\_Revenue\_Table
* **Desc Company\_Ranking\_Table;**

|  |  |
| --- | --- |
| **Attribute Name** | **Type** |
| CEO\_Ranking (Primary Key) | Integer |
| Company\_Name (Foreign Key) | Varchar |
| Company\_CEO (Foreign Key) | Varchar |
| Company\_Previous\_Ranking | Integer |
| Company\_Employees | Integer |

* Table: Company\_Ranking\_Table

1. **When should you use an entity versus attribute?**

* Here in our conceptual model we have used Company Address as a separate entity as the address of each company will be different having different headquarters at different location, zip code, city and state.
* If we had just used address as an attribute, then it would have been difficult for storing proper details about each individual company and during data cleaning as well.
* For example, to find the number of companies in Boston it is easy to have address as an entity rather than an attribute.

1. **When should you use an entity or relationship, and placement of attributes?**

* Entity Relationship Diagrams is a pivotal data modelling tool and will help organize the data in our model into entities and define the relationships between the entities. The, process has proved to enable the analyst to produce a good database structure so that the data can be stored and retrieved in a most efficient manner.
* **Entity:** A data entity is anything real or abstract about which we want to store data. To store the object details of any real-world data we use entity. For example: In our case Company is an entity.
* **Relationship:** A data relationship is a natural association that exists between one or more entities. E.g. Company have company rankings. **Cardinality** defines the number of occurrences of one entity for a single occurrence of the related entity. E.g. A Company can have many processing in several industries and sectors.
* **Attribute:** A data attribute is a characteristic common to all or most instances of an entity. The attribute placements are based on primary key and foreign key.

1. **How did you choose your keys? What are unique?**

* The Primary Key and Foreign key has been established based on the conceptual schema.
* Company Name will be the Primary Key for the Company\_Details\_Table. Also, it will be Foreign Key for the other tables such as Company\_Revenue, Company\_address, etc.
* All the names of different companies are unique, so it can be easily set as Primary Key.
* The Company\_Address\_Table, Company\_Ranking\_Table have company address and company ranking as primary key as all tuples are unique since attribute is also unique for each company’s address and ranking.

1. **Did you model hierarchies using the “ISA” design element? Why or why not?**

* Our Model doesn’t have an ISA Relation. The ISA Relation would have been established had the Domain Company would have been segregated into different sub companies based on the company industries and sectors.
* For eg: A table created dedicatedly only for Finance group, Technology group etc. then we could have established the ISA Relation such as Finance Group Company is a Company, Technology Group Company is a company.

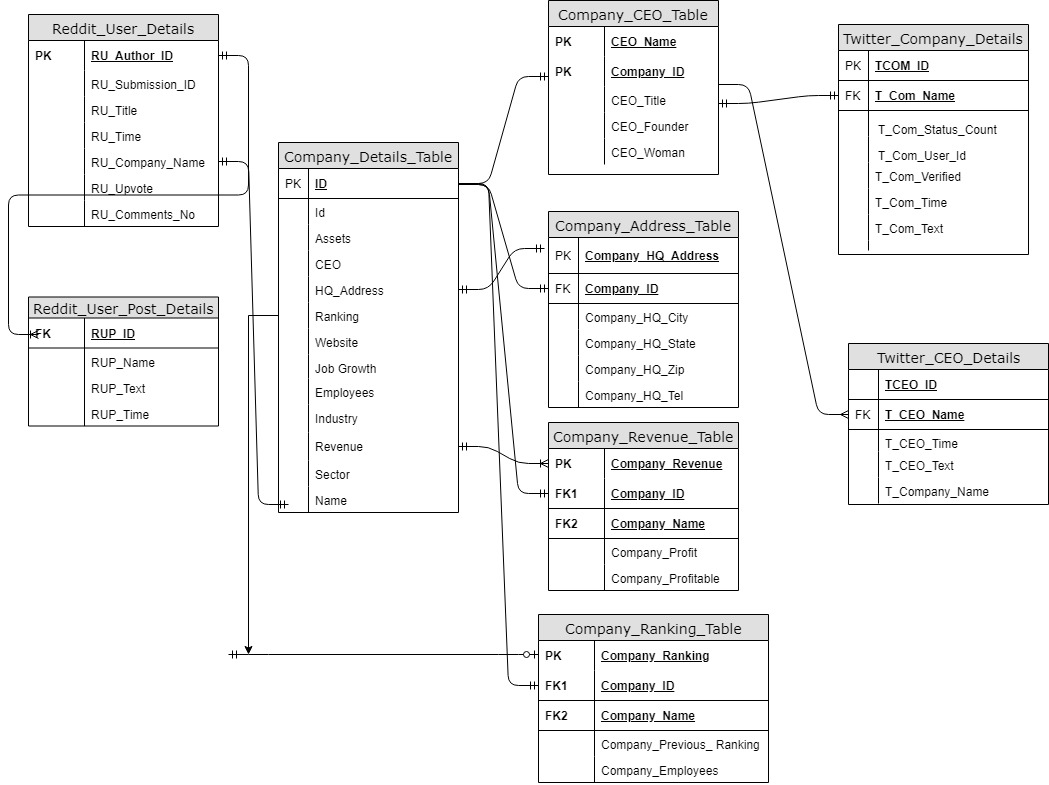
1. **Were there design alternatives? What are the tradeoffs: entity vs attribute, entity vs relationship, binary vs. ternary relationships?**

