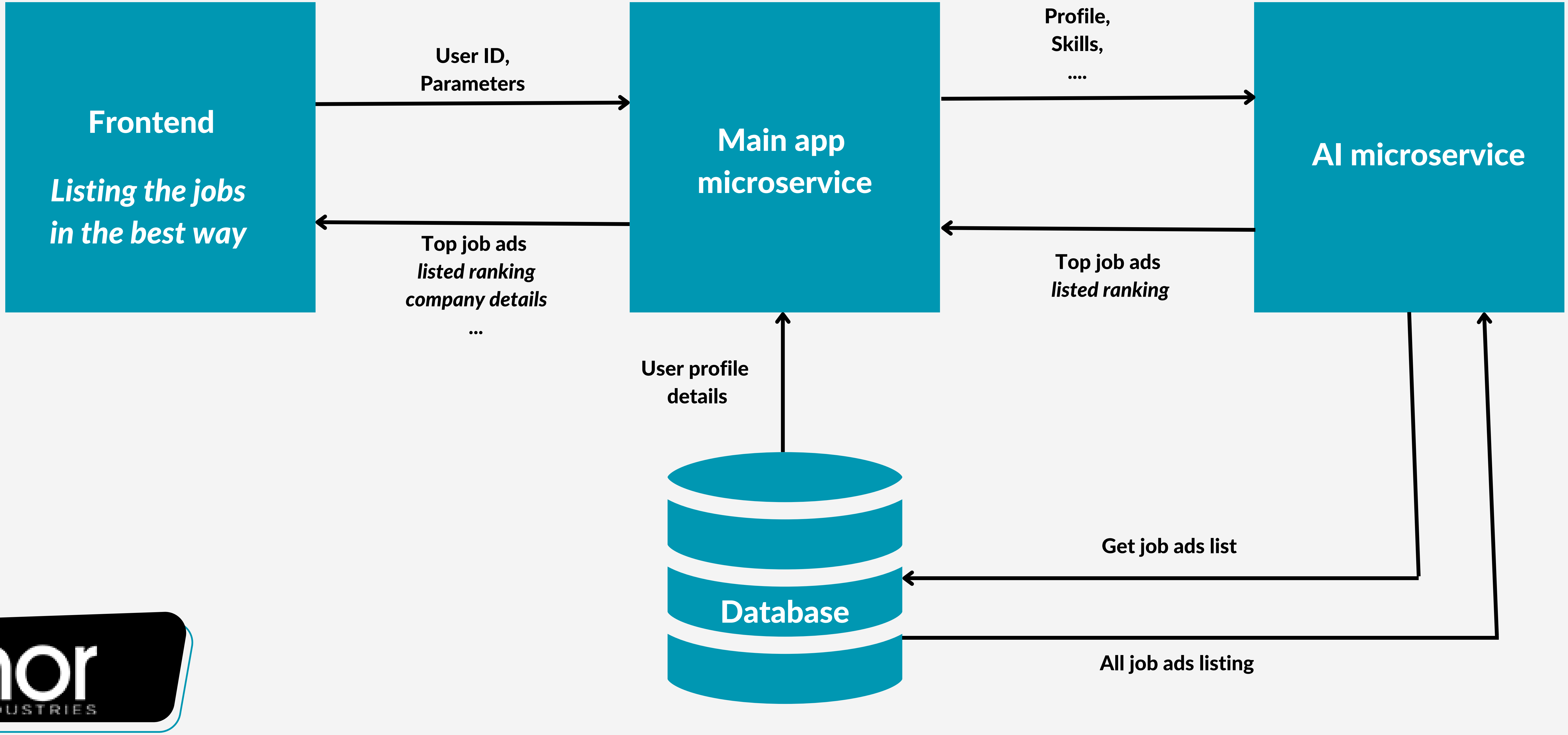
Job Recommendation module - System design

