Proyecto de ETL - 2da Parte

Kevin Santiago Artunduaga Vivas

Universidad Autónoma de Occidente

Cali, Colombia

Extract API

La API de la cual hice uso fue de Rapid API llamada "Job Salary Data" que sirve para obtener estimaciones de salarios/pagos laborales de las principales páginas de ofertas laborales como Payscale, Glassdoor y ZipRecruiter. Hice uso de esta API para poder complementar los salarios de los de los cuales no disponía información, cabe recalcar que no pude hacer el extract de todos los trabajos debido a que la API cuenta con un límite de consultas así que los que extraje fue de los trabajos más recurrentes que tenía en mi base de datos "datajobs_glassdoor" para ello primero hice un consulta en el query tool de pgAdmin con el siguiente query:

```
SELECT header_jobtitle, COUNT(*) as job_count
FROM datajobs_glassdoor
GROUP BY header_jobtitle
ORDER BY job_count DESC;
```

Recorte de como salio la consulta para los primeros 8 trabajos:

	header_jobtitle character varying	job_count bigint
1	Project Manager	4340
2	Software Engineer	3250
3	Business Analyst	2429
4	Product Manager	2125
5	Data Scientist	1955
6	Data Analyst	1752
7	Data Engineer	1506
8	Senior Software ...	1070
9	DevOps Engineer	813

Total rows: 1000 of 66214 Query c

Luego de esta consulta tomé 743 trabajos, de los cuales use para que la API hiciera la búsqueda de estos. Lastimosamente aunque la API tuviera una cantidad de consultas gratis tenía que ser cuidadoso con la cantidad de trabajos que buscará. Ahora bien para el código tenía que definir la localización de donde iba a tomar los trabajos, para ello preferí basarme en USA (United States) debido a que casi la mayoría de trabajos que tengo los tengo escritos en inglés y hay varios que provienen de ahí.

A pesar de que parecieran una poca cantidad de trabajos la API lo que hace es que toma 3 páginas de búsqueda de ofertas laborales y busca ese titulo en esas paginas, asi que como resultado por 1 trabajo arroja información a partir de Glassdoor, Indeed y Payscale. Por ende para la extracción quede con una cantidad de filas de 2229.

Para el código lo que hacía era poner los trabajos en una lista separados con una , y lo que la API tomará de esos trabajos los iba agregando a una lista para posterior a ello guardarlo en un csv llamado job_salaries.csv.

