Letra Obligatorio Taller Big Data.

Para el obligatorio deberá utilizar las herramientas usadas en el curso, levantadas en la máquina virtual de Azure. Deberá crear un conjunto de datos tabulares con por lo menos 4 tablas y deberá seleccionar 2 preguntas relativas a los datos que seleccionó, para contestarlas con las herramientas analíticas. La realidad sobre la cuál se va a trabajar es la expresada en el caso de estudio “Una comunidad para compartir desafíos deportivos - Data Fitness”.

- Se debe armar el conjunto de datos con https://www.mockaroo.com/.
- Se deberán ingestar los datos vía NiFi al HDFS creando una ubicación específica para los mismos.
- Luego se deberá realizar un análisis exploratorio de los datos identificando el tipo de datos que hay en cada columna y que significado tienen dentro del dominio de los datos. Dentro de un Jupyter notebook se mostrará una vista previa de las primeras filas, cantidad de columnas de cada tabla, nombre de cada columna, descripción de los datos de cada tabla, cómo está compuesto el esquema de los datos, revisar valores nulos o faltantes y limpiarlos si es necesario. Estos nuevos archivos se guardarán en el hdfs, en otra ubicación como la versión refinada de los datos.
- Una vez que termine con la exploración y limpieza de datos, deberá elegir una forma de modelarlos, esta puede ser, Normalizada, Diagrama Estrella, Data Vault, o OBT. Si para el modelo seleccionado necesita crear nuevos archivos, deben guardarse en una nueva ubicación dentro de HDFS.
- Después de elegir una forma de modelar los datos, deberá bosquejar cómo sería el esquema que relaciona las tablas (se sugiere usar draw.io, pero puede utilizar otra herramienta).
- Una vez tenga los dataframes de las tablas que va a utilizar en su modelo de datos, deberá guardarlas como tabla de HIVE en un esquema nuevo.
- A partir de las tablas de Hive se contestaran las preguntas seleccionadas en un nuevo notebook. Deberá utilizar spark con sintaxis SQL o con los distintos métodos que provee pyspark.
- También se deben crear visualizaciones dentro del notebook que ayuden a responder las preguntas seleccionadas y un reporte del estilo dashboard con la herramienta SuperSet.

La entrega final consiste en un informe donde se muestre mediante capturas de pantalla lo realizado. Se debe explicar el dominio de los datos seleccionados, qué significa cada columna de cada tabla y qué tipo de dato debe tener. Se plantearán y responderán las preguntas seleccionadas y se explicará la arquitectura del datalake. las tecnologías utilizadas y cómo contribuyen a su solución de big data para hacer análisis de datos. También se deben entregar los notebooks utilizados, los dashboards creados en SuperSet y todo lo que resulte relevante para el trabajo que los alumnos consideren necesario.
El grupo deberá ser de un máximo de tres integrantes.