* There were no design alternatives and for the tradeoff’s also we were able to justify the entity vs attribute for the Company\_Address entity, also entity vs relationship has been indicated during company details in foreign key and for binary vs ternary relationship, we have attributes in several entities like Company\_Profiftable, CEO\_women where we get proper binary data such as yes or no.

1. **Where are you going to find real-world data to populate your model?**

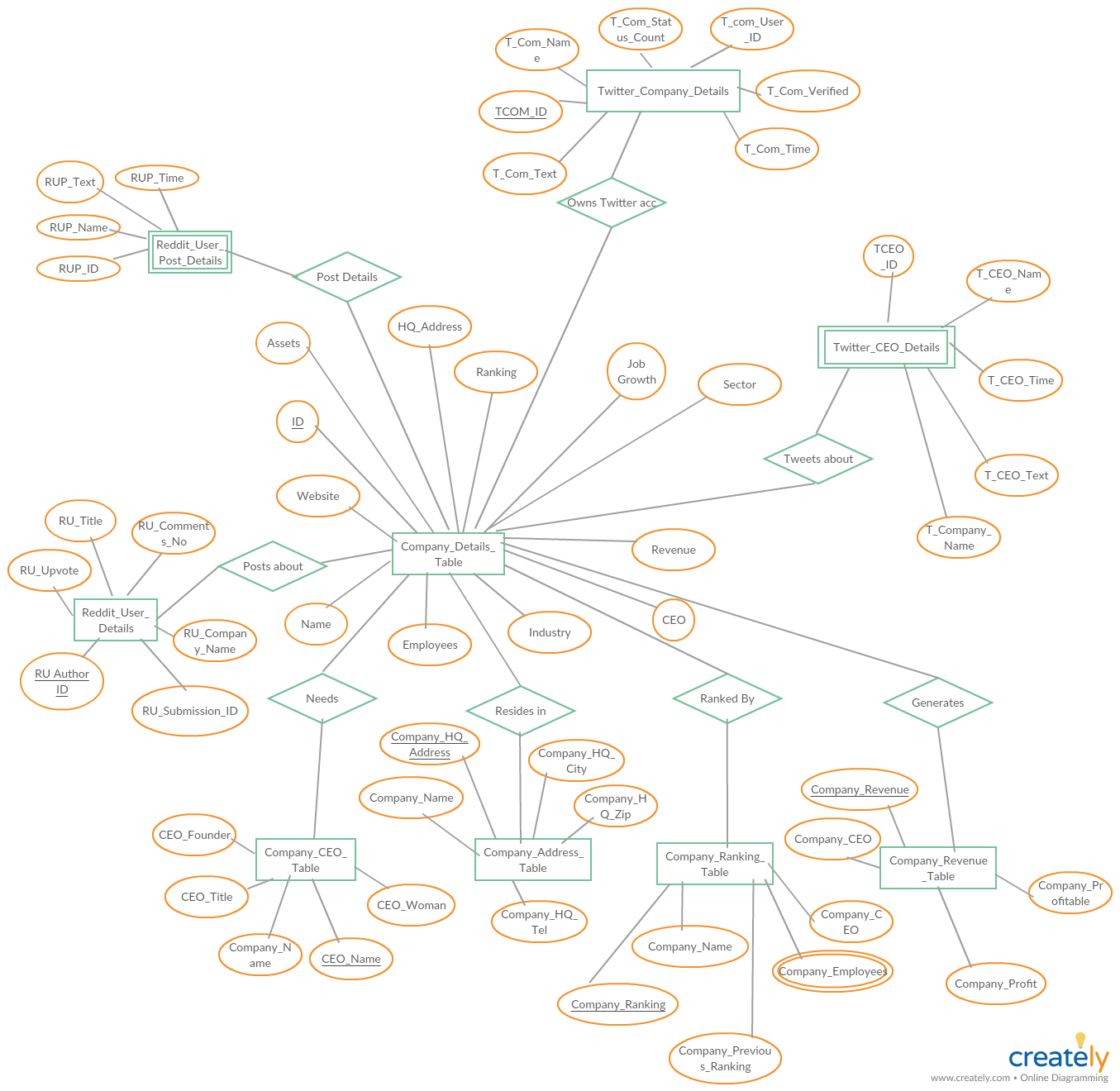
* The real-world data to populate our model will be available on fortune 500 company’s website where the company’s data will change continuously also the User’s comments and CEO’s comments for the company will be posted gradually.