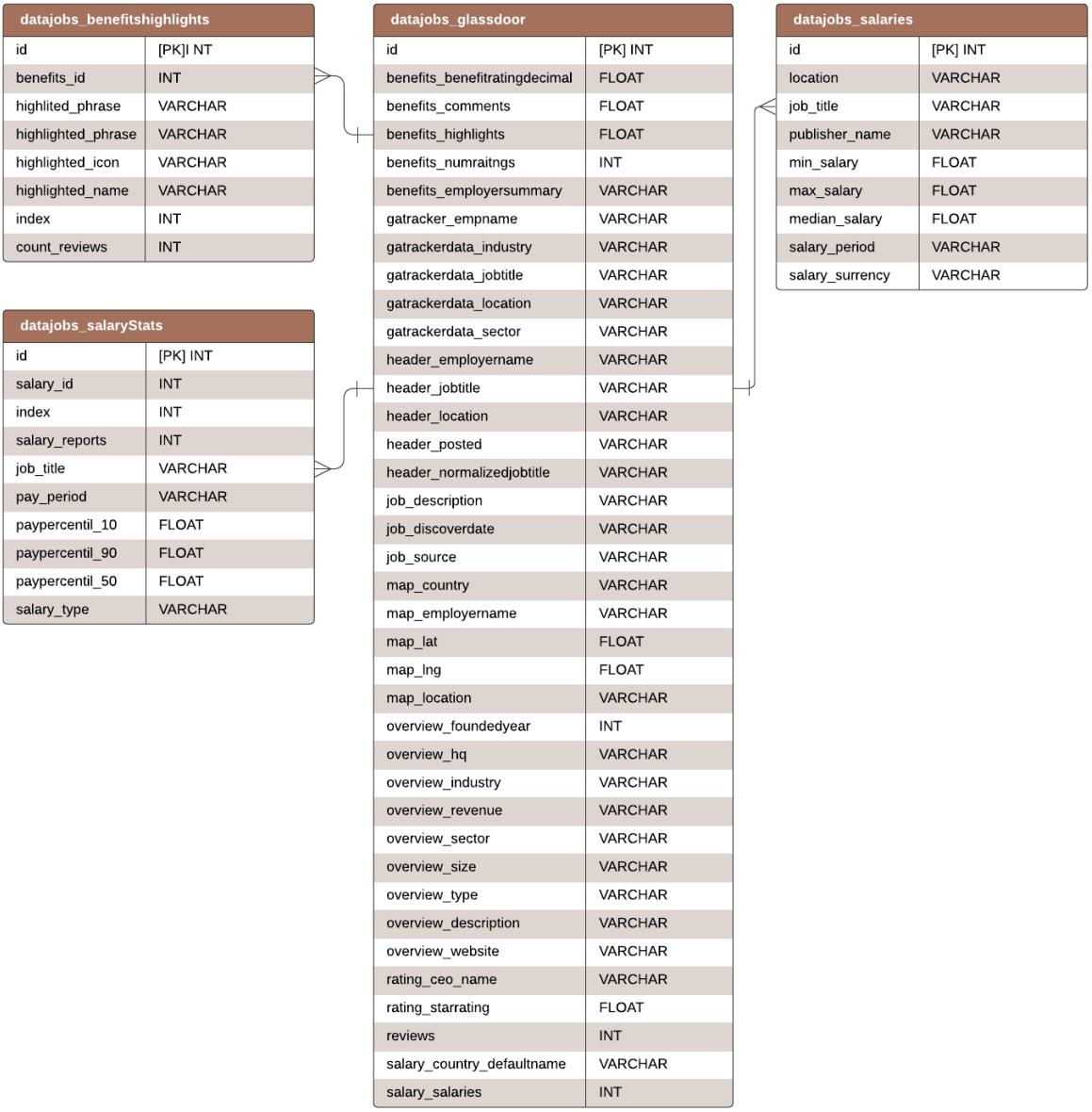
Transform

Para las transformaciones del data warehouse final yo tenía preparadas dos tablas que pude traer de Kaggle que se supone que estaban en conjunto con la principal, ahora bien estas dos tablas las tome como mis 2 dimensiones en donde una hacía parte de los beneficios destacados y otra de los salarios de los trabajos. Ahora bien a la hora de yo querer hacer mirar la tabla de datajobs_salaries contaba con algo raro y es que al hacer la limpieza de esa tabla me quedaron 338.391 filas y mi dataset principal cuenta con solo 165.290 filas, es decir el dataset de salarios posee id duplicados en donde al yo querer hacer la revisión de esto pude notar que en esos id duplicados hay distintos títulos de trabajos algo bastante peculiar porque en datajobs_glassdoor no hay ni un id que hace referencia a la tabla de trabajos duplicado, así que tuve que abandonar la idea de tomarla como una decisión, ya que no supe de manera está conectada la tabla de salarios con la principal.

Ahora bien para la dimensión de benefitsHighlights al parecer esta si contaba con la conexión de los id correcta, pero a pesar de que esta cuente con una amplia cantidad de filas 159.800 no todos los trabajos tienen los beneficios destacados sino que un trabajo puede contar con varios beneficios para ser más exactos hice un query que me revisara esto y la cantidad de benefits_id únicos en datajobs_benefitshighlights es de 33.034, esto aunque parezca una cantidad poca en realidad es bastante porque es información relevante e importante de a la hora uno mirar una oferta laboral.

