**Analítica de Grandes Datos**

**Departamento de Ciencias de la Computación y la Decisión**

**Facultad de Minas**

**Universidad Nacional de Colombia**

Trabajo Nociones de Arquitectura de la Información

Versión: aaaa.mm.dd HH:mm

*Observación: Cada vez que agregue nuevos elementos al documento, o que modifique algún componente del informe, revise la coherencia y consistencia con los otros elementos que hacen parte del mismo.*

**Responsables**

|  |
| --- |
| Nombre Completo – Documento de Identificación |
| 1.Maria Claudia Ordoñez Ceballos - 43868082 |
| 2.Yubar Daniel Marin Benjumeda - 1216722806 |
| 3.Alejandro Agudelo Galeano - 1152209359 |
| 4.Gipsy Tatiana Naranjo Dueñas - 1016077759 |
| 5. |
| **REPO EN GITHUB: https://github.com/ydmarinb97/arquitectura** |

*Realiza este trabajo considerando los datos que generan los sistemas transaccionales e información no estructurada de tu dominio (si trabajas por ejemplo para TCC tu dominio es la mensajería; también puedes explorar en la página* [*https://www.kaggle.com/datasets*](https://www.kaggle.com/datasets) *o* [*https://arxiv.org/*](https://arxiv.org/)*). Considera tener acceso a esta información, de al menos 10 MB (puede ser uno o varios archivos de texto), y* ***tener al menos cuatro clases conceptuales. Este documento también debe almacenarse en el REPO. Plazo Máximo de Entrega 23 de Mayo, NO SE recibirá por correo electrónico, envío por*** [***https://forms.gle/h7ty3yZykaUq5m7y6***](https://forms.gle/h7ty3yZykaUq5m7y6)

1. **Comprensión del negocio**
   1. **Descripción del contexto del negocio.**

Describa en máximo 250 palabras el contexto en el cual se generan los datos y cuál es el proceso que los genera.

Según la revista Forbes, hoy en día todas las compañías, sin importar el producto o servicio que ofrecen son empresas de tecnología y las que entiendan esto son las compañías que finalmente moldearán el mundo. Es por esto, que hoy las empresas están demandando un gran volumen de servicios de tecnología ya sea contratando personas de forma directa o contratando compañías que prestan servicios de tecnología.

Por otro lado, la pandemia generada por el Covid – 19 introdujo una “nueva realidad” que trae consigo cambios en las formas de trabajo que permite que dicho talento pueda ser en cualquier parte del mundo independientemente de donde esté ubicada la empresa que lo busca.

Esta situación, lleva a que la industria de servicios de tecnología tenga una gran demanda de talento humano especializado y que se de una “pelea” de las empresas por atraer y retener el mejor talento. Parte de la estrategia para que una empresa logre retención y atracción de talento, es tener la información de lo que pasa en el mercado, es decir, comprender que está ofreciendo la competencia para tener una base comparativa de cuál es su situación respecto a los otros competidores, en términos de los cargos que están ofreciendo, el salario, las condiciones de trabajo y el lugar donde están buscando. La base de datos a trabajar, comprende información respecto a diferentes ofertas laborales del mercado que permite informarse de dicha situación con el fin de entender cómo se está comportando el mercado.

* 1. **Identificación del problema:**

Delimite en máximo 150 palabras la problemática, así como identificar los requisitos, supuestos, restricciones y beneficios de la solución de este.

La alta demanda de talento humano especializado en desarrollo de productos de software tiene hoy a las empresas en una dura “pelea” por el poco que existe. Compañías que tradicionalmente han buscado este talento en Estados Unidos, han empezado a buscar este talento en Latinoamérica debido al trabajo casi 100% remoto que ha traído esta nueva realidad, la cercanía en términos de husos horarios y la capacidad técnica. Este fenómeno, ha generado una burbuja salarial en estos roles, se han generado cambios en las escalas salariales de más del 15% de un mes a otro, imponiendo presión en los empleadores tanto para retener el talento, como para atraer el nuevo. Si las compañías no tienen el talento que es su “materia prima”, no podrían sobrevivir.

Es necesario entonces, tener información en tiempo real de lo que está pasando en el mercado laboral para definir estrategias efectivas a tiempo.

**Determinación de objetivos:**

Describa en máximo 150 palabras las metas a lograr al proponer una solución basada en un modelo de datos o de analítica (cómo y qué tipo de ventaja competitiva se ganará).

Con esta solución, se busca tener información actualizada diariamente de las ofertas laborales requiriendo talento de desarrolladores de software con el fin de entender qué empresas lo están buscando, en dónde lo están buscando, que tipo de roles se están moviendo más y cuáles son las condiciones contractuales y de salario de estas ofertas. Esto con el fin de analizar y entender que cambia por país, por compañía, o por rol. Con esto se espera lograr saber cuáles son las bases salariales por país y explorar los países donde se pueda tener una mejor oferta que los competidores, entender que competidores están buscando talento dónde está mi estrategia de crecimiento para estar preparado, definir cuáles son los roles que tienen más demanda para enfocar mis estrategias de retención, y detectar cambios en las escalas salariales para reaccionar, si es del caso, a tiempo.

* 1. **Evaluación de la situación actual:**

Describa en máximo 150 palabras el estado actual antes de implementar la solución de analítica, a fin de tener un punto de comparación que permita medir el grado de éxito de la solución.

La información que se tiene para el análisis de lo que está pasando en el mercado laboral, se obtiene hoy de forma reactiva, es decir, se recopila información cuando a un empleado le hacen una oferta y entrega la información de las características de la oferta, o en las entrevistas de selección que hace el equipo de talento humano cuando una persona no acepta una oferta y acepta otra. Esta situación genera que las estrategias sean reactivas, que la información no sea de gran volumen y no permite tener una mirada global del mercado ni de los competidores, lo que se ve reflejado en altos índices de rotación y de procesos de retención que son desgastantes y costosos ya que no es posible prevenir la rotación ni estar preparados para ella.

