Proposal for NoSQL Project

1. **Team member in alphabetical order of the last name:**
   > Sapna Khatri, Omkar Shinde

2. **What problem you want to work on?**
   > Impact of COVID-19 on employment and its relation to layoffs.

3. **Which NoSQL database you want to use?**
   > We are going to use MongoDB.
   > Our data is in Tabular format, we will convert it to JSON.

4. **Which public dataset will you use? Provide the URL of the dataset. Does the dataset have at least 20 attributes and 10,000 samples?**
   > There are 20+ attributes and 40k + samples
   > Need to combine multiple datasets for better analysis-
   > https://www.kaggle.com/datasets/swaptr/layoffs-2022
   > https://www.uslayoffs.org/home
   > https://layoffdata.com/data/
   > https://layoffs.fyi/
   > Covid Dataset- https://github.com/nytimes/covid-19-data (2020 to 2023)

5. **Describe the N+1 non-trivial data analysis tasks you plan to design and implement. Do they extract new and valuable information from the raw data in the dataset? Why are the tasks non-trivial?**
   > We are a team of 2 so in total 3 analysis-
   Following are a few non-trivial data analysis tasks which we plan to take up

   1. Relationship between funds received to a company, covid cases and layoff during that time.
   2. Relation between companies that laid off x% of employees and covid cases in those regions.
   3. Comparison of layoffs to Previous Years after covid impact within a company/industry (% layoff).
   4. Year wise layoffs according to state, industry and covid cases. (Extra- Place holder)

- The first analysis involves examining the relationship between the funds received by a company, covid cases, and layoffs during that time. This analysis requires a comprehensive understanding of the financial data of the company, the number of covid cases in the area, and the number of layoffs that occurred. It also involves determining the causal relationship between these variables.
- The second analysis involves examining the relationship between the percentage of employees laid off by a company and covid cases in the region. This analysis requires an understanding of the demographic and economic characteristics of the region, as well as the industry and size of the company. It also involves determining the causal relationship between these variables, which can be complex and multi-directional.
- The third analysis involves comparing the layoffs that occurred within a company or industry after the covid impact to previous years. This analysis requires a thorough understanding of the historical data on layoffs and industry trends, as well as the specific factors that influenced the covid-related layoffs. It also requires identifying any confounding attributes that may have influenced the results.
- The fourth analysis involves examining the year-wise layoffs according to state, industry, and covid cases. This analysis requires a comprehensive understanding of the data on layoffs, covid cases, and industry trends over time. It also involves identifying any patterns or trends in the data and determining the factors that may have influenced these trends.

6. **Why is it doable in this semester? Describe your plan, strength, and familiarity with the language/tool of your choice?**
   > We can complete this analysis in this semester with careful planning, diligent work, and efficient use of resources.

   Plan:

   1. Collect data from sources mentioned above (government agencies, Kaggle, git)
   2. Clean and process the data, identify any missing data, normalize and transform the data into a suitable format of analysis.
   3. We will use Python and MongoDB for analysis and matplotlib for any visual representation.
   4. Using these tools, we will identify patterns or trends in the data and determine the factor that may have influenced these trends.
   5. Draw conclusions and make recommendations on this analysis.

   Strength:

   - Omkar is a senior software developer with 4 years of experience in Oracle, Shell, Python and has been continuously working with high volumes of data. He has extensive experience in the ETL process and a deep understanding of data flow in the Beverage/Alcohol pricing industry, which has equipped him with a strong skill set in data manipulation, data warehousing, and database management.
   - Sapna is a Senior Software Engineer with 2 year of experience in full stack development. She has experience with Microsoft SQL Server and was responsible for statistical analysis on user data. She was also responsible for handling data redundancy and null value. Not only she handled data management but also made sure to validate all the constraints to avoid future risk.

   Language and Tools:

   - Python (pymongo, json)
   - MongoDB (MongoDB Atlas cluster)
   - Pandas
   - Matplotlib
   - Tableau