Aunque parezca que la dimensión de datajobs_salaries no es la mejor por no contener los id adecuados para la conexión de las tablas esta puede conectarse a partir de los títulos de trabajo parecidos, porque esta contiene una columna de job title en donde puede ser tomada como valor para referenciar los trabajos que estén en el dataset principal, para ello le cambie el nombre a esta y la llame datajobs_salaryStats porque lo más importante es que cuenta con las medidas estadísticas de los percentiles de los salarios y la informacion que saque a partir de la API la llame como datajobs_salaries, ahora bien para los datos que me saco la API a a hora de crear la tabla en Postgresql elimine la columna de publisher_link debido a que esta no es necesaria.

Tables schema

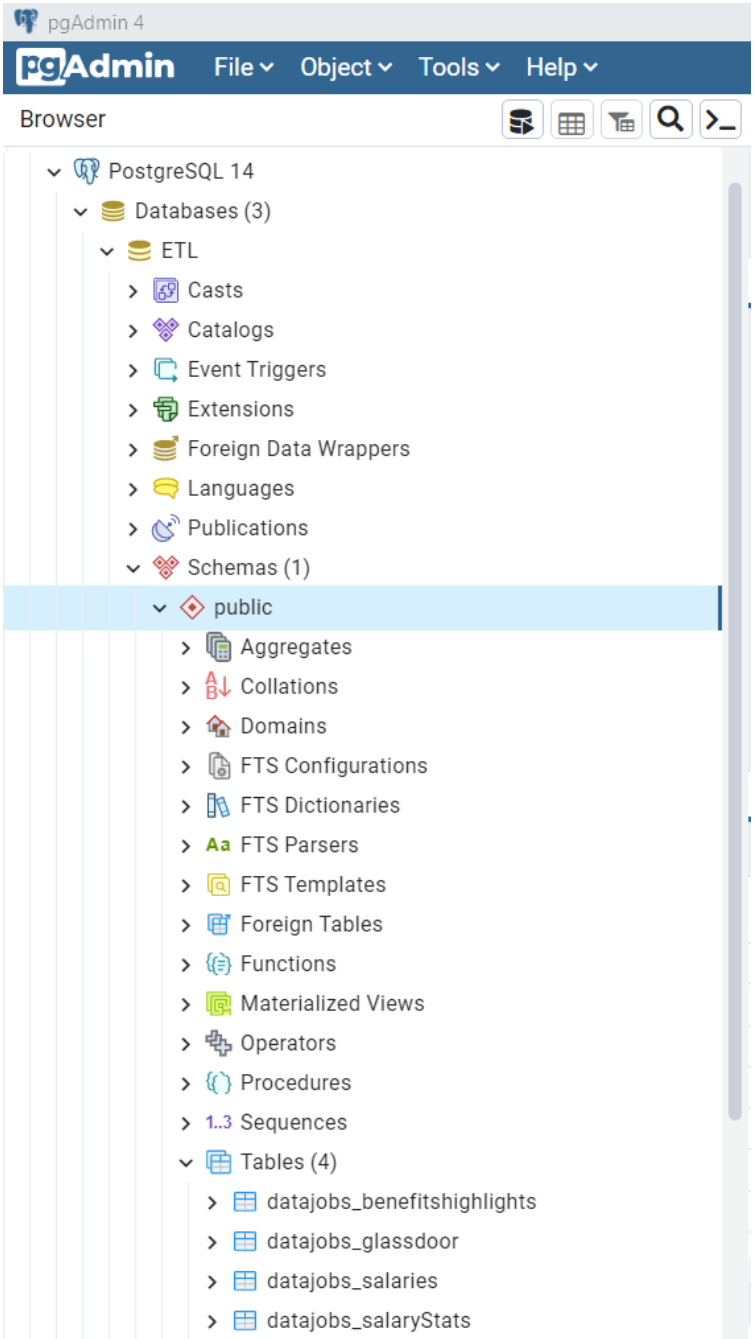


Análisis

Como resultado quedó un data warehouse bastante completo para las ofertas laborales, que aunque la dimensión de datajobs_salaries no haya quedado con

todos los títulos de trabajo disponibles cuenta con la mayor cantidad de trabajos más recurrentes en donde se puede hacer un buen análisis de cómo varias en cuanto a cómo se da el salario dependiendo del trabajo.

