PRACTICA BIG DATA ARCHITECTURE

1. Introducción.

El objetivo de esta práctica es dar recomendaciones al cliente de KcTelecom en el método de pago. Para ello será necesario hacer las siguientes tablas de indicadores:

|  |
| --- |
| KPIS\_CLIENTES  Id\_cliente (int)  Nota (f(N\_facturas), int)  Influencer (f(Twitter), bool)  Id\_profesion (int)  Id\_ciudad\_estado (int) |

|  |
| --- |
| INDICADOR\_ACEPTO\_CRIPTOS  Aceptamos (f(Marketcap), bool) |

A partir de estos indicadores se le dará al cliente la posibilidad de pagar cuando convenga las facturas con marketing de afiliados (cuando a partir del análisis se considere que el cliente puede atraer a muchos clientes) o criptomonedas (siempre y cuando el mercado se encuentre bajo, para ganar cuando suba).

Se establecen además indicadores de calificación, de zona de residencia y tipo de profesión para tener mayor conocimiento del cliente en otros análisis.

1. Diseño.

Para ello necesitaré traerme a HIVE la tabla CLIENTE y FACTURACION del datawarehouse de KcTelecom y datos web (de RRSS y de mercados criptomonedas). A nivel lógico, los pasos que se seguirían para crear los dos tablones de indicadores serían los siguientes:

|  |
| --- |
| CLIENTES\_TRANS  Id\_cliente (int)  Id\_ciudad\_estado (int)  Id\_profesión (int) |

|  |
| --- |
| CLIENTES\_FACTURAS  Id\_cliente (int)  Nota ( f(N\_facturas), int)  Id\_profesión (int)  Id\_ciudad\_estado (int) |

|  |
| --- |
| CLIENTES |

|  |
| --- |
| KPIS\_CLIENTES  Id\_cliente (int)  Nota (int)  Influencer (bool)  Id\_profesion (int)  Id\_ciudad\_estado (int) |

|  |
| --- |
| FACTURACION |

|  |
| --- |
| FACTURAS\_TRANS  Id\_cliente (int)  N\_facturas (int) |

|  |
| --- |
| INFLUENCER  Id\_cliente (int)  Influencia (f(Twitter), bool) |

|  |
| --- |
| Twitter (web) |

|  |
| --- |
| INDICADOR\_ACEPTO\_CRIPTOS  Aceptamos (f(Marketcap), bool) |

|  |
| --- |
| Coinmarketcap (Web) |

Todos los datos estructurados necesarios en la practica irán directos a un área de RAW DATA en HDFS. En este caso se guardarán en formato texto. No obstante, si fuera un entorno real las guardaría en formato ORC y las particionaría por el campo id\_cliente para ahorrar espacio y tiempo.

Después me llevo estos datos a tablas gestionadas por HIVE en el área de STAGING, para luego, hacer las transformaciones correspondientes. Aquí conservaré el formato texto y de igual forma conservaría las particiones y ORC en un caso con más volumen de datos.

Carga inicial: Las tablas estructuradas (CLIENTES y FACTURACION) se traerán de MySql Workbench a RAW DATA en HDFS haciendo una copia mediante Sqoop en forma “snapshot”.

Por otro lado, los datos web de Twitter se traerán a la zona de CREACIÓN\_KPIS en HIVE mediante un cliente gestionado por IMPYLA y consumido por un notebook de Jupyter. Este notebook contendrá el script (escrito en Python) que automatizaría el WebScrapping del Twitter de cada cliente dejando los datos en arrays. De la misma manera se hará para recoger la tabla indicadora de aceptación de criptomonedas.

De esta manera el esquema general quedaría así:

HADOOP

COINMARKETCAP

WEBHCAT/IMPYLA

TWITTER

GENERACION KPIS

KPIS

WEB

HDFS

BBDD

STAGING

RAW DATA

CLIENTES

SQOOP

HIVE

FACTURAS

Estrategia periódica: actualizaría todos los datos cada tres semanas exceptuando el indicador de criptomonedas que lo haría cada semana y media.

Gobierno del dato: para controlar que las tablas FACTURAS y CLIENTE no las cargue mas de una vez pondré de apellido la fecha en el que se cargan al nombre de las tablas. Para evitar duplicidades en las tablas que provengan de la web seguiré el mismo procedimiento. Para automatizar todos los procesos haría uso de herramientas ETL que ejecutaran los scripts correspondientes.

1. Construcción.

Para conseguir la infraestructura de la practica he configurado una Sandbox de GCloud con Hortonworks obteniendo un mono clúster Hadoop.

En la construcción del diseño planeado he llegado hasta la creación de la tabla CLIENTES\_FACTURAS, pues como los usuarios de la tabla clientes no son reales no tienen cuentas de Twitter. Además, después de mucho investigar e intentar tampoco consigo programar un notebook que automatice la búsqueda de nombres en Twitter.

En cuanto al indicador de aceptar pagos en criptomonedas también se me complica el código.

Las transformaciones que he llevado a cabo son:

Nota = f(N\_facturas) = 1(N\_facturas >= 50), 2(50 > N\_facturas >= 35) o 3(N\_facturas < 35).