Universidad Técnica Federico Santa María
Campus Santiago San Joaquín
Bases de Datos Avanzadas INF325
Primer Semestre 2019

Tarea II

Profesor: José Luis Martí Lara, jmarti@inf.usm.cl
Ayudante: Ignacio Tampe Palma, ignacio.tampe@sansano.usm.cl

1. Motivación

Se busca introducir al tema de las Bases de Datos Columnares y almacenamiento Clave-Multivalor a través del desarrollo de un caso de estudio grupal utilizando el motor de datos HBase con el objetivo de conocer las características propias de este tipo de ambientes NoSQL. Adicionalmente, se deberá realizar un trabajo de análisis y visualización sobre parte del contenido de la base de datos construida.

2. Actividad: Base de datos Columnar

Las Bases de Datos del tipo Columnar permiten el almacenamiento y gestión de datos de modo semi-estructurado que favorece el análisis de los datos en su conjunto por sobre su análisis individual (agregación). Con el propósito de adquirir una noción general de las bases de datos NoSQL del tipo Columnar, cada grupo de trabajo deberá instalar la base de datos HBase y poblarla con un set de datos correspondiente a registros de accidentes automovilísticos en Francia entre los años 2005-2016, con información respecto a los datos personales de los implicados, las características del suceso, del lugar donde ocurrió y del vehículo.

Disponible en: `https://www.kaggle.com/ahmedlahlou/accidents-in-france-from-2005-to-2016`

El dataset contiene distintos archivos con categorías de información, cada archivo se puede considerar como una familia de columnas distinta las cuales están enlazadas por medio del ID del accidente.

Su misión es cargar el dataset en HBase usando las familias de columnas presentadas y los datos multivaluados e.g. Las personas involucradas en el accidente. Deberá detallar el proceso realizado para el modelado e inserción en su resumen ejecutivo.

3. Actividad: Análisis exploratorio

Considerando la cantidad de atributos e información para cada accidente ocurrido, se dividirá el análisis en distintos enfoques que, al combinarlos, permitirán un mejor entendimiento de las características de accidentes. Como equipo deberán elegir que enfoque tomar para sus conclusiones,los cuales están orientados a las distintas familias de columnas, esto no significa que su análisis estará basado exclusivamente en esa familia, debe cruzar información con las otras familias pero sus conclusiones deben ser respecto a la familia elegida (cupos limitados).

Según el enfoque elegido, deberán realizar análisis estadístico de las columnas que lo permitan. Además deberán plantear visualizaciones que ayuden al entendimiento de los datos. los enfoques disponibles se presentan a continuación:

3.1. Enfoque sobre características del accidente

En este caso, se busca analizar las características propias siniestro, esta investigación se orienta a la familia de columnas del archivo *caracteristics.csv* donde se provee información del lugar de los hechos, condiciones climáticas, etc.

Se buscan realizar visualizaciones que permitan responder preguntas como:

- ¿Cómo son las condiciones climáticas donde ocurren la mayoría de los accidentes?
- ¿Cómo varían los tipos de colisiones a lo largo del tiempo?
- ¿Cómo se distribuyen geospacialmente los accidentes según año?

3.2. Enfoque sobre las personas implicadas

Este enfoque busca analizar a las personas involucradas en el accidente, esta investigación se orienta a la familia de columnas del archivo *users.csv* junto con parte de la información de *vehicles.csv*

Se buscan realizar visualizaciones que permitan responder preguntas como:

- ¿Qué relaciones existen entre la edad de los peatones su localización/acción cuando ocurrió el accidente?
- Compare la relación entre género y razones de viaje.
- ¿Cómo influye el equipamiento de seguridad en la gravedad del accidente?

3.3. Enfoque sobre el lugar de los hechos

En esta ocasión se busca analizar las características detalladas del camino y la superficie, esta investigación se orienta a la familia de columnas del archivo *places.csv*
Se buscan realizar visualizaciones que permitan responder preguntas como:

- ¿Qué tipo de carretera deja más heridos al año en promedio?
- ¿Cómo influye la condición de la superficie y el perfil longitudinal en la frecuencia de accidentes en las vías?
- Cual es el efecto de las ciclovías la seguridad de la gente?

Para cada visualización deberá realizar su sendo análisis. Además, deberán **plantear 2 nuevas interrogantes** que hayan surgido en el equipo y realizar el análisis/visualización necesaria para resolverla. Tras esto, se deben realizar conclusiones generales sobre el análisis realizado.

Para hacer este análisis deberán conectarse con la base de datos a través de Python, utilizando Jupyter Notebook y con visualizaciones en Plotly.

4. Evaluación y Entrega

Fecha de entrega: Viernes 10 de Mayo 23:55.
Formato: Archivo comprimido en `tar.gz` con el Jupyter notebook y su versión en PDF.

4.1. Evaluación

- Definir y crear base de datos: 10 ptos.
- Resumen ejecutivo: 10 ptos.
- Estadísticas columnares: 5 ptos.
- Análisis exploratorio: Por cada ítem (5 en total)
  - Visualización: 8 ptos.
  - Análisis y conclusión: 7 ptos.

  Se valorará la diversidad y complejidad de sus gráficos, considere que el gráfico debe proveer información de forma explícita, sin necesidad de leer el análisis.

4.2. Consideraciones

- Equipos de 3 personas.
- Deberán elegir su enfoque desde: `https://doodle.com/poll/s48xihqqi9eatkrz` (cupos limitados, máximo una categoría con 4 grupos)
- Puede que el dataset necesite limpieza de datos.
- Copiar toda la base de datos a memoria en Python no es válido, se deben hacer *queries* desde la base de datos para cada ítem del análisis.
- La media no tiene relevancia estadística si no es a luz de su varianza, considere incluir este dato en su análisis columnar o como barras de error. Los Boxplots le pueden ayudar en este proceso.
- Consultas por moodle o correo.