Nombre del Proyecto	Word Count / Tokenizer
Fecha de Entrega (Final)	Semana 10
Fecha de Entrega (Avance Inicial)	Semana 5
Método de Entrega	Subir código fuente y datos a la plataforma Presentación presencial
Valor	Ver sílabo
Método de Desarrollo	3 integrantes por grupo

Descripción

El objetivo de este trabajo es aprender cómo usar la aplicación de **Map/Reduce** en un **sistema de archivos distribuidos de Hadoop**, desarrollando conocimientos básicos sobre cómo encontrar conjuntos de elementos frecuentes para realizar posteriores análisis para minería de datos.

Para poder llevar a cabo el proyecto se deberá desarrollar dos diferentes aplicaciones:

- WordCount
- Frequencyanalysis

Se proporcionará un archivo de 1 semana de consulta en Google como ejemplo para encontrar conjuntos de elementos frecuentes. El archivo de consulta deberá ser analizado, reducido y procesado utilizando las aplicaciones mencionadas anteriormente. Debido al gran tamaño del dataset y la gran cantidad de consultas que se procesarán, **se sugiere un soporte mínimo de 5000 (Minimum Support).**

Referencias Adicionales para el Proyecto:

- Se deberá realizar el **procesamiento** del dataset sobre una implementación de Hadoop (a realizar por el estudiante).

Sugerencias de Programación

- Se sugiere desarrollar el proyecto en Java.

Aplicación

WordCount

En la elaboración de este proyecto, se deberá entregar la aplicación WordCount. Esta aplicación básicamente recibe un archivo de texto y devuelve otro archivo que enumera cada palabra encontrada en el archivo de entrada y la cantidad de veces que dicha palabra apareció.

La aplicación deberá usarse dos veces en dicho proyecto:

- Para contar la cantidad de veces que cada palabra aparece en el archivo de consulta.
- Para contar la cantidad de veces que cada conjunto de dos elementos aparece en el archivo de consulta.

Frequency Analysis

Se deberá desarrollar una aplicación para poder lograr los objetivos del proyecto y encontrar los conjuntos de dos elementos frecuentes en el archivo de consulta. La aplicación deberá contener tres clases:

- Main: deberá ejecutar todo el proceso
- DataPreprocessor: deberá pre-procesar el archivo de consulta y reducirlo para facilitar el manejo.
- FrequencyAnalyzer: deberá de encontrar los conjuntos de elementos uno-frecuente y dos-frecuente en el archivo de consulta.

Método de Evaluación

- Preprocesamiento (15%)
    - No es válido incluir los términos escritas en duro en el fuente (datos quemados) ´hardcoded´
        - Deben incluir un diccionario de datos.
        - El no hacerlo dará lugar a una penalización de 100% en el preprocesamiento.

- Sobre programas entregados, habrá créditos parciales de acuerdo a la ponderación establecida en esta sección.
    - Implementación Hadoop (30%).
        - Cluster (10% adicional)
    - Realizar las validaciones y pre procesos en el dataset (15%).
    - Se evaluará la utilización de estándares de programación.

    - Se evaluará la utilización y desarrollo de las estructuras de datos / librerías / sistemas de archivos especificadas en este proyecto. No utilizar, desarrollar e implementar estas estructuras para que cumplan la función para la cual fueron definidas, será penalizado (20% por cada una).
    - Se tomará en cuenta la creatividad. (10% extra sobre el total del proyecto)
        - Conclusiones y recomendaciones sobre hallazgos interesantes.
    - Presentación de resultados interesantes: (20%)
        - Top 10 o Top 20 itemsets de 1 palabra
        - Top 10 o Top 20 itemsets de 2 palabra
    - No respetar el minimum Support definido dará lugar a una penalización de 100%

- Cada grupo deberá realizar tres críticas puntuales y constructivas de la presentación realizada por el resto de compañeros.
    - Puntos de mejora.
    - Ponderación (en base a 100%).

Entregas Tardías

- No se aceptarán proyectos después de la fecha de entrega.
- No se aceptarán reclamos sobre el proyecto si no se presentan a revisión en la fecha estipulada.

Fase Inicial de Revisión de Proyecto

De acuerdo con lo estipulado en el sílabo, se tendrá revisión preliminar del proyecto.

- 25% de la nota total.
- Presentar un avance concreto
    - Implementación de Hadoop y al menos 3 datasets candidato para MapReduce

Método de Presentación

- Cada grupo deberá realizar una pequeña presentación (15 – 20 mins) mostrando todas las etapas de desarrollo de su proyecto:
    - Introducción
    - Presentación de las aplicaciones desarrolladas
    - Mostrar paso por paso la ejecución de su aplicación
    - Resultados y conclusiones
    - Hallazgos interesantes