ETL Project Proposal (Group: ETL too hard)

# Question to be answered:

1. The correlation between unemployment rate and the amount of job postings (Trendline)
2. Find the most promising/growing industry in 2016

Ratio1= job postings / unemployment rate

Ratio2= job posting incremental rate per month

# Data used:

1. 2016 US job postings on Dice.com (Source from Kaggle, approximately 22000 rows) <https://www.kaggle.com/PromptCloudHQ/usbased-jobs-from-dicecom>
2. US unemployment rate by county (1990-2016) (Source from Kaggle) <https://www.kaggle.com/jayrav13/unemployment-by-county-us/data#output.csv>
3. US state acronym -- Provided
4. Career industry list -- Provided

# Data cleaning (Uncompleted):

## Job postings cleaning:

1. Clean up and strip the location column
2. Remove non-US jobs
3. Group the dates by month
4. Group the industry by job titles

## Unemployment rate table cleaning:

1. Group by State and link with job postings
2. Take the average unemployment rate
3. Link with the ‘US state acronym list’ (no dataset found yet)
4. Remove all years other than 2016

# Database:

1. PostgreSQL