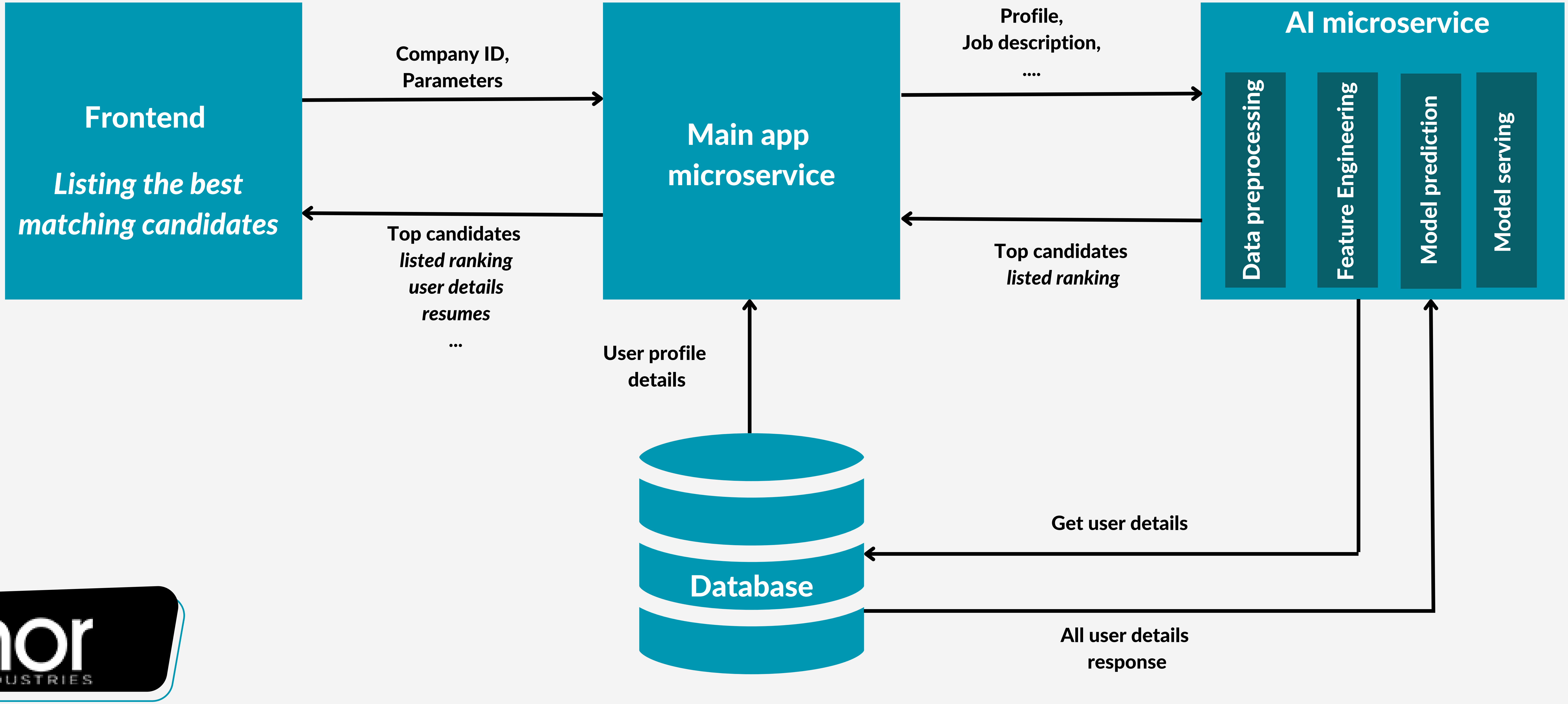
thor

INDUSTRIES

1.Job seeker scenario



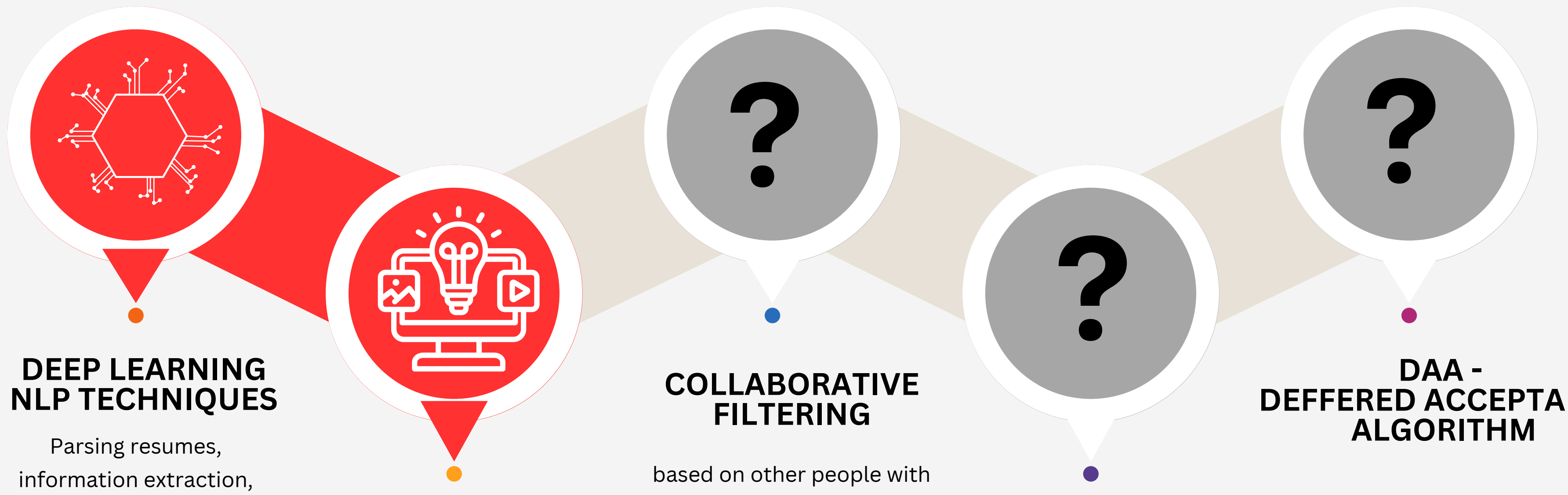
2.Company scenario



Benefits of this approach

- Independently built, tested and deployed
- Knows everything about the users/companies
- Parses resumes
- Parses job descriptions
- Runs recommendation tasks fast
- Scalable level: medium

Where we at?



DEEP LEARNING NLP TECHNIQUES

Parsing resumes, information extraction, recommending based on similarity metrics through weighted sum approach

CONTENT BASED FILTERING

based on content-based recommendation strategy where recommendations are generated based on (similar) jobs the user has previously browsed

COLLABORATIVE FILTERING

based on other people with similar features and behaviours, the model will make recommendation analysing and identifying such patterns

CONTENT BASED + COLLABORATIVE FILTERING

combining both content-based filtering and collaborative filtering method and as a final recommendation model will be a hybrid model

