Actividad grupal: Análisis exploratorio con Apache Hive sobre HDFS

**Objetivos**

A través de esta actividad, los alumnos podrán poner en práctica los conceptos estudiados en las clases de teoría acerca de la tecnología Apache Hive, aplicándolos a un análisis exploratorio de un conjunto de datos. El objetivo es comprobar que dicha tecnología es muy propicia para realizar análisis sin necesidad de conocimientos de programación, al contrario de lo que ocurría en las actividades anteriores, sino que es suficiente con conocer el lenguaje SQL, al que muchos profesionales ya están acostumbrados.

Por tanto, Apache Hive se erige como una tecnología accesible para un público más amplio, que facilita la consulta de datos almacenados en sistemas de almacenamiento distribuidos, como HDFS, caso que ocupará esta práctica. Además de esto, se pretende que los estudiantes desarrollen su creatividad y sus capacidades como científicos de datos. Para ello, se enfrentarán a un nuevo conjunto de datos que tendrán que procesar para, posteriormente, utilizar su propia iniciativa a fin de extraer conclusiones de ellos sin una guía exacta de qué se pide sobre este análisis.

**Descripción de la actividad**

El grupo deberá, en primer lugar, analizar el conjunto de datos propuesto para encontrar cualquier tipo de incoherencia (valores faltantes, fuera de rango…) y tratarla de la forma más adecuada. Después, habrá de realizar un segundo análisis, desde una perspectiva de negocio, para plantearse y, al mismo tiempo, responder las preguntas que consideren los miembros que más valor aportan a un departamento especializado en el sector en cuestión.

No se requiere un análisis de negocio en profundidad ni pormenorizado, sino solamente practicar y comprobar cómo el conocimiento general adquirido en la asignatura es suficiente para aportar valor desde el punto de vista de negocio, atendiendo exclusivamente a lo que dicen los datos. Un análisis básico basta si se consigue extraer alguna conclusión relevante sobre estos, lo cual será puntuado muy positivamente.

Dado el *dataset* anexo, junto con el planteamiento y la descripción del problema que también se adjuntan, se pide:

* Subir los ficheros de datos al sistema de ficheros local.
* Crear una carpeta en el directorio raíz de HDFS con un nombre de grupo a vuestra elección (y diferente del propuesto). Por ejemplo: «/analytics\_team/».
* Copiar los ficheros de datos subidos inicialmente al sistema de ficheros local en la carpeta recién creada en HDFS.
* Crear una tabla por cada uno de los ficheros de datos copiados en HDFS, con el esquema más adecuado. Al menos, una de las tablas debe ser externa.
* Mostrar las cinco primeras filas de cada tabla cargada. Adjuntar una captura de pantalla con la consulta y resultado.
* Contar el número de filas de cada tabla. Indicar este número en el informe.
* Estudiar el rango (máximo y mínimo) de cada variable numérica. Indicar estos datos en el informe.
* Estudiar las diferentes categorías de las principales variables categóricas y el número de filas correspondientes a cada categoría. Indicar las conclusiones en el informe.
* Buscar valores inexistentes o anómalos (por ejemplo, valores negativos en variables que solo pueden ser positivas; valores excepcionalmente altos o excepcionalmente bajos o negativos) en las principales variables. Indicar cualquier anomalía encontrada.
* Realizar alguna operación *join* entre las tablas creadas que el equipo considere relevante. Guardar el resultado en vistas y **explicar en el informe el valor añadido de realizar dicho(s) *join(s)*.**
* Obtener cualquier agregación (medias, sumas, desviaciones típicas…) que el equipo considere relevante desde una perspectiva de negocio, guardarla en vistas y **explicar en el informe el motivo y las conclusiones que se deducen**.
* Obtener cualquier agrupación y agregación (cualquier agregación aplicada en diferentes grupos) que el equipo considere relevante desde una perspectiva de negocio, guardarlas en vistas y **explicar en el informe el motivo y las conclusiones que se deducen**.

**Entregable**

* Fichero *apellidos\_nombre.txt* con todas las consultas SQL realizadas para resolver los apartados anteriores. Se recomienda incluir también el nombre y los apellidos del alumno dentro del *script*, así como los del resto de los miembros del equipo, como comentario en la primera línea. Cada alumno tendrá que realizar una entrega individual, aunque se entregue el mismo fichero txt.
* Informe en pdf del proceso llevado a cabo, **con énfasis en las capturas requeridas y en los aspectos de negocio, así como en la explicación de las conclusiones derivadas de las consultas realizadas**. Cada análisis se ha de vincular al apartado correspondiente en el *script* entregado.

**Rúbrica**

|  |  |  |  |
| --- | --- | --- | --- |
| Análisis exploratorio con Apache Hive sobre HDFS | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Criterio 1 | La gestión de directorios y ficheros en HDFS es correcta. | 2 | 20 % |
| Criterio 2 | Las consultas SQL y las capturas de funcionamiento son correctas. | 3 | 30 % |
| Criterio 3 | Las preguntas de negocio son relevantes. | 2 | 20 % |
| Criterio 4 | El informe es claro y conciso, y aporta conclusiones claras a partir de los datos extraídos de las consultas en Hive | 2 | 20 % |
|  |  | **9** | **90 %** |

El **Foro de esta actividad** tiene un peso del **10 %**. Cada grupo debe presentar un resumen del desarrollo de su trabajo en el foro de la actividad. Cada miembro debe explicar su aportación individual al grupo. Además, se requiere que cada estudiante haga **al menos dos sugerencias**, críticas o comentarios al trabajo de otros grupos.

**Extensión** máxima del informe en pdf: 5 páginas, fuente Calibri 12, interlineado 1,5.

**Entrega de la actividad grupal**

Al finalizar la actividad grupal, **todos los miembros del equipo entregarán la misma actividad** a través del apartado «Envío de actividades» del aula virtual. El informe y el *script* deben ir nombrados, respectivamente, como se muestra a continuación, sin tildes ni apóstrofes ni ningún otro carácter que pudiera resultar conflictivo:

* APELLIDO1\_APELLIDO2\_NOMBRE\_actividad\_grupal.pdf
* APELLIDO1\_APELLIDO2\_NOMBRE\_actividad\_grupal.txt

Todos los miembros del equipo deben hacer la entrega en el aula virtual y han de adjuntar el mismo documento.

Se debe indicar en el informe y en el *script* el nombre de todos los componentes del equipo. En el informe, además, se adjuntará cumplimentada la siguiente tabla de valoración individual:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Sí | No | A veces |
| Todos los miembros se han integrado en el trabajo del grupo. |  |  |  |
| Todos los miembros participan activamente. |  |  |  |
| Todos los miembros respetan otras ideas aportadas. |  |  |  |
| Todos los miembros participan en la elaboración del informe. |  |  |  |
| Me he preocupado por realizar un trabajo cooperativo con mis compañeros. |  |  |  |
| Señala si consideras que algún aspecto del trabajo en grupo no ha sido adecuado. |  |  |  |