Base de datos con las transformaciones



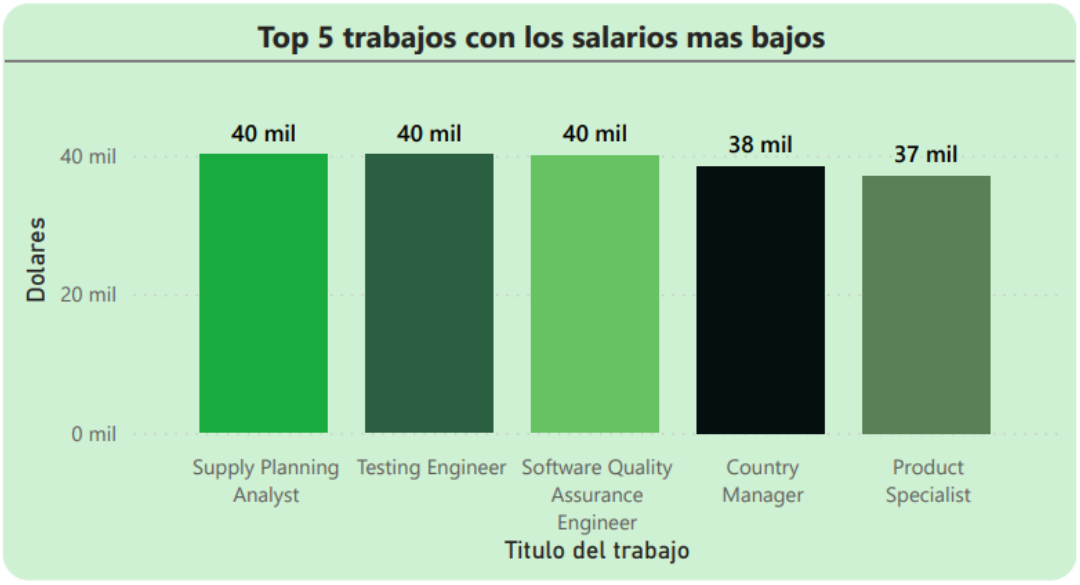
Gráficos

Los gráficos que use fueron una tabla para mostrar sus respectivos promedio de salario de su salario mínimo, la mediana del salario y el salario máximo.

job_title	Promedio de min_salary	Promedio de median_salary	Promedio de max_salar
(Senior Web Application Developer	56.870,00	65.967,00	75.365,(
.Net Developer	93.169,00	110.057,00	136.800,(
.NET Software Engineer	103.967,00	121.974,00	155.400,(
Account Manager	45.550,00	60.250,00	100.400,(
Actuarial Analyst	68.356,00	76.652,00	85.249,(
Agile Business Analyst	43.723,00	51.186,00	58.950,(
Agile Project Manager	43.842,00	52.948,00	62.355,(
Analyst	57.014,00	77.373,00	115.054,(
Application Developer	83.535,00	103.630,00	138.174,(
Application Engineer	80.238,00	105.273,00	145.027,(
Total	**65.334,76**	**76.913,10**	**94.136,8**

Luego hice un top de los 5 trabajos con los salarios más altos, en donde los trabajos que tienen los salarios más altos son, Staff Software Engineer, Software Engineering Manager, Senior Machine Learning Engineer,Computer Vision Engineer y Principal Software Engineer.

También hice una gráfica parecida pero con un top de los salarios más bajos a partir de min_salary estos trabajos que al parecer cuentan con los salarios más bajos del mínimo son; Supply Planning Analyst, Software Quality Assurance Engineer, Country Manager y Product Specialist.



Ahora bien cabe recalcar que estos salarios son tomados a partir de páginas de ofertas laborales y no es que describan exactamente cual es el valor de cada uno de los trabajos, además de que no se sabe cómo la API hace la debida extracción de estos datos.