| | | |
|---|---|---|
| Universidad de los Andes<br>Facultad de Ingeniería | Ingeniería de Sistemas y Computación<br>**Maestría**<br>MINE4102 – Análisis de información sobre Big Data<br>Semestre: 201820 | Accredited ABET Engineering Accreditation Commission |

# Taller 2 – Repositorios NoSQL y análisis básico de contenido

## Objetivo

- Utilizar un entorno de tecnología de almacenamiento NoSQL en la construcción de soluciones altamente escalables para el procesamiento de información.
- Utilizar un repositorio de datos NoSQL para la consulta de documentos sobre *dataset* reales.
- Utilizar herramientas de análisis sintáctico en textos.
- Experimentar con infraestructuras que permiten la escalabilidad de procesamiento a través de la paralelización de procesos

## Prerrequisitos

- Herramientas y lenguajes para desarrollo de aplicaciones y de clientes Web. Por ejemplo, Java, JSP, Phyton, etc.
- Conocimiento básico de Unix y ambientes de virtualización
- Conocimiento básico de la arquitectura y funcionalidad de repositorios NoSQL
- Conocimiento básico de técnicas de modelaje de contenido de documentos.

## Metodología

- Se trabaja de acuerdo con los lineamientos generales del curso.
- Se realiza una entrega por grupo
- Utilice para el documento las pautas de elaboración de documentos técnicos que encuentra en Sicua+.

## Enunciado

El desarrollo de este taller se enmarca en el análisis de una fuente de datos escalable, con contenido textual. Su grupo selecciona temas de coyuntura nacional o internacional, para los cuales sea viable descubrir e incorporar grandes cantidades de información. Por ejemplo, en la coyuntura actual podrían considerarse temas como el paro estudiantil en Colombia, los acontecimientos relacionados con catástrofes naturales como huracanes, los concursos estilo “Yo me llamo”, artistas o deportistas polémicos, etc. Una fuente interesante pueden ser las noticias y columnistas de opinión. El contenido base debe ser producido en español.

Su análisis puede enfocarse en los temas que tienen impacto generalizado sobre la opinión pública. Seleccione por lo menos tres temas generales en los cuales centra su análisis y plantee al menos 5 preguntas “de negocio” que sería interesante resolver sobre dichos datos.

Tenga en cuenta que TODOS los aspectos de su solución deben ser escalables. Su diseño debe siempre ser aplicable en escenarios de Big Data.

1. **Recolección de datos y selección de tecnología NoSQL**
   - Realice un proceso de recolección de datos en Twitter, que le permita configurar un *dataset* lo más completo posible sobre los temas seleccionados. Considere los tuits generados por tuiteros relacionados geográficamente con los temas seleccionados. Para ello, considere cuentas de influenciadores locales y globales, personajes de gran visibilidad, prensa nacional e internacional, etc. Considere mínimo 20 cuentas y por lo menos 10 temas concretos de coyuntura, en el marco indicado anteriormente.

- Debe considerar que la fuente de datos establece límites diarios por IP y por cuenta. Por ello, **evite** hacer la recolección de datos desde la máquina asignada al curso. Dado que está en una subred privada, todas las máquinas salen a internet por la misma IP y el límite será alcanzado muy rápidamente.

- Establezca una tecnología NoSQL con la cual realiza su experimentación. En el *cluster* del curso dispone de MongoDB y HBase. Si desea experimentar con algún otro producto infórmelo en los primeros 3 días de asignación del taller para evaluar la capacidad y factibilidad de instalarlo allí.
- Revise el *dataset* de tuits, estudie el formato de los datos, contenido y alcance.
- Guarde todos los datos en el repositorio seleccionado. La base de datos DEBE TENER como prefijo <GrupoNN> donde NN es el identificador de su grupo.
  - Pueden unir *datasets* generados por varios grupos de forma que constituyan un *dataset* más significante. Deben eliminar elementos repetidos. Esto debe ser acordado previamente por los grupos e informado con anticipación al profesor.
- Construya consultas que le permitan evaluar el alcance del repositorio en su expresividad, forma de procesar los datos y entrega de resultados. Elabore al menos 3 consultas complejas.