1. **Comprensión de los datos**
   1. **recolección de datos**

Describa en máximo 150 palabras los datos a utilizar identificando las fuentes, las técnicas empleadas en su recolección, los problemas encontrados en su obtención y la forma como se resolvieron los mismos. Además, adjunte los datos (archivos de texto, etc.) agréguelos en el github **(REPO EN GITHUB)** en un solo archivo, por favor comprímalo(s). Llame el archivo T1.2.1.Datos.zip

Los datos fueron tomados de la página kaggle, esta base de datos fue construida por el equipo de datos de la empresa DataStock, contiene información sobre características de empleos publicados en distintas paginas en estados unidos, tiene un peso de 232 MB y un total 29982 registros. Existe una base de datos más completa pero no es de libre acceso.

Los datos fueron descargados por medio de la api de kaggle utilizando python. Uno de los retos más grandes a la hora de manipular los datos es que estaban disponibles en formato .ldjson, un formato no muy común a la hora de almacenar datos. Para la lectura y manejo del archivo se construyó una función en python que lee linea por linea del archivo y luego la transforma en un data frame.

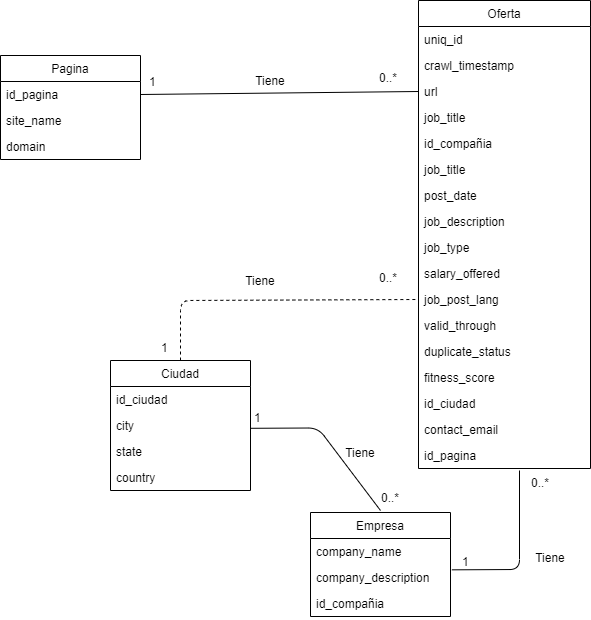
* 1. **Descripción de datos (diccionario):**

Diligencia la siguiente tabla, puede agregar otra columna si lo considera necesario.

|  |  |  |
| --- | --- | --- |
| Nombre del atributo / variable | Formato o Tipo de Dato | Descripción |
| uniq\_id | string | Id único del post de la oferta |
| crawl\_timestamp | string | Fecha y hora de creación de la oferta |
| url | string | Url del post de la oferta |
| job\_title | string | Título de la oferta |
| company\_name | string | Compañía que ofrece la oferta |
| city | string | Ciudad donde se requiere al oferta |
| state | string | Estado donde se requiere la oferta |
| country | string | País donde se requiere la oferta |
| post\_date | string | Fecha cuando se publicó la oferta |
| job\_description | string | Descripción del trabajo |
| job\_type | string | Tipo de oferta(tiempo completo,...) |
| inferred\_salary\_time\_unit | string | Salario estimado en unidades |
| company\_description | string | Descripción de la empres que postea el trabajo |
| salary\_offered | string | Salario estimado |
| job\_board | string | Tipo de trabajo |
| geo | string | Ubicación del trabajo |
| job\_post\_lang | string | Lenguaje del trabajo |
| valid\_through | string | Fecha limite des postulación |
| html\_job\_description | string | Descripción del trabajo en html |
| inferred\_iso2\_lang\_code | string | Lenguaje de la posición inferido |
| inferred\_iso3\_lang\_code | string | Código Lenguaje de la posición inferido |
| site\_name | string | Nombre del sitio donde se publicó la oferta |
| domain | string | Dominio del sitio donde se publicó el trabajo |
| postdate\_yyyymmdd | string | Fecha en la que se postuló el trabajo |
| has\_expired | string | Si puede expirar o no el post |
| last\_expiry\_check\_date | string | Última fecha en la que se chequeo la expiración del trabajo |
| latest\_expiry\_check\_date | string | Última fecha en la que se chequeo la expiración del trabajo |
| duplicate\_status | string | Se encarga publicado en otro sitio |
| postdate\_in\_indexname\_format | string | Fecha en la que se publicó la oferta |
| inferred\_city | string | Inferencia de la ciudad donde se desempeñará el trabajo |
| inferred\_state | string | Inferencia del estado donde se desempeñará el trabajo |
| inferred\_country | string | Inferencia del país donde se desempeñará el trabajo |
| fitness\_score | int | Puntaje de aptitud del candidato otorgado por el sitio web |
| contact\_email | string | Correo de contacto |

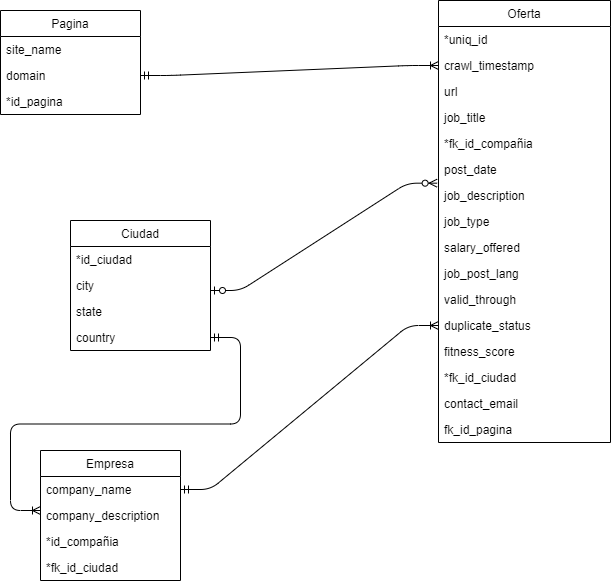
* 1. **Modelo del domino**

***Observación****: Incluya el gráfico del modelo del dominio que representa la estructura de datos de su problema.*

**

1. **Modelo Entidad-Relación**
   1. **Toma de pantalla del modelo E-R**

***Observación****: lo que se pide, puede usar https://draw.io o Microsoft Visio® y modele usando la notación de Barker.*

**

* 1. **Sentencia o consulta de creación del tabla(s)**

***Observación****: Escriba el código en el Sistema de Gestión de Bases de Datos Relacionales de su elección (se recomienda SQLite por simplicidad, mediante https://sqlitebrowser.org/) para crear las tablas que corresponda con su conjunto de datos específico. Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.3.2.Creacion\_Tablas.sql*

