Tecnológico de Monterrey
Escuela de Ingeniería y Ciencias

Campus LATAM

Nombre del trabajo:

6.2 Fase 3c. Almacenamiento de información en bases de datos NoSql

Curso:

Análisis de Grandes Volúmenes de Datos

Alumnos:

Juan Carlos Alvarado Carricarte, A01793486

Bryan Rodolfo Alvarado Cruz, A01793670

Eduardo Gabriel Arévalo Aguilar, A01793897

Profesor:

Alberto De Obeso Orendain

Tutor:

Luis Angel Lozano Medina

Fecha de entrega:

28 de febrero

Resumen

En este trabajo se describe el proceso de almacenamiento de los datos obtenidos a partir de información demográfica de la región de Jalisco, México, desde Databricks a una base de datos NoSql en MongoDB. Se hace uso de las herramientas adecuadas para interactuar con MongoDB Atlas y el proceso de migración. Además, se presentan cinco consultas que muestran que los datos fueron cargados exitosamente y muestran algunas nuevas características demográficas de la región de Jalisco que se calcularon en la actividad pasada. El uso de MongoDB permite un almacenamiento escalable y flexible de grandes cantidades de datos, lo que lo convierte en una opción atractiva para el almacenamiento de información demográfica y otros tipos de datos similares.

Palabras clave: NoSql, MongoDB, consultas, Apache Spark, Databricks, datos demográficos, fuentes de datos, JSON, México, Jalisco.

Almacenamiento de información en bases de datos NoSql

Uno de los desafíos dentro del ecosistema de Big Data, es el de almacenar esas grandes cantidades de datos y acceder a ellos de manera eficiente, esto ha llevado al desarrollo de nuevas tecnologías que nos permiten hacer estas tareas de la forma más óptima posible. Una de estas tecnologías es MongoDB, una base de datos NoSQL basada en documentos que permite el almacenamiento y la recuperación de grandes cantidades de datos de manera escalable y flexible. En este trabajo se presenta la migración de datos de la información generada a partir de los datos demográficos de la región de Jalisco, México, descargando la información de Databricks para luego migrarla a MongoDB. Se describe la estructura del documento JSON con los campos que tiene cada una las nuevas fuentes a cargar, el proceso de migración y se presentan algunas consultas que muestran esta información almacenada.

Migración de datos a MongoDB

Para llevar a cabo el proceso de migración de datos de Databricks a MongoDB se realizó lo siguiente:

- Se tomaron las tablas nuevas generadas en el proceso de creación de las nuevas características con Apache Spark en la plataforma de Databricks.
- Se exportaron los datos desde Databricks en formato CSV.
- Se utilizó la herramienta mongoimport para realizar la migración de los datos desde el equipo local a MongoDB.
- Para cada tabla se creó una nueva colección en MongoDB.

- Se presentaron cinco consultas donde se muestran que los datos fueron cargados exitosamente. Las consultas también evidencian las características de la región de Jalisco que fueron calculadas.

Tabla anos_de_la_compania

Esta tabla contiene la característica de los *años que tiene una compañía creada* en la región de Jalisco, se hizo el cálculo a partir de la tabla de DENUE. Su estructura en MongoDB puede evidenciar en la siguiente consulta:

```
Atlas atlas-mtntgc-shard-0 [primary] test> db.anos_de_la_compania.find().limit(3)
[
  {
    _id: ObjectId("640eb64a788ddd1e4b79147c"),
    id_denue: 1789827,
    anos_de_la_compania: 9
  },
  {
    _id: ObjectId("640eb64a788ddd1e4b79147d"),
    id_denue: 9337856,
    anos_de_la_compania: 2
  },
  {
    _id: ObjectId("640eb64a788ddd1e4b79147e"),
    id_denue: 9203369,
    anos_de_la_compania: 4
  }
```

Tabla tamano_empresa

Esta tabla clasifica el tamaño de la empresa acorde a su número de empleados. Dependiendo del rango en número de empleados, clasificamos como pequeña, mediana o grande empresa.

```
Atlas atlas-mtntgc-shard-0 [primary] test> db.tamano_empresa.find().limit(3)
[
  {
    _id: ObjectId("640e87a6bc2e9f3cd18b361a"),
    id_denue: 1789827,
    per_ocu: '31 a 50 personas',
    tamano_empresa: 'Pequeña'
  },
  {
    _id: ObjectId("640e87a6bc2e9f3cd18b361b"),
    id_denue: 9316641,
    per_ocu: '6 a 10 personas',
    tamano_empresa: 'Pequeña'
  },
  {
    _id: ObjectId("640e87a6bc2e9f3cd18b361c"),
    id_denue: 9337856,
    per_ocu: '0 a 5 personas',
    tamano_empresa: 'Pequeña'
  }
]
```