**Conceptual Diagram**



**Figure: 1**

**E-R Diagram**

****

**Figure: 2**

**Contributions**

* **Tweepy API – Aditya Kamat**
* **Reddit API – Nikunj Doshi**
* **CSV Data Extraction – Aditya Kamat**
* **Data Cleaning – Aditya Kamat**
* **SQL Queries – Aditya Kamat, Nikunj Doshi**
* **Report Preparation – Nikunj Doshi, Aditya Kamat**
* **GitHub – Aditya Kamat, Nikunj Doshi**

**Conclusions**

* We were able to establish a relationship for the model and show the relationship for the domain etc.
* The Producer and Consumer were used from social media API’s, Producer were shown in form of Twitter API’s and Consumer were shown in form of Reddit API for the users. The extracted data from twitter and reddit API were used to extract the status count, comment count, title, user id, company text.
* Through the conceptual schema, and the ER diagram we have developed the model for different companies name we were able to relate the features of companies and different social media users. Also, we were able to clean the data from the csv files and eliminated all duplicate and null values.

**Citations**

* **For WEB API**
* <https://stackoverflow.com/questions/47925828/how-to-create-a-pandas-dataframe-using-tweepy>
* <https://tweepy.readthedocs.io/en/v3.5.0/code_snippet.html>
* <https://www.reddit.com/dev/api/>
* **For Web Scrapping**
* <https://www.bing.com/videos/search?q=web+scrapping+youtube&view=detail&mid=DA3EB1EAE99DC9ADBD97DA3EB1EAE99DC9ADBD97&FORM=VIRE>
* <https://github.com/nikbearbrown/INFO_6210/blob/master/Week1/NBB_%20BeautifulSoup.ipynb>
* <http://fortune.com/fortune500/>
* **For SQL Lite**
* <https://www.sqlite.org/docs.html>
* <https://www.programmableweb.com/api/reddit>
* **For CSV File**
* <http://pandas.pydata.org/pandas-docs/version/0.17/generated/pandas.DataFrame.drop_duplicates.html>
* <https://stackoverflow.com/questions/9785049/python-how-to-use-json-dumps-on-windows>
* <http://book.pythontips.com/en/latest/enumerate.html>
* <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.loc.html>
* <https://stackoverflow.com/questions/14037540/writing-a-python-list-of-lists-to-a-csv-file>
* <https://github.com/nikbearbrown/INFO_6210/blob/master/Movie_DB_Example/TMDB_Movie_Data_Assignment_Example.ipynb>
* <https://stackoverflow.com/questions/34682828/extracting-specific-selected-columns-to-new-dataframe-as-a-copy>
* **For E R Diagram File**
* [**http://web.cs.ucdavis.edu/~green/courses/ecs165a-w11/2-er.pdf**](http://web.cs.ucdavis.edu/~green/courses/ecs165a-w11/2-er.pdf)
* [**http://users.csc.calpoly.edu/~jdalbey/205/Lectures/HOWTO-ERD.html**](http://users.csc.calpoly.edu/~jdalbey/205/Lectures/HOWTO-ERD.html)

**Licence**

Copyright 2019 Nikunj Doshi, Aditya Kamat

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.