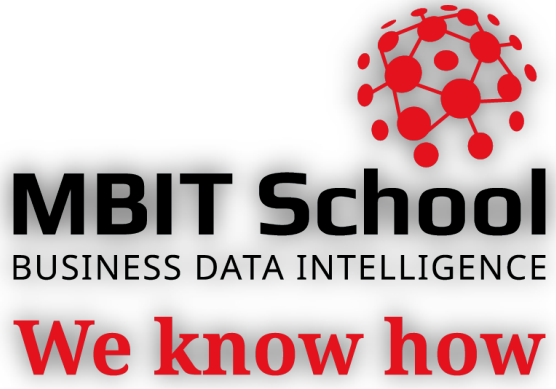

Módulo 2. Ecosistema Spark

Entregable: Práctica Apache Spark

Programa: Máster Executive en Big Data, Cloud y Analytics

Periodo académico: 2019-2020

Autor/es: Jacinto Arias

1. Descripción del trabajo

Disponemos de un stream de AWS kinesis que recibe datos sobre la meteorología de cada provincia española cada 15 minutos.
El objetivo el trabajo será el de crear una libreta de databricks que se conecte al stream y descargue, pre-procese y almacene los datos conforme a una serie de tareas.

2. Tareas

1. Creación de la libreta de databricks

Crea una libreta de databricks. Esta libreta será el entregable que tendrás que subir al campus. Es importante que la edites bien, añadiendo código markdown para documentar y que sea agradable de leer. Parte de la evaluación dependerá de lo bien estructurada que esté la liibreta!

2. Conexión al stream

Aprovecha el código de la práctica de spark streaming para conectarte al stream. Deberás configurar spark utilizando las mismas claves y configuración utilizadas para el stream de kinesis de twitter que vimos en clase. Tan solo deberás cambiar el nombre del stream por `mbit-weather`

3. Preprocesamiento

Siguiendo con las prácticas vistas en clase necesitaremos preprocesar el stream para extraer los datos en json desde el propio stream. Para ello necesitaréis un esquema de los datos.

Los datos de este stream tienen el siguiente formato:

```
{
  'created_at': number,
  'name': string,
  'lon': number,
  'lat': number,
  'weather': {
    'clouds': number,
    'rain':  number,
    'wind_speed': number,
    'humidity': number,
    'temperature': number,
    'status': number
  }
}
```

Por ejemplo:

```
{
  'created_at': 1579865281,
  'name': 'Valencia',
  'lon': -0.38,
  'lat': 39.47,
  'weather': {
    'clouds': 40,
    'rain': 0,
    'wind_speed': 1.5,
    'humidity': 87,
    'temperature': 285.26,
    'status': 'Clouds'
  }
}
```

Debes configurar el esquema y extraer estos datos a un dataframe en streaming. Este dataframe deberá ser plano y tener el siguiente esquema en Spark:

```
root
 |-- created_at: timestamp (nullable = true)
 |-- name: string (nullable = true)
 |-- lat: double (nullable = true)
 |-- lon: double (nullable = true)
 |-- temperature: double (nullable = true)
 |-- rain: double (nullable = true)
 |-- wind_speed: double (nullable = true)
 |-- clouds: double (nullable = true)
 |-- humidity: double (nullable = true)
 |-- status: string (nullable = true)
```

4. Limpieza y aumentado

Los grados de la variable temperatura están indicados en grados Kelvin, para poder continuar transformaremos la columna para que los datos pasen a grados Celsius.

5. Agregación en tiempo real

Configura una agregación del dataframe para que nos muestre los siguientes datos en ventanas de una hora para cada municipio:

- Temperatura media
- Temperatura máxima
- Temperatura minima
- Humedad media
- Número de observaciones

6. Serialización

Los datos de kinesis están disponibles solo 24 horas en el stream. Para poder mantener un histórico vamos a serializar los resultados. Para ello primero aumentaremos nuestro dataframe añadiendo las columnas `month` y `day` estas columnas se extraerán de la hora de inicio de la ventana.

Pista: podeis utilizar las funciones `month` *y* `dayofmonth`

Una vez obtenidas estas dos columnas almacenaremos el dataframe en formato parquet en algún directorio de DBFS. Particionaremos el dataframe por estas dos nuevas columnas.

7. Analítica

Por último vamos a obtener dos vistas, ahora sobre el dataframe serializado:

- Agrupación por días: Un dataframe agrupado por municipio y día del mes y otro agreagando solo por municipio y mes. Obtendremos las siguientes variables agregadas
    - Temperatura media
    - Termperatura máxima
    - Temperatura mínima
    - Humedad media
    - Número de elementos agregados
- Mostrar los 10 municipios con las temperaturas más bajas
- Mostras los 10 municipios con las temperaturas más altas.

8. Extras

Si quieres practicar y subir nota, ahí van algunos ejercicios extra.

Una vez terminada la libreta, ejecuta el stream durante varios días para recopilar nuevos dato y ver como se va poblando el fichero serializado.

Recuerda que los datos del stream cambian cada 24 horas. Para mostrar este resultado de cara a la entrega, lista y muestra la carpeta con el fichero serializado. También lo demostrarás al hacer la agregación, ya que habrá datos de más días!

"Analítica Freestyle" Practica con Spark y añade algún resultado interesante que puedas sacar con estos datos utilizando lo visto en clase. Más interesante si utilizas otro tipo de columna, sugerencias: latitud y longitus, "status" que representa la descripción del tiempo…

Recuerda explicar bien usando markdown cualquier extra que hagas para poderlo identificar.

9. Entrega

Exportar y subir el resultado como un notebook de databricks (fichero dbc). El trabajo puede hacerse hasta en grupos de 3. Todos los alumnos del grupo tendrán que subir el trabajo e indicar con quien lo han realizado.