Tabla promedio_alfabetas_por_cantidad_empresas

Para esta tabla se creó la medida de promedio de analfabetas extraídos de la tabla de educación y la cantidad de empresas que existen en cada municipio, con el fin de posteriormente evaluar si hay alguna correlación entre estas dos variables. Consulta:

```
Atlas atlas-mtntgc-shard-0 [primary] test> db.promedio_alfabetas_por_cantidad_empresas.find().limit(3)
[
  {
    _id: ObjectId("640e8792e76773ca454a4e5d"),
    cve_mun: 39,
    municipio: 'Guadalajara',
    cantidad_empresas: 99738,
    promedio_alfabetas_mayores_15: 98.271947982632
  },
  {
    _id: ObjectId("640e8792e76773ca454a4e5e"),
    cve_mun: 98,
    municipio: 'San Pedro Tlaquepaque',
    cantidad_empresas: 24625,
    promedio_alfabetas_mayores_15: 97.5796467155587
  },
  {
    _id: ObjectId("640e8792e76773ca454a4e5f"),
    cve_mun: 101,
    municipio: 'Tonalá',
    cantidad_empresas: 19420,
    promedio_alfabetas_mayores_15: 97.6907482179232
  }
]
```

Tabla promedio_poblacion_por_hogar

Esta tabla fue generada luego de combinar los dos únicos indicadores de la tabla hogares, los cuales nos indicaba la cantidad de de todos los hogares y la población total de los mismos, lo que nos permitió calcular el promedio de habitantes por cada hogar en cada municipio.

```
Atlas atlas-mtntgc-shard-0 [primary] test> db.promedio_poblacion_por_hogar.find().limit(3)
[
  {
    _id: ObjectId("640e875ed499a729bebc72ad"),
    cve_municipio: 54,
    promedio_poblacion_por_hogar: 2.9905027932960895,
    unidad_promedio_poblacion: 'Promedio de población por hogar'
  },
  {
    _id: ObjectId("640e875ed499a729bebc72ae"),
    cve_municipio: 34,
    promedio_poblacion_por_hogar: 2.9650986342943852,
    unidad_promedio_poblacion: 'Promedio de población por hogar'
  },
  {
    _id: ObjectId("640e875ed499a729bebc72af"),
    cve_municipio: 104,
    promedio_poblacion_por_hogar: 3.130925507900677,
    unidad_promedio_poblacion: 'Promedio de población por hogar'
  }
]
```

