Universidad de los Andes Facultad de Ingeniería	Ingeniería de Sistemas y Computación **Maestría** MINE4102 – Análisis de Información sobre Big Data Semestre: 2016-2	Accredited ABET Engineering Accreditation Commission

Taller 2 – Procesamiento escalable de fuentes semiestructuradas

Objetivo

- Utilizar los entornos Hadoop y Spark (opcional) en la construcción de soluciones altamente escalables para la el procesamiento de información.
- Utilizar el esquema de procesamiento *Map Reduce* en la construcción de soluciones altamente escalables para la el análisis básico de información semiestructurada y enlazada.
- Experimentar con infraestructuras que permiten la escalabilidad de procesamiento a través de la paralelización de procesos

Prerrequisitos

- Herramientas y lenguajes para desarrollo de aplicaciones Web. Por ejemplo, Java, JSP, Phyton, etc.
- Conocimiento básico de Unix y ambientes de virtualización
- Conocimiento básico de expresiones regulares (regEx) y descubrimiento de información enlazada.
- Conocimiento básico de Hadoop y Spark (Opcional)
- Conocimiento básico de *Map Reduce*
- Conocimiento básico de técnicas de descubrimiento de información semiestructurada enlazada.

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

1. **Descubrimiento de información en fuentes semiestructuradas utilizando *Map Reduce***
   Construya de forma dinámica un grafo de artículos en Wikipedia relacionados con un criterio de filtrado de datos.

   - Revise los artículos de Wikipedia (versión en inglés) que se relacionan con personajes. Encuentre la estructura básica que permite distinguir que el artículo se refiere a una persona y cómo presenta sus datos biográficos.
   - Construya el grafo de personajes en Wikipedia, nacidos en un rango de fechas dado, de tal forma que los enlace con otros personajes relacionados con ellos a 2 pasos de distancia (sin importar su fecha de nacimiento). Las relaciones pueden darse a uno o dos pasos de distancia, bien sea por personajes o por las fechas.
   - Logre que la visualización del grafo permita ver sus elementos. Cuando se revisa una relación debe encontrarse el detalle que la justifica. Visualice los personajes de acuerdo con su país de nacimiento.
   - Establezca la fortaleza de la relación entre países. Dos países están relacionados cuando personajes que allí nacieron se encuentran relacionados. La fortaleza está dada por la frecuencia de las relaciones en el periodo de tiempo indicado.
   - Los parámetros de consulta sobre la aplicación pueden ser el nombre de un personaje, el país de interés y el rango de fechas. Las consultas pueden incluir todos o alguno de los parámetros. El resultado debe quedar almacenado en el cluster. Pueden ser varios países, si el grupo lo considera interesante.
     - El procesamiento puede hacerlo directamente en Hadoop o utilizando Spark. Su decisión debe estar claramente documentada.
   - Construya una aplicación Web que permita:
     - ✓ Introducir los parámetros de interés y correr el proceso de construcción del grafo de respuesta utilizando Map Reduce.

- ✓ En la medida de lo posible, ver resultados incrementales del proceso, si este tarda mucho en terminar.
- ✓ Tomar un grafo de resultados, que se encuentra en Hadoop y visualizarlo.

RESTRICCIONES

- Para realizar el proceso DEBE utilizar el *data set* de Wikipedia que encuentra en el *cluster* Hadoop asignado al curso. La ruta del archivo será publicada en Sicua+.
- EN NINGÚN CASO debe hacerse sobre la versión en línea de Wikipedia. EN NINGÚN CASO debe hacerse copia alguna del data set ni deben descomprimirlo. Hacer esto compromete la viabilidad de TODO el taller para el curso completo. DEBE procesarse en su estado comprimido, tal como lo encuentran.
- DEBE utilizar estrategias *Map Reduce* escalables en la solución del problema.
- DEBE utilizar Hadoop como repositorio de la información (tanto los fuentes como los resultados).
- Debe realizar la visualización de forma dinámica sobre los resultados obtenidos y almacenados en Hadoop.

Entregable

- Muestre los resultados solicitados en una aplicación Web sencilla, que ofrezca las ***funcionalidades*** solicitadas. No olvide relacionar el número de grupo y sus integrantes en la página Web de resultados.
- Elabore un documento de **máximo 3 páginas** en el cual relacione:
  - Forma de acceso a la aplicación Web resultante.
  - Carpeta donde se encuentran los resultados del procesamiento realizado a través de la aplicación Web
  - **Métodos y tecnología** concretos utilizados en cada uno de los retos propuestos
  - **Estrategia Map Reduce** que le utiliza en la solución y tecnología utilizada.
  - El **algoritmo básico** para resolver cada uno de los retos, de manera que puedan percibirse los elementos interesantes para poner en valor en la solución
  - **Análisis de resultados obtenido.**

Aspectos que el grupo decide

1. Herramientas para visualización Web de grafos
2. Lenguaje y ambiente de desarrollo

Requerimientos técnicos

1. La interacción con el usuario debe ser en una aplicación Web gráfica, sencilla pero intuitiva y bien presentada.
2. Desarrolle y despliegue la aplicación solicitada en el ambiente UNIX provisto en el curso.

Evaluación

La evaluación se hace así:

Entregable	Porcentaje en la evaluación
Proyecto de software desarrollado, aplicación funcional y demostración	70%
Informe y sustentación	30%

El cumplimiento de las restricciones técnicas es parte integral de los dos entregables. No satisfacerlos invalida LOS DOS 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 10 minutos por grupo en horario definido en Sicua+.

Entregables

Fecha de entrega: **septiembre 22 de 2016, a las 14:00**

Entregables

.

Archivo de la entrega: <Taller2_NN_login1_login2_login3>.zip.

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

Contenido: Archivo zip con el proyecto de software y archivo .pdf con el informe de análisis. Nombre del archivo de análisis: Taller2_NN_login1_login2_login3.pdf

El no seguimiento del formato de entrega del taller tiene una penalización de **0.5/5.0** en la nota final. La no presentación a la sustentación de los resultados produce una nota final en la tarea de 0.0/5.0. El grupo COMPLETO tiene UNA oportunidad de sustentación.