PARCIAL II – Big Data e Ing. de datos -Máximo 2 estudiantes

1. Desarrollar un sistema de scraping diario a las páginas de bbc.com y cnn.com utilizando AWS Lambda y Zappa:

a) Debe descargar cada día la página principal de cada periódico.

b) En S3 debe quedar la información con la estructura:

• s3://bucket/headlines/raw/periodico=xxx/year=xxx/month=xxx/day=xxx

c) Una vez llega el archivo a la carpeta raw se debe disparar un lambda que procese los

datos que llegaron. Este proceso debe extraer la categoría, el titular y el enlace para cada

noticia. Utilizar la biblioteca Beautifulsoup. Estos datos se deben guardar en un csv en la siguiente ruta:

• s3://bucket/news/final/periodico=xxx/year=xxx/month=xxx/day=xxx

d) Se debe crear la respectiva base de datos sobre la carpeta final usando Hive en EMR

e) Crear las respectivas particiones

f) Verificar que se puedan realizar consultas sobre Presto

Nota:

\* Hay que crear un proyecto de zappa para cada función.

\*\* Se debe entregar el código en github(utilizar ramas, commits, código limpio, código comentado,

pruebas unitarias si aplica, scripts de creación para Hive).