## 2. Análisis de coyuntura sobre Twitter

*Visualizar la temáticas, polaridad, localización, tendencias e interacción entre usuarios en un dataset tomado de Twitter.*

- Realice un análisis de polaridad de los tuits.
  - Elabore un análisis de sentimientos en por lo menos 3 niveles (positivo, negativo, neutro). Sería preferible hacer uno en más de tres niveles, para ello busque las herramientas adecuadas. Dados los temas seleccionados, es factible utilizar otra escala de polaridad para el análisis de sentimientos, que sea apropiada y significante.
  - Establezca un subconjunto significativo de datos, realice el proceso de anotación de la polaridad.
  - Construya un modelo que permita la clasificación automática de los tuits, de acuerdo con su modelo de polaridad.
  - Evalúe los resultados obtenidos del proceso automático de análisis.
- Realice un análisis del histórico de seguidores de las cuentas que seleccione.
  - Identifique si hay un cambio significativo de seguidores y analice los contenidos que ellos publican. Identifique si hay seguidores o interacciones con usuarios que son potencialmente identificados como perfiles robot.
  - Haga un análisis de retuits, respuestas y citas para apoyar su conclusión sobre el punto anterior. Incluya en su análisis la característica de tuit favorito (anotado con ♥).
- Haga un análisis de apoyo, contradicción o matoneo hacia los personajes monitoreados, o a quienes se refieren a los temas de coyuntura seleccionados. Para ello:
  - Revise las palabras con las cuales se asocia a los personajes o temas más relevantes, de acuerdo con el contenido de los tuits y las respuestas que reciben o que producen.
  - Establezca una relación entre los términos de tendencia (*Trending topics*) y los temas y personajes que su grupo decide analizar.
  - Revise la relación entre las temáticas del tuit original y las de sus respuestas. Para las respuestas, analice el grado de apoyo o contradicción hacia la persona o hacia el tema.
  - Analice la objetividad de quienes responden, a partir de la relación temática de las respuestas y la polaridad de las respuestas que quien responde suele ofrecer. Para ello es importante que el análisis de polaridad tenga más de 3 niveles.
  - Establezca un modelo que mida el apoyo o matoneo hacia los personajes seleccionados y hacia quienes responden u opinan.
  - Esté al tanto de la evolución de las noticias diarias durante el periodo de recolección de datos y analice qué tanto influyen en la expresión de los tuiteros con respecto a las temáticas y los personajes monitoreados.
  - A partir de los resultados obtenidos, encuentre los *clusters* de apoyo a cada uno de los temas, los *clusters* de rechazo, los clusters de generadores de matoneo o de quienes lo reciben.
  - Identifique los usuarios que concentran su participación en esa red social en intervenciones calificables como matoneo.

- Identifique qué tanto se mencionan en las fuentes de noticias los temas que ocupan a la opinión pública, dentro de sus temas monitoreados.
- Para los personajes públicos o sitios geográficos que sean mencionados, enriquezca la información sobre esa entidad a partir de lo que se encuentra de ella en Wikipedia. Para ello, utilice el API que dicha fuente ofrece.

Para tener en cuenta:

- ✓ En las palabras incluya por una parte los hashtags y por otra parte las palabras significantes encontradas en cada uno de los tuits.
- ✓ Los tuits pueden incluir enlaces a sitios Web. Tenga cuidado de NO hacer crawling sobre sitios que no lo permitan.

### 3. Análisis de polaridad sobre un *dataset* anotado

- Tome el *dataset* que se le entrega, que contiene un conjunto de tuits con polaridad anotada. Revise el formato de los datos.
  - El sitio de descarga de este *dataset* es publicado en Sicua+.
- Realice el proceso de análisis de polaridad básico. Produzca un modelo a partir del *dataset* entregado y compare su resultado con el modelo generado para la segunda parte de este taller.
- Compare sus resultados de análisis de polaridad con el producido por los evaluadores en el *dataset*. Tenga en cuenta que no siempre se tienen resultados completos de los evaluadores.
- Realice un análisis de frecuencia de los usuarios, interacciones, *hashtags* encontrados, citas y menciones.

### 4. Análisis de escalabilidad en NoSQL

Una vez tenga claro y funcional el proceso solicitado en el punto 2, procese los datos en diferentes escenarios de escalabilidad. Para cada escenario documente el tiempo que tardan los procesos de *consulta* y análisis.

