Kevin Kao
Harshavardhan Suresh
Divya Vishnumurthy

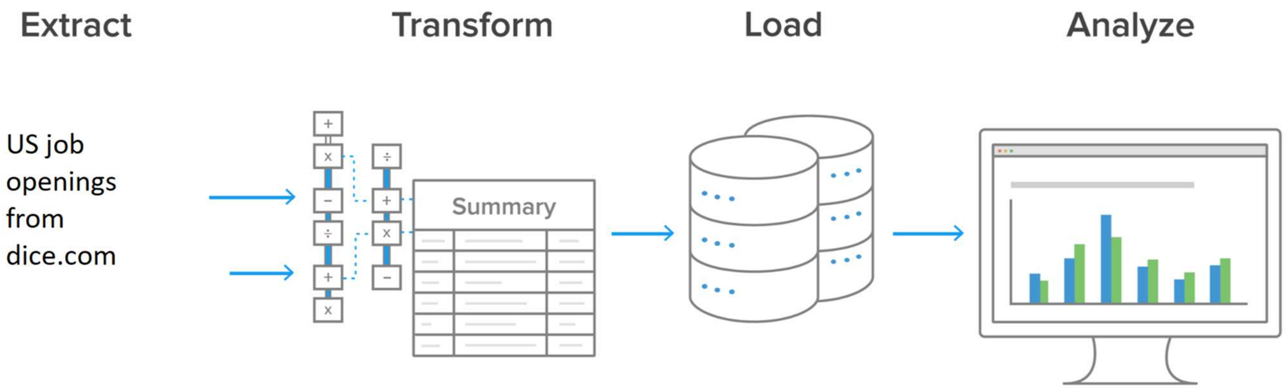
ETL Project

US Technology jobs, 2019

ETL is an important part of today's business intelligence (BI) processes and systems. It is the IT process from which data from disparate sources can be put in one place to programmatically analyze and discover business insights.

In our project we were interested in understanding the job market for data analysts. We pulled data from dice.com where all the technology job openings are posted. We then wanted to check if a certain city, state had more requirements than others. Also, what are the top three skills that companies are looking for. Our transformed and loaded data can also answer questions like: if there is a relation between job title and travel. If a job title allows to telecommute and what % of time, list of companies that look for certain skills and where they are located.

We used both SQL and NOSQL Databases to load our data based on the data type we were transforming.



01

Data Extraction: Kaggle dataset- US Technology job openings from dice.com

02

Data Transformation: Python pandas and dataframe, dictionaries and string manipulations.

03

Data Loading: SQLAlchemy, pymongo, SQL-Postgres and NOSQL-MongoDB.

About the Dataset

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. The data set contains 22K records.

Columns in the dataset	Fields of interest
advertiserurl company employmenttype_jobstatus jobdescription jobid joblocation_address jobtitle postdate shift site_name skills uniq_id	company jobdescription joblocation_address jobtitle shift skills uniq_id

Challenges

1. Find a dataset which fit the requirements.
2. The irrelevant data had to be filtered out as needed and the comma separated skills had to be split.
3. Jobaddress_location – Contained multiple elements separated by a comma and in some records, they had entire address. We had to clean and transform the data into two columns containing city and state only.
4. Shift column- Contained multiple elements separated by a "|" and some rows were empty. We had to clean those and transform them into two dataframe columns – Travel and Telecommute.
5. Job Description column- Contained a paragraph of the job requirement and this field is more like a document. We tried to fit it onto SQLDB, but it fit better in NOSQL DB.
6. The skills data was gnarly and did not have the values in a uniform format. While the column contained skills as comma separated values, there were also values that were either referencing columns or contained the job description rather than the skills required.
7. Once the data was cleaned up, a judgement call had to be made to confirm on the type of database (SQL vs NOSQL). Given that the number of skills varied, we chose mongo DB, NOSQL as it was easier to reference the skills and pull out the relevant job details.