***Se recomienda repasar SQL en*** [***https://www.w3schools.com/sql/default.asp***](https://www.w3schools.com/sql/default.asp)

*CREATE SCHEMA IF NOT EXISTS `trabajos` DEFAULT CHARACTER SET utf8 ;*

*-- -----------------------------------------------------*

*-- Tabla `trabajos`.`ciudad`*

*-- -----------------------------------------------------*

*CREATE TABLE IF NOT EXISTS `trabajos`.`ciudad` (*

*`idciudad` INT NOT NULL AUTO\_INCREMENT,*

*`city` VARCHAR(45) NOT NULL,*

*`state` VARCHAR(45) NULL,*

*`country` VARCHAR(45) NULL,*

*PRIMARY KEY (`idciudad`))*

*ENGINE = InnoDB;*

*-- -----------------------------------------------------*

*-- Tabla `trabajos`.`empresa`*

*-- -----------------------------------------------------*

*CREATE TABLE IF NOT EXISTS `trabajos`.`empresa` (*

*`idempresa` INT NOT NULL AUTO\_INCREMENT,*

*`company\_name` VARCHAR(255) NOT NULL,*

*`company\_description` VARCHAR(255) NULL,*

*`idciudad` INT NOT NULL,*

*PRIMARY KEY (`idempresa`),*

*INDEX `fk\_ciudad\_empresa\_idx` (`idciudad` ASC) VISIBLE,*

*CONSTRAINT `fk\_ciudad\_empresa`*

*FOREIGN KEY (`idciudad`)*

*REFERENCES `trabajos`.`ciudad` (`idciudad`)*

*ON DELETE NO ACTION*

*ON UPDATE NO ACTION)*

*ENGINE = InnoDB;*

*-- -----------------------------------------------------*

*-- Tabla `trabajos`.`pagina`*

*-- -----------------------------------------------------*

*CREATE TABLE IF NOT EXISTS `trabajos`.`Pagina` (*

*`idpagina` INT NOT NULL AUTO\_INCREMENT,*

*`site\_name` VARCHAR(255) NOT NULL,*

*`domain` VARCHAR(255) NOT NULL,*

*PRIMARY KEY (`idpagina`))*

*ENGINE = InnoDB;*

*-- -----------------------------------------------------*

*-- Tabla `trabajos`.`oferta`*

*-- -----------------------------------------------------*

*CREATE TABLE IF NOT EXISTS `trabajos`.`oferta` (*

*`uniq\_id` VARCHAR(80) NOT NULL,*

*`crawl\_timestamp` VARCHAR(45) NOT NULL,*

*`url` VARCHAR(255) NOT NULL,*

*`idempresa` INT NOT NULL,*

*`post\_date` VARCHAR(45) NOT NULL,*

*`job\_description` TEXT NOT NULL,*

*`salary\_offered` VARCHAR(255) NULL,*

*`job\_post\_lang` VARCHAR(45) NULL,*

*`valid\_through` VARCHAR(45) NULL,*

*`job\_type` VARCHAR(255) NULL,*

*`duplicate\_status` VARCHAR(45) NULL,*

*`fitness\_score` INT NULL,*

*`idciudad` INT NOT NULL,*

*`contact\_email` VARCHAR(255) NULL,*

*`idpagina` INT NOT NULL,*

*PRIMARY KEY (`uniq\_id`),*

*INDEX `fk\_oferta\_ciudad\_idx` (`idciudad` ASC) VISIBLE,*

*INDEX `fk\_oferta\_empresa\_idx` (`idempresa` ASC) VISIBLE,*

*INDEX `fk\_oferta\_pagina\_idx` (`idpagina` ASC) VISIBLE,*

*CONSTRAINT `fk\_oferta\_ciudad`*

*FOREIGN KEY (`idciudad`)*

*REFERENCES `trabajos`.`ciudad` (`idciudad`)*

*ON DELETE NO ACTION*

*ON UPDATE NO ACTION,*

*CONSTRAINT `fk\_oferta\_empresa`*

*FOREIGN KEY (`idempresa`)*

*REFERENCES `trabajos`.`empresa` (`idempresa`)*

*ON DELETE NO ACTION*

*ON UPDATE NO ACTION,*

*CONSTRAINT `fk\_oferta\_pagina`*

*FOREIGN KEY (`idpagina`)*

*REFERENCES `trabajos`.`Pagina` (`idpagina`)*

*ON DELETE NO ACTION*

*ON UPDATE NO ACTION)*

*ENGINE = InnoDB;*

* 1. **Sentencias para Insertar datos**

***Observación****: Escriba el código para insertar los datos en cada una de las tablas creadas. Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.3.3.Insertar\_Datos.sql*

*-- -----------------------------------------------------*

*-- Tabla `trabajos`.`ciudad`*

*-- -----------------------------------------------------*

*mysql –u root –p*

*USE trabajos;*

*LOAD DATA LOCAL INFILE 'C:/Users/ydmar/Documents/Mestria/Analitica de grandes datos/careerbuilder-job-listing-2020/ciudad.csv'*

*INTO TABLE ciudad*

*FIELDS TERMINATED BY ','*

*OPTIONALLY ENCLOSED BY '"' ESCAPED BY '"'*

*LINES TERMINATED BY '/n'*

*IGNORE 1 ROWS;*

*-- -----------------------------------------------------*

*-- Tabla `trabajos`.`empresa`*

*-- -----------------------------------------------------*

*mysql –u username –p*

*USE trabajos;*

*LOAD DATA INFILE 'C:\Users\ydmar\Documents\Mestria\Analitica de grandes datos\careerbuilder-job-listing-2020\empresa.csv'*

*INTO TABLE trabajos*

*FIELDS TERMINATED BY ','*

*ENCLOSED BY '"'*

*LINES TERMINATED BY '/n'*

*IGNORE 1 ROWS;*

*-- -----------------------------------------------------*

*-- Tabla `trabajos`.`oferta`*

*-- -----------------------------------------------------*

*mysql –u username –p*

*USE trabajos;*

*LOAD DATA INFILE 'C:\Users\ydmar\Documents\Mestria\Analitica de grandes datos\careerbuilder-job-listing-2020\oferta.csv'*