DAA - DEFFERED ACCEPTANCE ALGORITHM

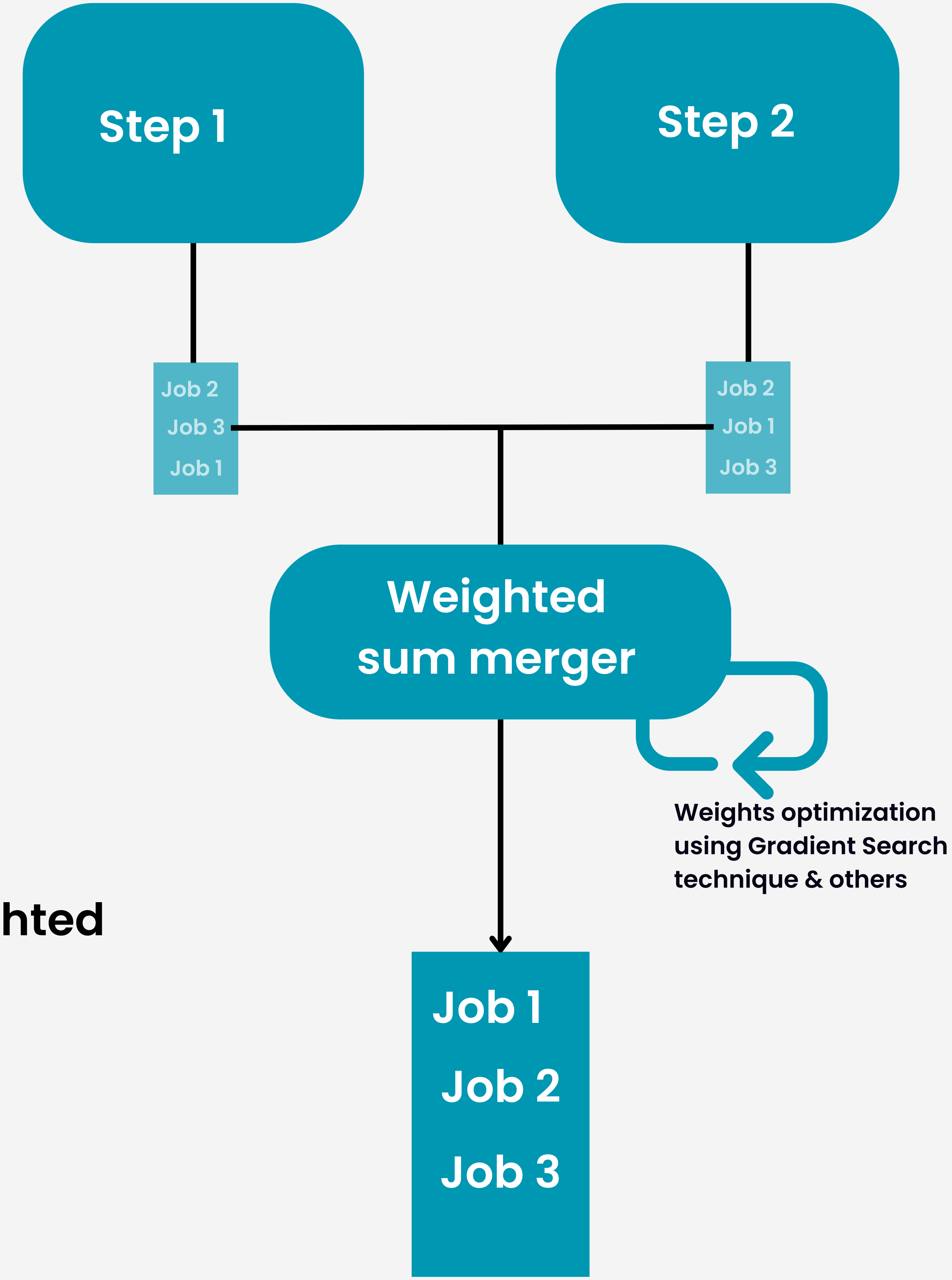
Content based filtering

Two-step recommendation

1. Using NLP techniques which involve:
   ->TF-IDF and Word2Vec Based Job Matching
   -> Similarity metrics: cosine similarity, jaccard and euclidean
   ->Threshold value optimization

