Universidad de los Andes Facultad de Ingeniería	Ingeniería de Sistemas y Computación **Maestría** MINE4102 – Análisis de Información sobre Big Data Semestre: 201820	Accredited ABET Engineering Accreditation Commission

Taller 1 – Procesamiento escalable

Objetivo

- Utilizar los entornos Hadoop y Spark en la construcción de soluciones altamente escalables para 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 de expresiones regulares (regEx) y descubrimiento de información enlazada.
- Conocimiento básico de Hadoop y Spark
- Conocimiento de *Map Reduce*

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

En este taller usted cumple el rol de quien toma decisiones de negocio con respecto a la planeación y regulación del sistema que opera con la información de la cual dispone. Usted debe desarrollar herramientas que resuelvan las preguntas de negocio indicadas, mediante técnicas de procesamiento escalable.

1. **Entendimiento de los datos**

   Para este taller trabaja sobre el dataset público correspondiente al registro de viajes en taxi en la ciudad de New York durante varios años. Lo encuentra en el cluster de trabajo en el curso y puede ser descargado en forma libre de: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml. En el sitio Web del publicador de los datos encuentra los diccionarios de datos que los describen. Analice el contenido de esos datos antes de iniciar el desarrollo del taller, entienda los valores, la estructura y la variedad en los registros a medida que pasa el tiempo.

2. **Consultas básicas utilizando *Map Reduce y Hadoop***

   Resuelva las siguientes preguntas de negocio utilizando técnicas de Map Reduce sobre Hadoop:

   RF 1. Encuentre cuál es el sitio de la ciudad hacia el cual se dirigen la mayor cantidad de vehículos en una cierta franja de horas del día, para cada día de la semana.

   RF 2. Encuentre cuál es el valor promedio de un viaje, para los viajes realizados en día domingo, en un mes dado.

   RF 3. Encuentre los destinos de los viajeros que salen de cada uno de los aeropuertos en una franja horaria y día de la semana, para un aeropuerto dado.

RF 4. Encuentre los sitios donde hay mayor demanda de viajes en cierta franja horaria y día de la semana. Muestre esos datos discriminados para cada uno de los meses que se encuentran en el dataset

3. Análisis de los datos utilizando estrategias *Map Reduce*

Los siguientes son requerimientos analíticos sobre los datos indicados. Para resolverlos, primero entienda los datos, planee su método de solución utilizando técnicas escalables, diseñe su estrategia de implementación y luego sí proceda a desarrollarla.

Usted debe desarrollar **DOS** de ellos, utilizando Map Reduce. Puede escoger cuáles desarrollar. También, puede escoger implementarlos sobre directamente sobre Hadoop o Spark.

RA 1. Construya una herramienta (por ejemplo, un tablero de control básico) que permita, de forma dinámica, encontrar los patrones de viaje de los usuarios de los aeropuertos de la ciudad de NY.

Analice los datos con respecto a origen del viaje, longitud del trayecto, medio de pago, valor pagado, horas pico y valle, día de la semana, cantidad de pasajeros, distancias recorridas vs. precios.

De esta forma, el tomador de decisiones puede llegar a responder preguntas como las siguientes:

¿Qué tanto se afecta el precio del viaje por cuenta de las congestiones de tráfico en horas pico?

¿Los viajes se hacen siguiendo rutas típicas, o, eventualmente, entre los mismos origen y destino hay viajes con distancias recorridas significativamente diferentes?

¿Hay relación espacial entre el origen de los pasajeros y el aeropuerto que utilizan, o los pasajeros atraviesan la ciudad permanentemente?

¿En qué momento del día hay mayor cantidad de vehículos circulando con un solo pasajero, y de dónde provienen? Si esto pasa, tendría sentido hacer rutas de buses, por ejemplo…

RA 2. Construya una herramienta que permita construir matrices Origen-Destino, de forma dinámica, por franjas horarias. Debe permitir encontrar los puntos de atracción en esa franja (son generadores de muchos viajes, bien sea en origen como en destino). Debe ser posible, de forma interactiva, mover las franjas horarias y ver cómo "se mueve" la ciudad. El movimiento de franjas horarias debe permitir cambiar de día de la semana o de día del mes.