*INTO TABLE trabajos*

*FIELDS TERMINATED BY ','*

*ENCLOSED BY '"'*

*LINES TERMINATED BY '/n'*

*IGNORE 1 ROWS;*

*-- -----------------------------------------------------*

*-- Tabla `trabajos`.`pagina`*

*-- -----------------------------------------------------*

*mysql –u username –p*

*USE trabajos;*

*LOAD DATA INFILE 'C:\Users\ydmar\Documents\Mestria\Analitica de grandes datos\careerbuilder-job-listing-2020\pagina.csv'*

*INTO TABLE trabajos*

*FIELDS TERMINATED BY ','*

*ENCLOSED BY '"'*

*LINES TERMINATED BY '/n'*

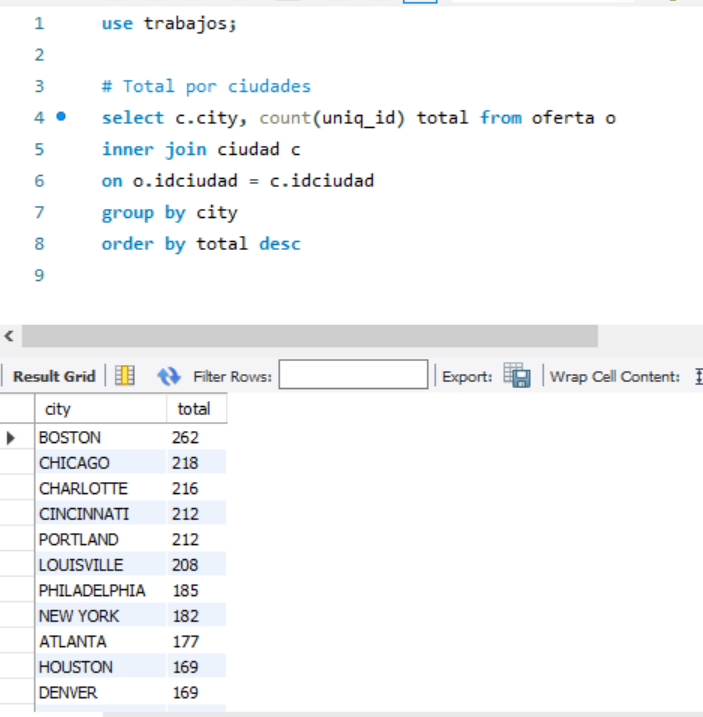
*IGNORE 1 ROWS;*

* 1. **Sentencia de consulta**

***Observación****: realice la exploración básica de los datos, conteos totales y por categorías, máximos, promedio y mínimos.* Es decir, aplique estadística descriptiva con el fin de conocer las propiedades de los datos y entenderlos lo mejor posible. Use solamente sentencias SQL. Anexe las tomas de pantalla donde evidencie la sentencia SQL y su correspondiente ejecución. Además, *Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.3.4.Consultar\_Datos.sql*

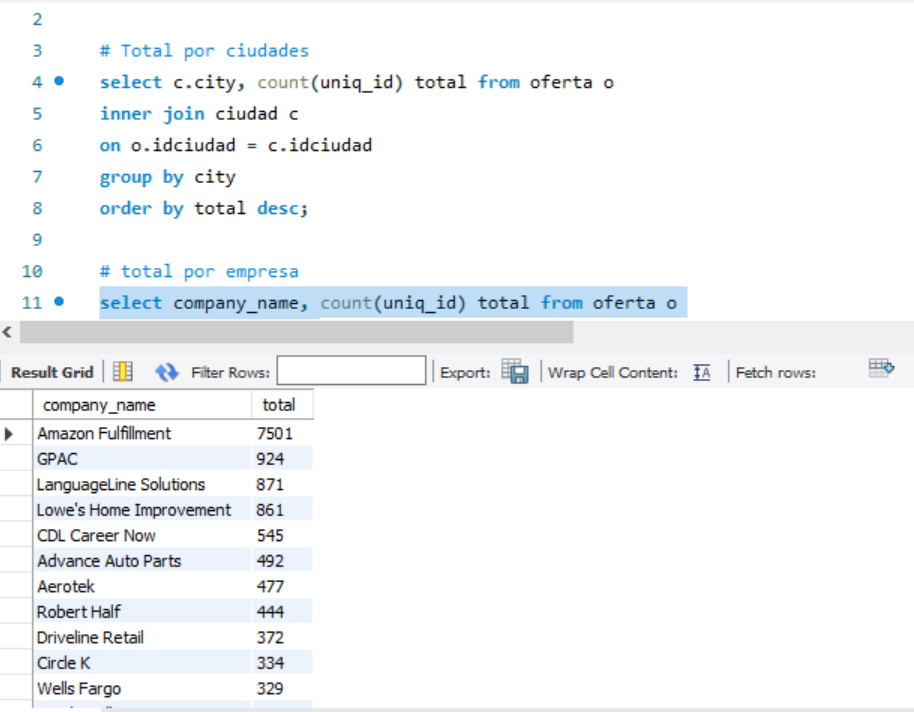
**Total por ciudad**

*El siguiente query resume el total de ofertas laborales para cada ciudad en la que se solicita el trabajo. La ciudad para la que más se ofrecen ofertas es Boston seguido de Chicago*

**

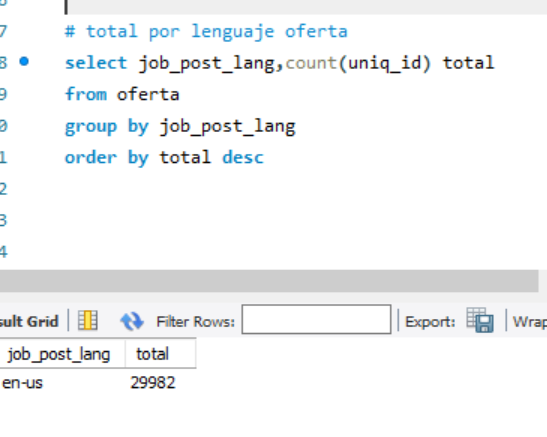
**Total por empresa**

*El siguiente query resume el total de ofertas laborales para cada empresa en la que se solicita el trabajo. La empresa que más ofrece ofertas es amazon, seguido de GPAC*

