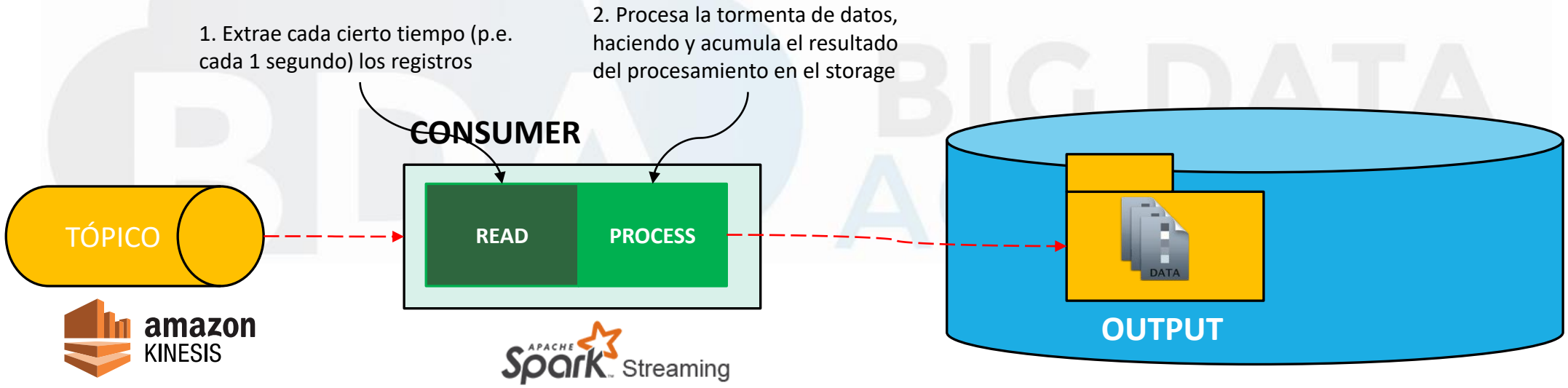
Arquitectura de procesamiento en Real-Time

BIG DATA ACADEMY

Alonso Melgarejo [alonsoraulmgs@gmail.com]

Arquitectura de Procesamiento en Real-Time

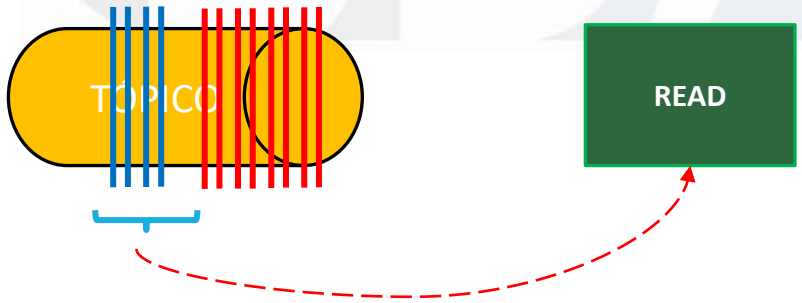
El consumer es el encargado de extraer los registros desde el tópico, procesarlos y almacenar el resultado del proceso en el storage.



Formas de lectura de datos en real-time

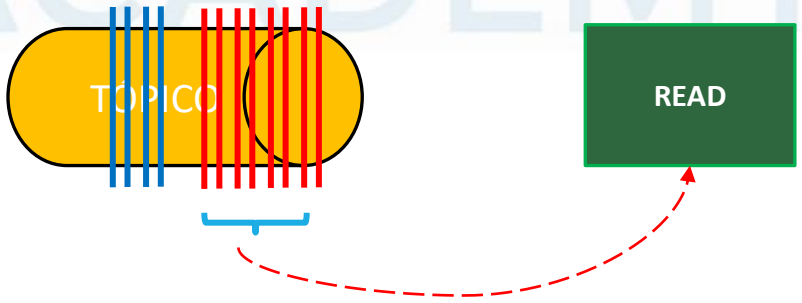
Lectura “latest”

Al iniciar el proceso, se ignoran todos los registros que el tópico tenga previamente y **se procesan sólo los nuevos registros que vengan**



Lectura “earliest”

Al iniciar el proceso, **se procesan todos los registros que el tópico tenga previamente** y luego se procesan los nuevos registros que vengan



Almacenamiento de datos en real-time

Luego de procesar los registros, el dataframe resultante lo guardamos en el sistema de archivos

1. Escribimos el dataframe en real time
2. El dataframe se escribe en “parquet”
3. Directorio donde se escribe el archivo resultante
4. Modo de escritura en el archivo
5. Directorio de checkpoint para recuperar los registros si es que se pierden en la red
6. Cada cuento se ejecuta el proceso de tiempo real
7. Función que ejecuta el proceso en bucle infinito

```
dfResultado.
writeStream.
format("parquet").
start("dbfs:///FileStore/resultado").
outputMode("append").
option("checkpointLocation", "dbfs:///FileStore/resultado/_checkpoints/resultado").
trigger(processingTime = "1 second").
awaitTermination()
```