- ✓ Utilice una instalación propia del repositorio standalone en su máquina virtual que descargó al inicio del curso. NO lo haga en la máquina de publicación de resultados.
- ✓ Utilice los repositorios que estarán disponibles en el cluster para la experimentación. Revise y documente la configuración que encuentra allí para el repositorio que utilice.
- ✓ Eventualmente puede hacer el análisis de escalabilidad, como plan B, con los resultados del punto 3.

### 5. Visualización de resultados

Desarrolle una aplicación Web que permita:

- Realizar las consultas y visualizar los resultados de forma dinámica. Para ello diseñe una forma de interactuar en forma de consultas sobre su *dataset*. Establezca consultas complejas, que le permitan mostrar su habilidad sobre los datos, el repositorio seleccionado y el análisis de contenido ofrecido.
- Visualizar la información básica de personajes o sitios geográficos mencionados. Debe poder relacionar dichas entidades por algún criterio encontrado en su información de base. Por ejemplo, todos los personajes mencionados que nacen en cierto rango de fechas, sitios que son geográficamente cercanos, etc.
- Visualizar los resultados del análisis de coyuntura. En UNA página de visualización muestre un tagCloud con los *hashtags* más frecuentes, visualizar la polaridad de los tuits relacionados y su contenido.
- Elabore una visualización que le permita ilustrar las relaciones Usuarios – Temas – Actividad en la red social, para aquello que monitorea.
- Elabore una visualización que le permita ilustrar los índices de apoyo, contradicción o matoneo hacia los temas y opinadores, así como la objetividad de quienes responden. Incluya de manera efectiva en la visualización los perfiles robot detectados.
- Visualizar los resultados de análisis de polaridad comparados con los encontrados en el *dataset* entregado.
- El sitio Web debe incluir explícitamente la citación de la fuente del *dataset*, tal como está especificado en la fuente.

### 6. Documentación de resultados

Elabore un documento de máximo 5 páginas en el cual relacione:

- Enlace a la aplicación Web que muestra los resultados.
- Caracterización de los datos recolectados (ficha técnica completa)
- Estrategia de solución, diseño y tecnología concretos utilizados en cada uno de los retos propuestos. Argumentación de cómo se logra un diseño de una solución escalable, en escenarios de Big Data
- Documentación detallada de las estrategias de análisis que utiliza en cada uno de los retos propuestos. Algoritmo básico utilizado para resolver cada uno de los retos, de manera que puedan percibirse los elementos para poner en valor la solución.
- Documentación de la evaluación de la calidad de los modelos de análisis y de la calidad de la solución global
- **Análisis** de resultados obtenidos, dificultades, logros y posibilidades de generalización de la solución. Analice la calidad de los resultados obtenidos desde el punto de vista de la información entregada al usuario. Analice problemas encontrados, mejoras posibles, retos por resolver para hacer un mejor trabajo de entrega de información al usuario. Proponga posibles extensiones de valor agregado

**7. BONO**

- Realice un análisis de geocodificación para los tuits que mencionan ciudades. Tenga en cuenta la geolocalización del tuitero y del tema que se trata, cuando esa información esté disponible.
- En UNA página de visualización muestre un mapa de los tuits geocodificados (sitio y cantidad de tuits), un tagCloud con los hashtags más frecuentes y una visualización de la polaridad que permita ver el contenido de los tuits relacionados.

## RESTRICCIONES

- EN NINGÚN CASO el ejercicio debe hacerse sobre la versión en línea de Twitter.
- NO debe hacer la recolección de datos desde las máquinas de la infraestructura del curso.
- No intente seguir los links que encuentra en los tuits.
- DEBE utilizar estrategias escalables en la solución del problema.
- DEBE utilizar un repositorio NoSQL para el almacenamiento y procesamiento de la información (tanto los archivos fuentes como los resultados).
- Debe realizar la visualización de forma dinámica sobre los resultados obtenidos y almacenados.