****

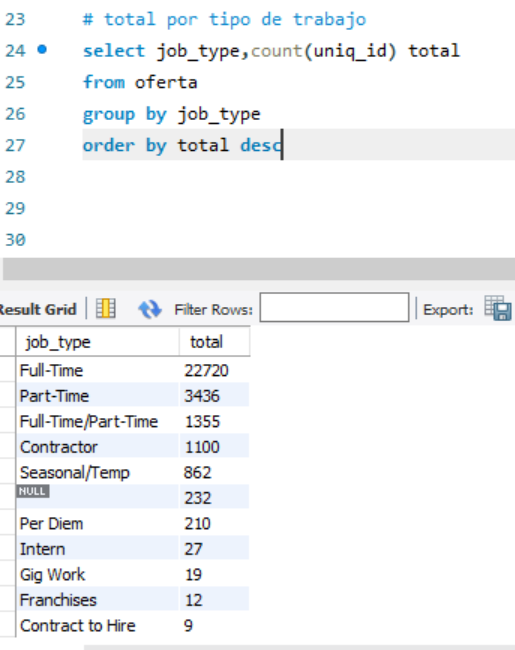
**Total por lenguaje**

*El siguiente query resume el total de ofertas laborales para cada lenguaje requerido en la aplicación. Podemos ver que el total de las ofertas solicitan que se desempeñe en inglés.*

****

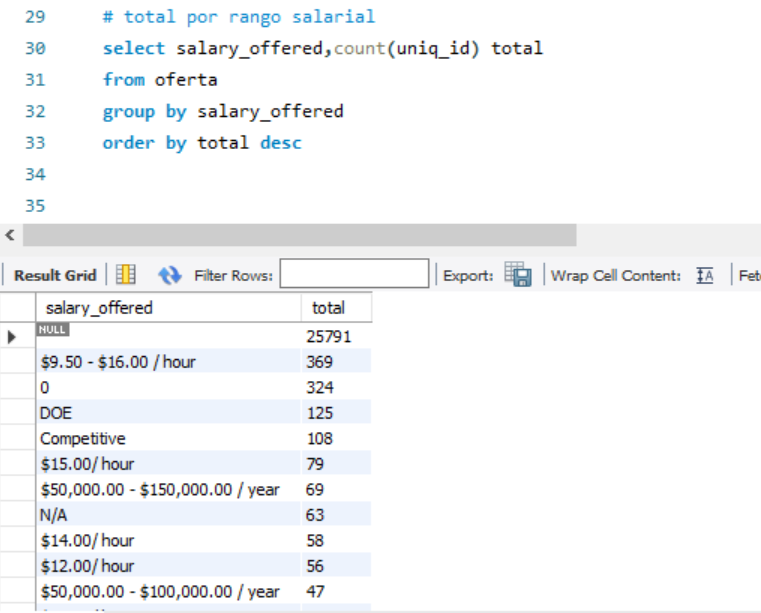
**Total por tipo de trabajo**

*El siguiente query resume el total de ofertas laborales para cada disponibilidad temporal requerida en la aplicación. Se puede observar como la mayor cantidad de ofertas requieren disponibilidad full-time, seguido de parcial-time*

****

**Total por rango salarial**

*El siguiente query resume el total de ofertas laborales para cada disponibilidad temporal requerida en la aplicación. Se puede observar como la mayor cantidad de ofertas requieren disponibilidad full-time, seguido de parcial-time*

****

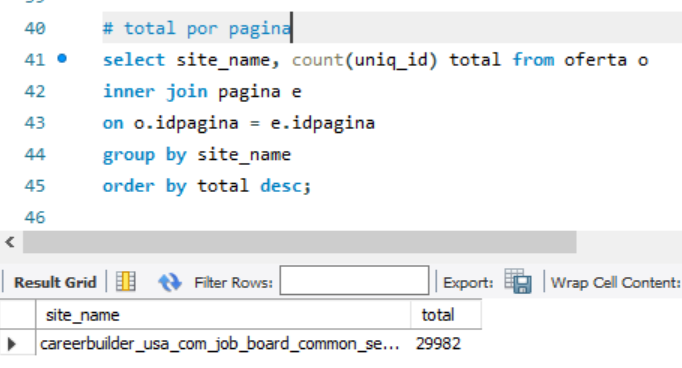
**Total mes año**

*El siguiente query resume el total de ofertas laborales para cada periodo mensual del año 2020. Se puede observar que el mes donde más ofertas se postearon fue el mes 5, además nos permite inferir que la información recolectada fue específicamente de cuatro meses del año 2020.*

****

**Total por página**

*El siguiente query resume el total de ofertas laborales extraídas por cada página, se puede observar que como esta base de datos hace parte de un estudio más grande solo contiene datos de una única página.*

****

1. **MongoDB**
   1. **Sentencia o consulta de creación del documento(s)**

***Observación****: Escriba el código en MongoDB para crear al menos 20 documentos que correspondan a su conjunto de datos específico. Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.4.1.Creacion\_Documentos.sql*

* *Si se quiere importar de manera masiva*

*cd C:\Program Files\MongoDB\Server\4.4\bin*

*mongoimport --type csv -d jobs -c jobs--headerline --drop C:/Users/ydmar/Documents/Maestria/Analitica de grandes datos/careerbuilder-job-listing-2020/trabajos.csv*