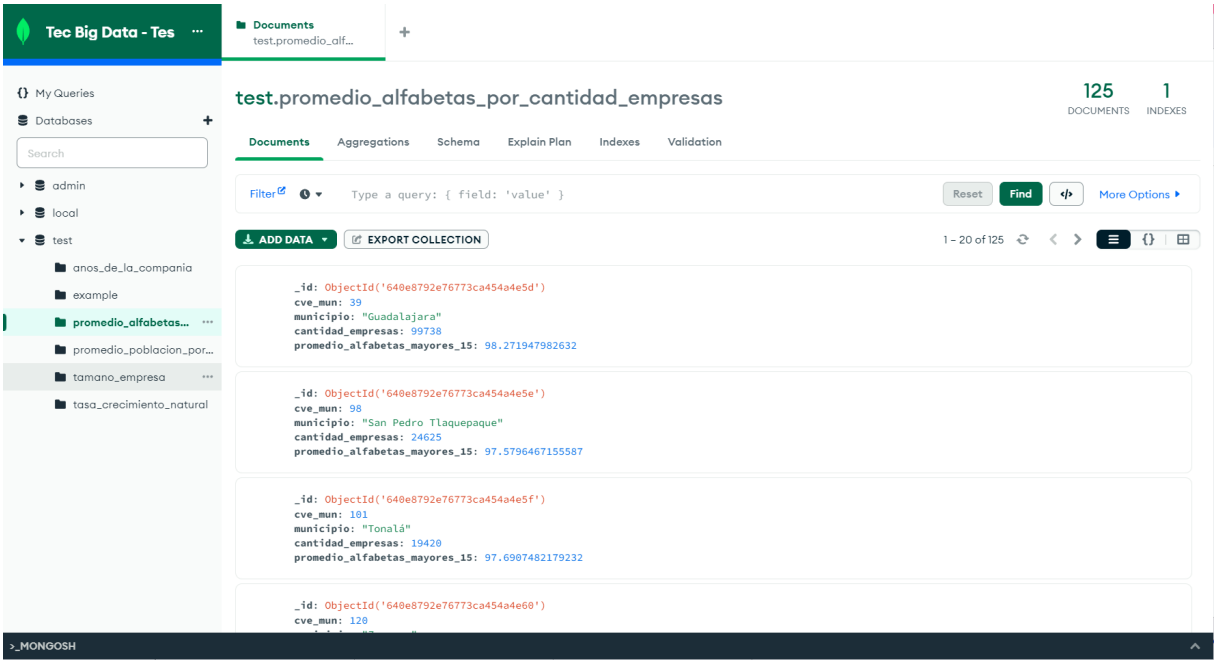
Tablas tasa_crecimiento_natural

Finalmente, con las tablas de natalidad, mortalidad y población, se calculó la *tasa de crecimiento poblacional* sin inmigrantes o tasa de crecimiento natural por cada mil habitantes en cada municipio para el año 2020.

Reemplazando valores, Tasa de crecimiento natural: = [(Nacimientos(2020) - Defunciones(2020)) / Población inicial(2020)] x 1000

```
Atlas atlas-mtntgc-shard-0 [primary] test> db.tasa_crecimiento_natural.find().limit(3)
[
  {
    _id: ObjectId("640e873ec2a13fc888b46993"),
    cve_municipio: 19,
    desc_municipio: 'Bolaños',
    nacimientos: 291,
    muertes: 43,
    poblacion_total: 7043,
    tasa_crecimiento_natural: 35.21226749964504
  },
  {
    _id: ObjectId("640e873ec2a13fc888b46994"),
    cve_municipio: 125,
    desc_municipio: 'San Ignacio Cerro Gordo',
    nacimientos: 522,
    muertes: 139,
    poblacion_total: 18341,
    tasa_crecimiento_natural: 20.8821765443541 77
  },
  {
    _id: ObjectId("640e873ec2a13fc888b46995"),
    cve_municipio: 61,
    desc_municipio: 'Mezquitic',
    nacimientos: 770,
    muertes: 115,
    poblacion_total: 22083,
    tasa_crecimiento_natural: 29.660825069057648
```

También se exploró la herramienta de MongoDB Compass, la cuál nos deja interactuar con nuestros recursos de MongoDB de manera práctica y sencilla.



MongoDB es una herramienta bastante poderosa a la hora de crear bases de datos NoSql, cuenta con una versión comunitaria que se puede descargar local y una en la nube con planes freemium, lo que le da gran flexibilidad a los usuarios de tomar usar recursos acorde a sus necesidades. También cuenta con herramientas complementarias como mongoimport que nos permite cargar archivos a nuestra base de datos, o como MongoDB Compass que nos proporciona de una interfaz gráfica e intuitiva para manejar nuestros datos. Es importante que los archivos que generamos después de los procesos de ETL y características, se agreguen en repositorios de datos o bases de datos que permitan acceso de manera segura a los usuarios tener una alta disponibilidad y seguridad de la información, es por eso que MongoDB se convierte en una opción bastante atractiva para cargar los resultados.

Referencias

Apache Spark™ - Unified Engine for large-scale data analytics. (s. f.).

https://spark.apache.org

Databricks. (2022, 26 abril). *Contact Us*.

https://www.databricks.com/

Geografía, E. D. N. I. Y. (s. f.-a). Directorio Estadístico Nacional de Unidades Económicas. DENUE.

https://www.inegi.org.mx/app/mapa/denue/default.aspx

Regiones de Jalisco | Gobierno del Estado de Jalisco. (s. f.).

https://www.jalisco.gob.mx/es/jalisco/regiones

Nayak, A. (2014). *MongoDB Cookbook.* Packt Publishing.

https://0-search-ebscohost-com.biblioteca-ils.tec.mx/login.aspx?direct=true&db=e000xww&AN=918199&lang=es&site=eds-live&scope=site