2. Using NLTK module and Topic modelling Classification technique

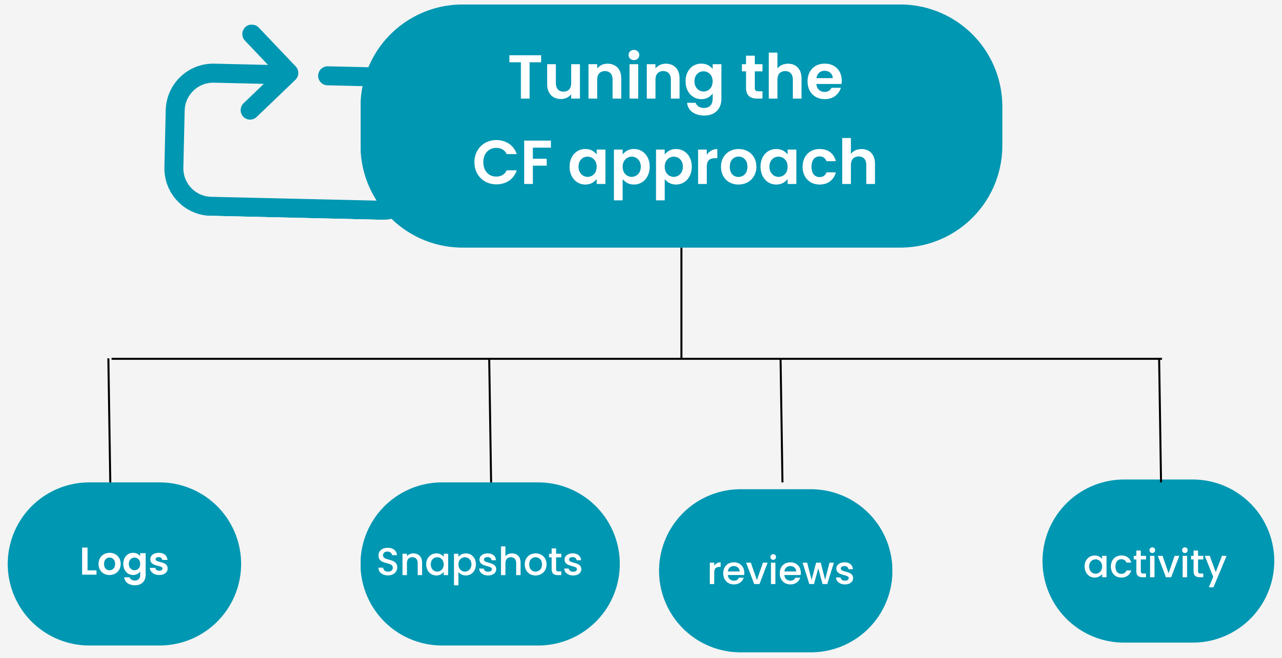
Final ranking list will be using a merging technique by applying weighted sum approach.



Tuning on metrics

After some time of work from the Recommender System, we can enhence the work, once more data and more users come into the system.