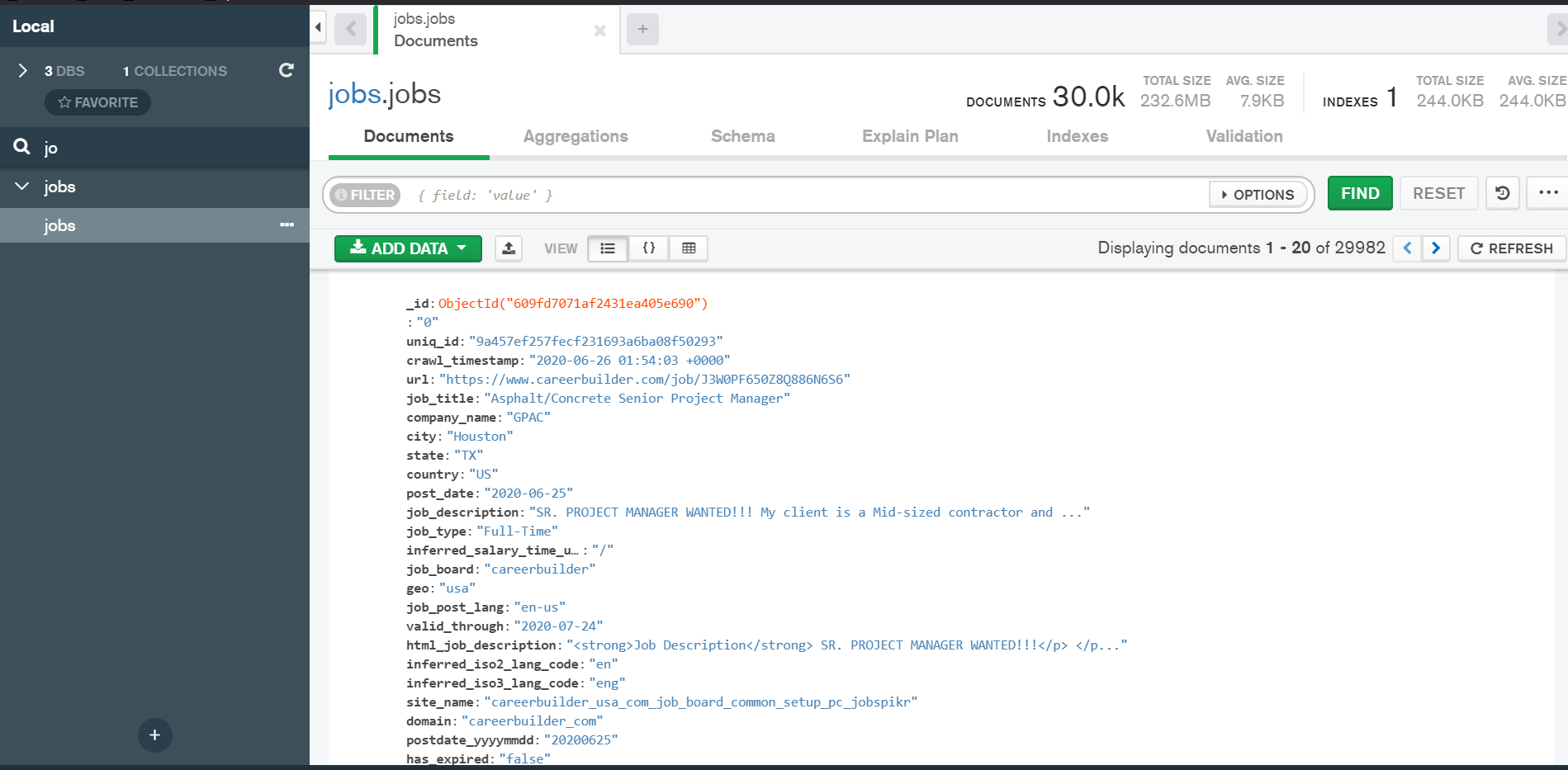
* *Ejemplo de como ingresar manualmente los datos*

