Revelio Labs Customer-Facing Data Scientist Assignment

The following assignment is adapted from a real client request we recently had at Revelio Labs. We recommend reading the assignment in its entirety before beginning any of the tasks. Please note that no actual data is provided in this assignment, but you are provided with the structure of the data. You only need to write the SQL code according to the table schemas. You are not meant to run the code, so minor errors are ok. Feel free to use any resources you want (e.g. Stack Overflow), and please don't hesitate to email us if anything is unclear!

Background: data and models

Within the Redshift cluster, the following important tables exist:

Table 1: position_dedup

Each row corresponds to a single piece of work history parsed from a resume or online job profile and mapped onto our taxonomies. It contains the following relevant columns:

- user_id: a unique integer ID for the user who holds the position
- position_id: a unique integer ID for the position
- company_id: a unique integer ID for the company at which the position was held
- title: position title (raw from profile)
- mapped_role: the best fit for the position out of 1500 job categories (see here for more on how this mapping works!)
- msa: the MSA in which the position was held
- startdate: the start date of the position, rounded to the nearest month
- enddate: the end date of the position, rounded to the nearest month (null for currently active positions)

user_id	position_id	company_id	title	mapped_role	msa	startdate	enddate
1	1235	3	Real Estate Salesperson	Salesperson	New York, NY	2014-06	2020-08
1	1236	1	Pharmaceutical Sales Rep	Salesperson	Ann Arbor, MI	2018-06	NULL
2	2536	4	Software Engineer	Software Engineer	San Francisco, CA	2019-05	2019-12
2	2537	6	Software Engineer II	Software Engineer	San Francisco, CA	2020-05	NULL
	…	…		…	…	…	…

Table 2: education_dedup

Each row corresponds to a single piece of education history parsed from a resume or online job profile and mapped onto our taxonomies. It contains the following relevant columns:

- user_id: a unique integer ID for the user with that education
- school: name of school
- degree: education degree pursued/being pursued
- startdate: the start date of the education, rounded to the nearest month
- enddate: the end date of the education, rounded to the nearest month (null for currently active education)

user_id	school	degree	startdate	enddate
1	Sunnyside High School	High School	2010-09	2014-06
2	Stanford University	Bachelor	2014-09	2018-05
2	Columbia University	Master	2018-09	2020-05
	…		…	…

Table 3: predicted_salaries

Revelio Labs has developed a model to predict salaries for positions (you can read about it here!). The model predicts a salary for each unique combination of company, mapped_role, msa, and year (between 2008 and 2022). Therefore the predicted_salaries table contains the following columns:

- company: the company for which the salary is being predicted
- mapped_role: the mapped role for which the salary is being predicted
- msa: the MSA for which the salary is being predicted
- year: the year for which the salary is being predicted (between 2008 and 2022)
- salary: the predicted salary

company_id	mapped_role	msa	year	salary
2	Salesperson	San Diego, CA	2014	124537.18
3	Software Engineer	New York, NY	2009	129415.74
…	…	…	…	…

Table 4: scaling_weights

The positions in our data are gathered from resumes and online job profiles and do not make up a representative sample of the global labor market. For instance, not everyone has an online job profile, and someone's likelihood of having one depends heavily on what their job is. Revelio Labs has developed a scaling model to counteract the effects of this sampling bias (you can read about it here!). This scaling model allows us to assign a weight to each position so that positions for roles that are underrepresented in our data are weighted more heavily. For instance, 90% of software engineers are covered by our data, but only 10% of truckers are. So every time we see a software engineer at a company, we'll count them as 1.11 software engineers. But every time we see a trucker, we'll count them as 10 truckers. Therefore the scaling_weights table contains the following columns:

- mapped_role: one of 1500 job categories (corresponds to mapped_role in position_dedup)
- weight: scaling weight for the role (i.e. each time we see a position for this role, we count it as *weight* positions)

mapped_role	weight
Software Engineer	1.11
Truck Driver	10.00
Teacher	2.15
…	…

Table 5: company_ref

This table contains various identifiers for each company that we track. It contains the following columns:

- company_id: the unique ID for the company
- name: the name of the company
- website: the url of the company's website
- isin: the ISIN identifier for the company's primary security

company_id	name	website	isin
1	Revelio Labs	reveliolabs.com	NULL
2	Apple	apple.com	US0378331005
3	Netflix	netflix.com	US64110L1061
…	…	…	…

Tasks

Imagine you received the following email from a client:

Hi Revelio Team,

We've recently been investigating salaries at tech companies and we're interested in using your data to explore how salaries have changed at different companies over time. In particular we're interested in receiving monthly timeseries data that shows how the average salary has changed for different companies at the MSA level. I've attached a file with around 1000 companies that we're interested in looking into.

<Attachment: requested_companies.csv>

Task 1:

Given what you know about the Revelio data, models, and tech stack, write some SQL code that will create a table containing the requested data. We're looking to see you get creative about how you can meet their needs!

Note 1: *You can assume the attached requested_companies.csv file has been loaded into a table on redshift and is called "client_requested_companies". When the final table you create is ready on redshift, assume you can unload it to Amazon S3 as a .csv file and provide the link to the customer.*
Note 2: *In your response, please provide the complete email communication that you would have with the customer, explaining the dataset you are delivering to them.*

Task 2:

Imagine that upon delivering the data, you receive the following reply:

Hi Revelio Team,

Thanks for the quick turnaround – so far the data has proved to be extremely useful. However, we would like to differentiate between contingent workers (part-time / interns) and full-time workers. We believe that by including contingent workers, the average salary is much lower than expected. Would you be able to split out these two types of workers in the calculation?

Please discuss in detail the steps you would take to identify contingent workers and include SQL code you would need to write. The output should be the same time series as in Task 1 but now separating out contingent/part-time vs full-time workers. In other words, for this task the time series will be by company, MSA, employee_type, and month. Remember to state any assumptions you make.

Note: *Similar to Task 1, please provide the complete email communication that you would have with the customer, explaining the dataset you are delivering to them.*