An idea of tuning the model and our techniques will include to monitor the performance of the system by different timelines and get:
-> logs from the activities of the users
-> snapshots from the users
-> reviews and feedback which can be manually or using any survey



Data

**After analysis of datasets that we can use I came up with 3 possible datasets,
which are with good reviews and used in different papers:**

https://www.kaggle.com/datasets/PromptCloudHQ/us-technology-jobs-on-dicecom -

-> this is a pre-crawled dataset, taken as subset of a bigger dataset (more than 4.6 million job listings) that was created by extracting data from Dice.com, a prominent US-based technology job board.

This dataset has following fields:

- advertiserurl
- company
- employmenttype_jobstatus
- jobdescription
- joblocation_address
- jobtitle
- postdate
- shift
- skills

Data

https://www.kaggle.com/datasets/shrutiambekar/46000-indian-companies-information-dataset

This dataset contains information about companies. The data was scraped using web scraping techniques and includes variables such as company name, industry, location, employee count, ratings, and reviews.

Source: The csv data was scraped from https://www.ambitionbox.com/

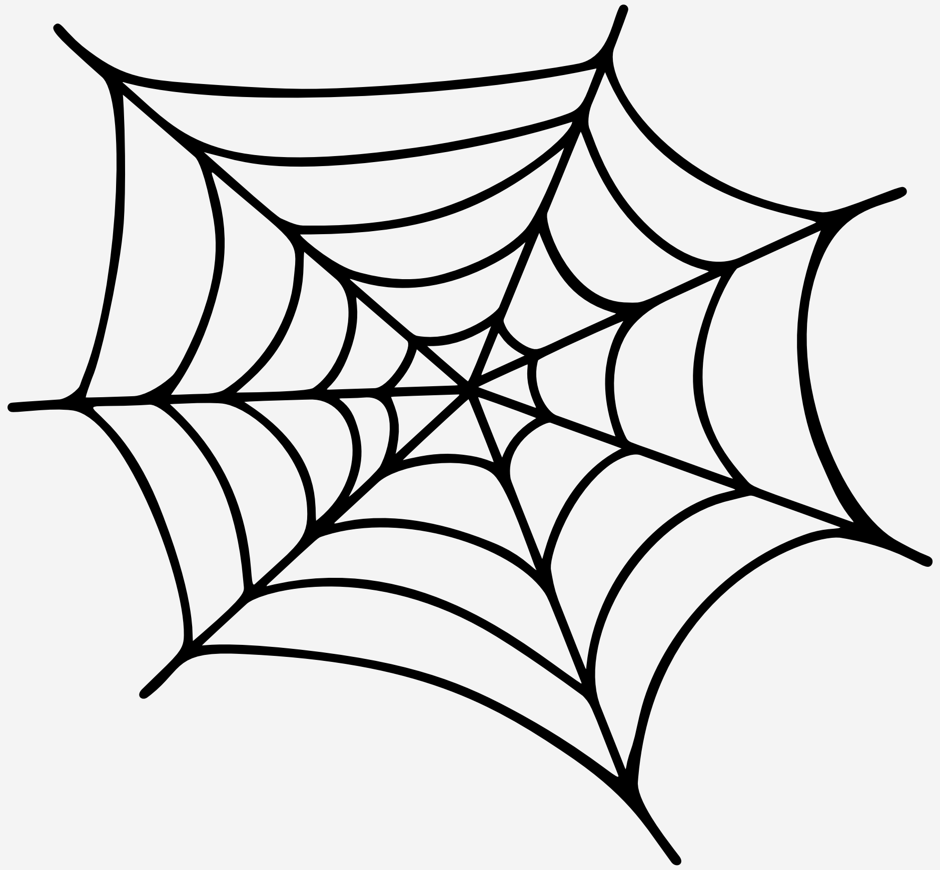
https://www.kaggle.com/datasets/stackoverflow/stack-overflow-2018-developer-survey

Each year, we at Stack Overflow ask the developer community about everything from their favorite technologies to their job preferences. This year marks the eighth year we've published our Annual Developer Survey results—with the largest number of respondents yet. Over 100,000 developers took the 30-minute survey in January 2018.

This year, we covered a few new topics ranging from artificial intelligence to ethics in coding. We also found that underrepresented groups in tech responded to our survey at even lower rates than we would expect from their participation in the workforce. Want to dive into the results yourself and see what you can learn about salaries or machine learning or diversity in tech? We look forward to seeing what you find!

Data

For better quality of model results, we can use web scrapping techniques in our own to try it with local sites in our country, so that the model can learn also cyrillic fonts and adapt the best results and accuracy for use cases in our country.



Technologies

-Python 3.10
-Flask
-XGBoost
-NLTK
-Word2Dec
-Huggingface
-Tokenization