*C:\Program Files\MongoDB\Server\4.4\bin*

*mongo.exe*

*use jobs*

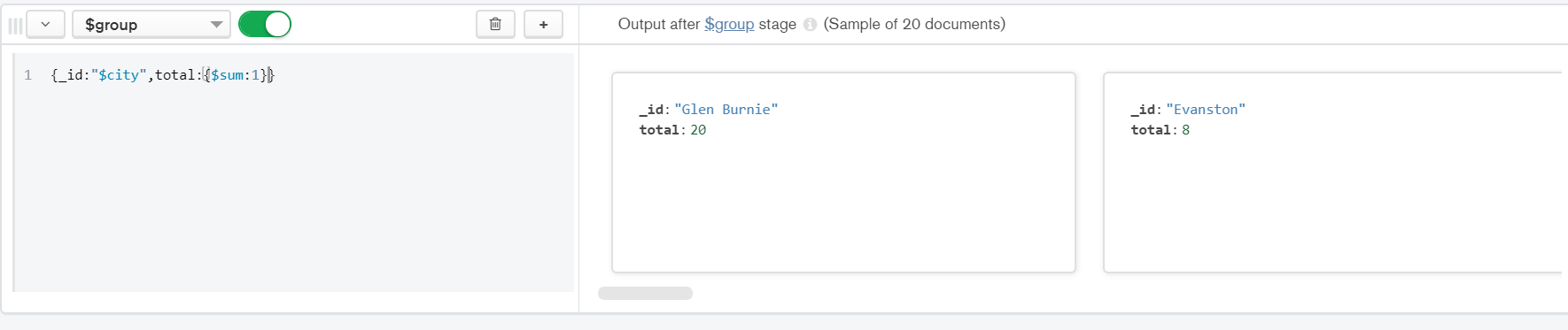
*jobs.jobs.insert({"\_id":{"$oid":"609fd7071af2431ea405e690"},"uniq\_id":"9a457ef257fecf231693a6ba08f50293","crawl\_timestamp":"2020-06-26 01:54:03 +0000","url":"https://www.careerbuilder.com/job/J3W0PF650Z8Q886N6S6","job\_title":"Asphalt/Concrete Senior Project Manager","company\_name":"GPAC","city":"Houston","state":"TX","country":"US","post\_date":"2020-06-25","job\_description":"SR. PROJECT MANAGER WANTED!!! My client is a Mid-sized contractor and only does work in the HOUSTON metro area. We are looking for someone with ASPHALT AND CONCRETE/ TX DOT EXPERIENCE. Qualifications: Previous asphalt paving experience preferred Project Management experience preferred Bachelor's Degree or equivalent experience; preferably in construction management Excellent math and geometry skills Knowledgeable in and proven abilities with estimating in ASPHALT AND CONCRETE or construction Excellent communication skills, both written and verbal Must have the ability to interact with external and internal customers and partners at a fast pace, multitasking environment, working independently with a high degree of problem-solving To be considered please apply with your RESUME at [ Email address blocked ] - Click here to apply to Asphalt/Concrete Senior Project Manager All qualified applicants will receive consideration without regard to race, age, color, sex (including pregnancy), religion, national origin, disability, sexual orientation, gender identity, marital status, military status, genetic information, or any other status protected by applicable laws or regulations. GPAC (Growing People and Companies) is an award-winning search firm specializing in placing quality professionals within multiple industries across the United States for the past 30 years. We are extremely competitive, client-focused and realize that our value is in our ability to deliver the right solutions at the right time. Asphalt and Concrete Project Manager Recommended skills Construction Management Self Motivation Social Skills Project Management Multitasking Construction Recommended Jobs Project Manager Houston,Tx Project Manager, Houston,Tx Asphalt/Concrete Senior Project Manager GPAC Houston Full-Time Apply to this job. Think you're the perfect candidate? Apply Now $('.external-apply-email-saved').on('click', function (event) { window.ExternalApply = window.open('/interstitial', 'ExternalApply-j3w0pf650z8q886n6s6'); }); Help us improve CareerBuilder by providing feedback about this job: Report this job Report this Job Once a job has been reported, we will investigate it further. If you require a response, submit your question or concern to our Trust and Site Security Team Select Report Type Cannot apply/broken link Fee or investment required Multiple positions in one posting Irrelevant/Excessive keywords Offensive content Other Possible Scam Provide additional details here. Cancel Job ID: GPRecProCoor CAREERBUILDER TIP For your privacy and protection, when applying to a job online, never give your social security number to a prospective employer, provide credit card or bank account information, or perform any sort of monetary transaction. Learn more. By applying to a job using CareerBuilder you are agreeing to comply with and be subject to the CareerBuilder Terms and Conditions for use of our website. To use our website, you must agree with the Terms and Conditions and both meet and comply with their provisions. |when applying to a job online, never give your social security number to a prospective employer, provide credit card or bank account information, or perform any sort of monetary transaction.","job\_type":"Full-Time","inferred\_salary\_time\_unit":"/","job\_board":"careerbuilder","geo":"usa","job\_post\_lang":"en-us","valid\_through":"2020-07-24","html\_job\_description":"<strong>Job Description</strong> SR. PROJECT MANAGER WANTED!!!</p> </p>My client is a Mid-sized contractor and only does work in the HOUSTON metro area. We are looking for someone with ASPHALT AND CONCRETE/ TX DOT EXPERIENCE.</p> </p> </p><p>Qualifications:</p><ul><li>Previous asphalt paving experience preferred</li><li>Project Management experience preferred</li><li>Bachelor's Degree or equivalent experience; preferably in construction management</li><li>Excellent math and geometry skills</li><li>Knowledgeable in and proven abilities with estimating in ASPHALT AND CONCRETE or construction </li><li>Excellent communication skills, both written and verbal</li><li>Must have the ability to interact with external and internal customers and partners at a fast pace, multitasking environment, working independently with a high degree of problem-solving</li></ul> </p>To be considered please apply with your RESUME at [ Link Removed ] </p><p>All qualified applicants will receive consideration without regard to race, age, color, sex (including pregnancy), religion, national origin, disability, sexual orientation, gender identity, marital status, military status, genetic information, or any other status protected by applicable laws or regulations.</p><p></p><p>GPAC (Growing People and Companies) is an award-winning search firm specializing in placing quality professionals within multiple industries across the United States for the past 30 years. We are extremely competitive, client-focused and realize that our value is in our ability to deliver the right solutions at the right time.</p> <strong>Job Requirements</strong> Asphalt and Concrete Project Manager","inferred\_iso2\_lang\_code":"en","inferred\_iso3\_lang\_code":"eng","site\_name":"careerbuilder\_usa\_com\_job\_board\_common\_setup\_pc\_jobspikr","domain":"careerbuilder\_com","postdate\_yyyymmdd":"20200625","has\_expired":"false","last\_expiry\_check\_date":"2020.06.26","latest\_expiry\_check\_date":"2020-06-26","duplicate\_status":"NA","postdate\_in\_indexname\_format":"2020.06.22","inferred\_city":"Houston","inferred\_state":"Texas","inferred\_country":"United states","fitness\_score":"10"})*

**

* 1. **Sentencia de consulta**

***Observación****: Realice la exploración básica de los datos, conteos totales y por categorías, máximos, promedio y mínimos.* Es decir, aplique estadística descriptiva con el fin de conocer las propiedades de los datos y entenderlos lo mejor posible. Use solamente sentencias SQL. Anexe las tomas de pantalla donde evidencie la sentencia SQL y su correspondiente ejecución. Además, *Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.4.2.Consultar\_Datos.sql*