La herramienta debe ser permitir identificar, para los puntos de atracción de viajes, cuáles son los sitios de la ciudad que pueden estar relacionados. Por ejemplo, presencia de hospitales, teatros, estadios, parques, escuelas, sitios turísticos, universidades, etc. Así mismo, es interesante ver si hay relación entre los puntos de atracción y el día del mes. Por ejemplo, si los días de pago o los festivos se comportan diferente a los demás.

Los datos que encuentra en sitios como: https://data.cityofnewyork.us/Health/Areas-of-Interest-GIS/mzbd-kucg pueden ser de utilidad.

RA 3. Realice un análisis de oferta y demanda en temporadas. Una temporada es, por ejemplo, en tiempo de verano, en navidad, en celebración de Acción de Gracias, San Valentín, etc.

Determine si hay estacionalidad en oferta y demanda por tipo de servicio. Determine cómo se afectó la estacionalidad con la entrada de los servicios tipo Uber.

RA 4. **(BONO)** Complemente sus análisis anteriores relacionando datos de clima, para revisar qué tanto se afectan los viajes en cuanto a demanda, origen-destino, etc. Complemente su análisis con la realización de eventos en la ciudad.

4. Visualización de resultados

Muestre los resultados solicitados en una aplicación Web sencilla, que ofrezca las funcionalidades solicitadas, en una interacción con el usuario efectiva, intuitiva y bien presentada. No olvide relacionar el número de grupo y sus integrantes en la página Web de resultados.

La aplicación Web debe permitir:

- ✓ Introducir los parámetros de interés y ejecutar los requerimientos correspondientes.
- ✓ Visualizar adecuadamente las respuestas, de forma que se logre validar la respuesta contra la fuente de datos original.

- ✓ La visualización debe permitir ver identificar los elementos y su detalle. Cuando se revisa un elemento o una relación debe encontrarse el detalle que la justifica. Permita mover fácilmente el eje del tiempo
- ✓ En la medida de lo posible, ver resultados incrementales del proceso, si este tarda mucho en terminar.

5. Escalabilidad de procesos con Hadoop y *Map Reduce*:

Ejecute las consultas básicas (RF1 a RF4) en escenarios diversos de escalabilidad y documente sus resultados de experimentación. Para la ejecución de los escenarios de experimentación se establecen franjas de ejecución exclusiva sobre los clusters, de manera que cada grupo pueda realizar las mediciones propias a su solución, sin interferencias. Estas franjas se coordinan según políticas definidas en clase.

Una vez tenga claro y funcional el desarrollo de cada uno de los requerimientos indicados, procese los datos en diferentes escenarios de escalabilidad. Para cada escenario documente el tiempo que tardan los procesos map y reduce respectivamente.

Establezca un *sub-dataset* de alrededor del 50% de los datos. Estos datos puede almacenarlos en su espacio en Hadoop. Documente cuál *sub-dataset* utiliza.

a. Utilice un *cluster* de 1 nodo y el *sub-dataset*. Instale este *cluster* en su máquina individual de resultados.
b. Utilice un *cluster* de 4 nodos y el *sub-dataset* definido para el paso anterior.
c. Utilice un *cluster* de 4 nodos y el *dataset* completo.
d. Utilice un *cluster* de 20 nodos y el *dataset* completo.

Analice sus resultados comparativos así:

- Escenario a y b
- Escenario b y c
- Escenario c y d

6. (BONO) Impacto de la infraestructura de desarrollo

Seleccione **UNO** de los requerimientos funcionales RF1 a RF4.

- Tome LA MISMA estrategia de solución que estableció en el punto 2, sobre Hadoop y Map Reduce. Ahora, desarrolle ESA estrategia utilizando HIVE y Spark.
- Para el desarrollo en HIVE, cada grupo debe crear y poblar su propio esquema, como tablas externas. NO debe duplicar los datos.
- Compare los resultados obtenidos considerando los siguientes criterios: dificultad de implementación, desempeño, escalabilidad de la solución.