- Para realizar los procesos de experimentación de escalabilidad deben reservar turnos para el uso del *cluster*, de manera que sólo un grupo esté procesando al tiempo. Se sugiere programar la ejecución de los Jobs de tal manera que puedan arrancar y dejar los datos sin que requieran intervención personal.
- Los turnos serán dispuestos en horarios nocturnos, cuando baja la carga de los medios de almacenamiento y carga de máquinas virtuales en el *datacenter*.
- La interacción con el usuario debe ser en una aplicación Web gráfica, sencilla pero intuitiva y bien presentada.
- Desarrolle y despliegue la aplicación solicitada en el ambiente UNIX provisto en el curso. En particular, la aplicación de demostración debe correr en la máquina Web prevista para publicación de resultados utilizando el cluster del curso

## Evaluación

- ✓ (5%) Recolección de datos y selección de tecnología NoSQL
  - Recolección y filtraje de datos según los criterios establecidos
  - Consultas que permiten tener una comprensión de lo recolectado
- ✓ Análisis de coyuntura sobre Twitter
  - (15%) Análisis de polaridad de los tuits
    - Anotación de datos
    - Modelo de análisis de sentimientos
    - Evaluación de la calidad del modelo
  - (15%) Análisis del histórico de seguidores de las cuentas, cuentas robot.
  - (20%) Análisis de apoyo, contradicción o matoneo
- ✓ (10%) Análisis de polaridad sobre un *dataset* anotado
- ✓ (5%) Escalabilidad en NoSQL

- ✓ Resultados
  - (15%) Visualización de resultados
  - (15%) Documentación de resultados
- ✓ (10%) BONO

**La solución en línea y la mostrada en el momento de sustentación DEBE CORRESPONDER DE MANERA EXACTA CON LO ENTREGADO EN SICUA+, si hay cualquier diferencia, por mínima que sea, puede ser considerado FRAUDE.**

**Los resultados de su solución deben ser comprobables en el momento de ejecución con los datos que se encuentran en las fuentes correspondientes.**

## Aspectos que el grupo decide

1. Herramientas para visualización Web
2. Lenguaje y ambiente de desarrollo
3. Tecnología NoSQL a utilizar, dentro de las instaladas en el cluster del curso
4. Consultas complejas que permiten expresar el alcance de la solución lograda.

El cumplimiento de las restricciones y de los requerimientos técnicos es parte integral de los entregables. No satisfacerlos invalida TODOS LOS entregables.

No presentar los resultados en la aplicación Web esperada, o no entregar el documento de análisis de resultados, invalida TODOS los entregables.

Se espera que cada miembro del grupo haga una contribución igualmente significante al desarrollo de esta actividad y a las tareas definidas al interior del grupo. El trabajo por debajo de este rango tiene una penalización proporcional sobre la evaluación global de la tarea

Los resultados serán sustentados en sesión de 20 minutos por grupo en horario definido en Sicua+.

## Entregables

Entregue en Sicua+ UN archivo con los resultados de su taller. El contenido debe ser:

1. Proyecto de software desarrollado. Elimine archivos temporales y archivos `.class`
2. Archivo en formato pdf con el informe de documentación de resultados.

Nombre del archivo de la entrega: Taller2_<NN>_<login1>_<login2>_<login3>.zip.

Nombre del archivo de análisis: Taller2_<NN>_<login1>_<login2>_<login3>.pdf

donde NN es el número del grupo y login1 y login2 son los correspondientes a los miembros del grupo en Uniandes.

El no seguimiento del formato de entrega del taller tiene una penalización de **0.5/5.0** en la nota final. La hora de entrega DEBE ser la indicada en el enunciado, independientemente de la disponibilidad eventual posterior del enlace de entrega en Sicua+. El taller se rige por las normas definidas en las reglas de juego de trabajos prácticos. En particular, entregas tardías tienen como evaluación 0.0/5.0.

El cumplimiento de las restricciones técnicas es parte integral de los entregables. No satisfacerlos invalida TODOS los entregables.

Los resultados serán sustentados en sesión de 20 minutos por grupo en horario definido en Sicua+. El grupo COMPLETO tiene UNA oportunidad de sustentación, para lo cual debe tomar oportunamente la cita correspondiente. La no presentación a la sustentación de los resultados hace que la nota final del taller sea 0.0/5.0.

Se espera que cada miembro del grupo haga una contribución igualmente significante al desarrollo de esta actividad y a las tareas definidas al interior del grupo. El trabajo por debajo de este rango tiene una penalización proporcional sobre la evaluación global del taller, de acuerdo con lo establecido en las reglas de juego de trabajos prácticos.

El *dataset* obtenido debe ser puesto a disposición del laboratorio CODICE, con la ficha técnica correspondiente.

**Fecha y hora límite de entrega: lunes 5 de noviembre, 14h00**