* *Total de ofertas para cada ciudad.*

**

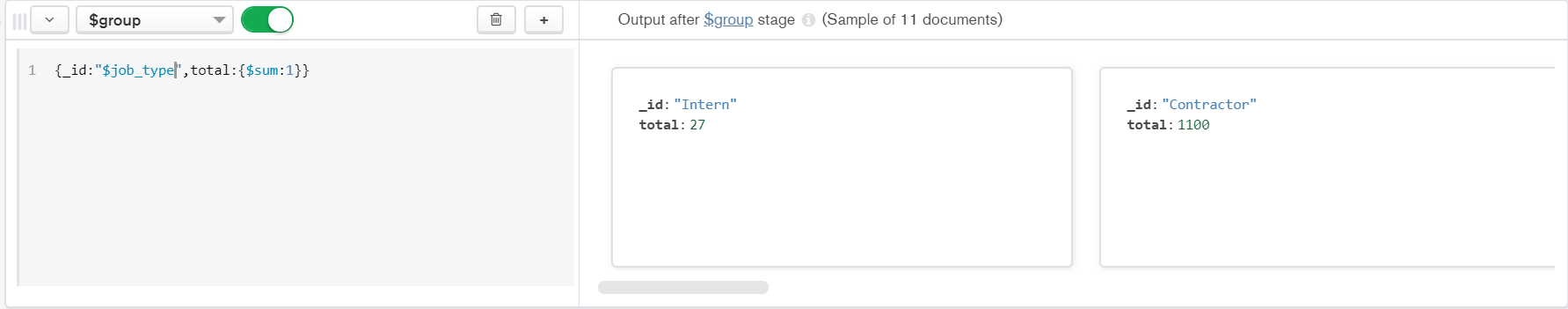
* *Total de ofertas por compañía*



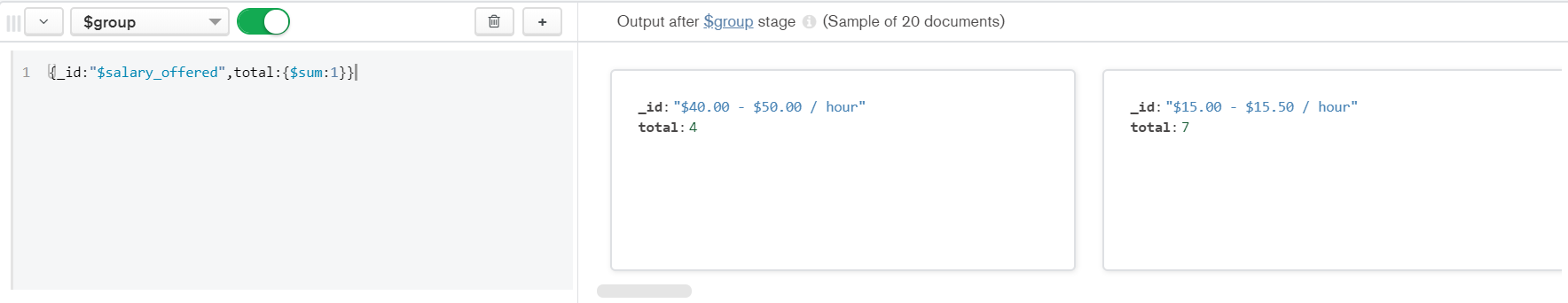
* *Total de ofertas por idioma*



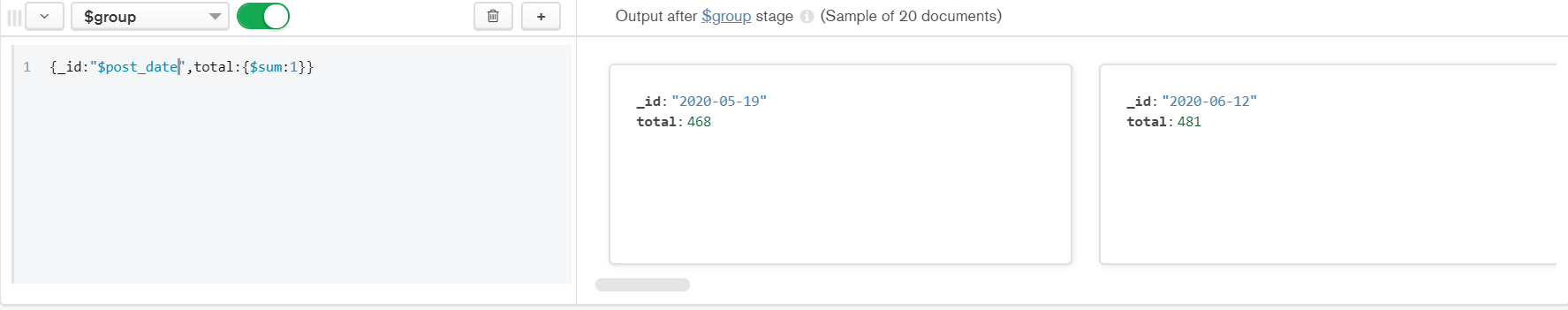
* *Total de ofertas por tipo de trabajo*



* *Total de ofertas por salario ofrecido*



* *Total ofertas por fecha de postulación*



1. **Análisis de lectura**

***Observación****: Considerando el artículo: “*The Definitive Guide to Graph Databases for the RDBMS Developer*” de Neo4J. Compartido en las carpeta de lecturas recomendadas. Analice y responda cada pregunta en máximo 150 palabras:*

1. ¿Cuáles son las limitaciones, que se pueden inferir de la lectura, para migrar los conjuntos de datos relacionales a NoSQL?

Uno de los puntos más importantes que se debe considerar antes de hacer la migración de un conjunto de datos relacionales a NoSQL, es el correcto entendimiento de la información y la estructura que se va a implementar en el NoSQL, es decir, el conocimiento acerca de los nodos, relaciones, propiedades, tipos de relaciones, etc. Una de las limitaciones que se generarían si no se realiza este paso es la pérdida de información y posiblemente retrasos en la migración.

Otra limitación que se puede presentar son los diferentes retos que se encuentran cuando se deben hacer copias de las bases de datos relacionales en formato csv, ya que muchas de estas no fueron diseñadas y optimizadas para exportar tamaños de información muy grandes en un periodo corto de tiempo. En conclusión, se debería realizar todo el análisis desde el planteamiento hasta la migración para evitar posibles limitaciones que generan diferentes problemas como la pérdida de la información, retrasos en el proceso y fallas en la arquitectura del proceso.

1. ¿Cuáles limitaciones adicionales se deben considerar, a parte de las mencionadas en el artículo?

Otra punto importante a tener en cuenta en la migración de la base de datos, es la duplicidad de la información, en una base de datos relacional se espera que exista una fila por cada una de las características de un clave , sin embargo, en el mundo real existe la posibilidad de modificaciones y por lo tanto sería necesario agregar una fila más para la misma clave para generar esta actualización. Se espera que con una base de datos NoSql sea más fácil está actualizaciones.

1. ¿Cuáles son las razones (criterios) que se deben considerar para migrar un conjunto de datos relacionados a NoSQL?

Con la evolución constante de la información y el desarrollo de la tecnología, se ha fomentado el crecimiento y la necesidad de almacenar más información, esta es útil para el análisis y el soporte en la toma de decisiones. La base de datos relacionales fue inicialmente diseñadas para tabular estructuras en forma de filas y columnas, sin embargo, esta estructura no está totalmente relacionada con la forma en la que existe en el mundo real la información. En el mundo real la información se encuentra en constante crecimiento en términos de velocidad, volumen y variedad. Por lo que inicialmente las bases de datos relaciones, no se encuentran con la capacidad de afrontar todos estos cambios. Por lo tanto para las empresas es indispensable que su arquitectura tenga esta facilidad de adaptación ya que va a permitir que se haga un correcto análisis de información, soporte la toma de las necesidad y sobre todo satisfaga las necesidades de sus clientes. Finalmente, es por esto que se debería migrar la información si su sistema actual no cumple con las necesidades de cada uno. Esta decisión dependerá de la empresa.