7. Documentación de resultados

Elabore un documento de **máximo 6 páginas** en el cual relacione:

- ✓ Enlace a la aplicación Web que muestra los resultados.
- ✓ **Métodos y tecnología** concretos utilizados en cada uno de los retos propuestos
- ✓ **Documentación detallada de las estrategias Map Reduce** que utiliza en la solución
- ✓ Elementos interesantes para poner en valor en la solución, elementos no logrados.
- ✓ **Análisis de resultados obtenidos**: análisis de la escalabilidad de cada una de las soluciones planteadas, 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.

Enlaces interesantes

Anonymizing NYC Taxi Data: Does It Matter?
NYC Subway Data - Do more people ride the New York City subway when it's raining versus not raining?
If Taxi Trips were Fireflies: 1.3 Billion NYC Taxi Trips Plotted

RESTRICCIONES

I. Para realizar el proceso DEBE utilizar el *dataset* que encuentra en el *cluster* Hadoop asignado al curso, en la carpeta data.

II. EN NINGÚN CASO debe hacerse sobre la versión en línea de los datos.

III. EN NINGÚN CASO debe hacerse copia alguna del *dataset* ni deben descomprimirlo. Hacer esto compromete la viabilidad de TODO el taller para el curso completo. DEBE procesarse tal como lo encuentran.

IV. DEBE utilizar estrategias *Map Reduce* escalables en la solución del problema.

V. DEBE utilizar Hadoop como repositorio de la información (tanto las fuentes como los resultados).

VI. Debe realizar la visualización de forma dinámica sobre los resultados obtenidos y almacenados en Hadoop.

VII. Desarrolle y despliegue la aplicación solicitada en el ambiente UNIX provisto en el curso.

Recomendaciones

Diseñe este componente de forma que pueda ser extensible, configurable e integrable con otros entornos Web. Estos entornos pueden ser talleres o tareas posteriores u otros sitios web que encuentren este servicio interesante y la solución apropiada.

Evaluación

La evaluación se hace siguiendo los lineamientos establecidos para los trabajos prácticos y talleres del curso. Los criterios de evaluación son los siguientes:

1. (20%) Desarrollo de los RF1 a RF4, punto 2 del enunciado (punto 2)
   - ✓ Interpretación de los datos, diseño adecuado de procesos Map Reduce básicos
   - ✓ Ingreso adecuado de parámetros y despliegue adecuado de respuestas
   - ✓ Escalabilidad de la solución.
2. (40%) Desarrollo de DOS de los RA1 a RA3 (punto 3)
   - ✓ Realización de las consultas, de forma dinámica, a partir de los criterios establecidos
   - ✓ Manejo de rangos de consulta deslizables, de forma dinámica
   - ✓ Manejo de la escalabilidad de las consultas
3. (10%) Visualización (punto 4)
   - ✓ Visualización adecuada de los resultados de las consultas
   - ✓ Visualización adecuada de los elementos que permiten verificar, DESDE LA APLICACIÓN WEB, la validez de los resultados mostrados
4. Resultados
   - ✓ (15%) Análisis y completitud de la experimentación de escalabilidad de procesos con Hadoop y Map Reduce (punto 5)
   - ✓ (15%) Informe completo, decisiones técnicas descritas y justificadas, análisis pertinente y adecuado (punto 7)
5. Bonos

   Los bonos se tienen en cuenta únicamente cuando **TODOS** los elementos esperados del taller están completamente desarrollados.
   - ✓ (10%) RA4
   - ✓ (5%) Análisis de los resultados sobre infraestructura (punto 6)

Para cada uno de los elementos solicitados, en la sustentación se evalúa que cada uno de los integrantes del grupo muestren dominio sobre el producto, los resultados alcanzados, los resultados no logrados y la aplicación de conceptos.

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.

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: Taller1_<NN>_<login1>_<login2>_<login3>.zip.

Nombre del archivo de análisis: Taller1_<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 30 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.

Fecha y hora límite de entrega: lunes 24 de septiembre, 14h00