Brainstorm

* Scrappear Twitter buscando comentarios sobre plataformas de venta de ropa online.
* Análisis del sentimiento de los comentarios.
* Identificar los puntos fuertes y las debilidades de otras plataformas.
* Visualizar los resultados del análisis (Kibana).
* Desarrollar una nueva plataforma propia que satisfaga las necesidades identificadas.
* Una vez creada la plataforma se podría repetir el análisis del sentimiento sobre comentarios de nuestra propia plataforma para continuar mejorando el servicio.

Diseño del DAaaS

### Definición la estrategia del DAaaS

*Definir el catálogo de servicios que proporcionará la plataforma DAaaS, que incluye incorporación de datos, limpieza de datos, transformación de datos, datapedias, bibliotecas de herramientas analíticas y otros.*

* Dashboard de visualización de los comentarios más frecuentes sobre las plataformas, con su correspondiente análisis del sentimiento, que nos permita identificar de forma sencilla qué aspectos valoran más los consumidores de las plataformas de venta de ropa online.

### Arquitectura DAaaS

*Definir la selección de componentes, la definición de procesos de ingeniería y el diseño de interfaces de usuario. Diseño y ejecución de Proofs-of-Concept (PoC) para demostrar la viabilidad del enfoque DAaaS.*

* **Google Cloud (Compute Engine):**

Las diversas máquinas requeridas a lo largo de proceso se levantarán en Google Cloud.

* **Extractor de datos de Twitter**:

Nos conectaremos a la API de Twitter, para lo que se requiere autorización por medio de la concesión de un permiso especial en la cuenta de Twitter (Developer Account), a fin de extraer aquellos tweets en que se mencionen plataformas de venta de ropa online (Zalando, Vinted, Asos…).

Dicha extracción se realizará en una máquina de Ubuntu, con Python instalado, y gestionada por medio de Google Cloud (Compute Engine). Dicha maquina permanecerá levantada y ejecutándose por un tiempo preestablecido, suficiente para extraer conclusiones que nos permitan identificar en qué aspectos hemos de centrarnos para desarrollar una plataforma que responda a las necesidades de los consumidores.

* **Google Cloud (Cloud Storage):**

Los datos extraídos se almacenarán en un Bucket específico de Google Cloud para después proceder al staging de datos.

* **Hadoop:**

Los datos se recogerán en un cluster con Hadoop instalado (Google Dataproc) para el procesamiento de los datos sirviéndonos de HDFS (Hadoop Distributed File System) y MapReduce a fin de optimizar los procesos. Aquí se llevará a cabo el tratamiento de datos, así como un análisis del sentimiento de los tweets que se recogerán y almacenarán en el propio Hadoop.

Dicho cluster contará con un nodo master y sus correspondientes workers, cuyo número será determinado en función del volumen de datos a procesar (auto-scaling) levantando y tumbando clusters de forma elástica.

* **Elasticsearch + Kibana:**

Conectar Hadoop a una máquina que tenga elastic instalado y levantado (Compute Engine) para enviar los resultados a un índice de elasticsearch que permita crear un dashboard en Kibana para visualizar los resultados del análisis, opiniones más frecuentes con su respectivo análisis del sentimiento.

* **Cloud Scheduler + Cloud Functions:** Automatización de procesos cloud.

### DAaaS Operating Model Design and Rollout

*Personalizar los modelos operativos DAaaS para cumplir con los procesos, la estructura organizacional, las reglas y el gobierno de los clientes individuales. Realizar seguimiento de consumo y mecanismos de informe.*

1. Solicitar la cuenta con derechos de “Developper Account” de Twitter para tener acceso a su API.
2. Arrancar una máquina en Google Compute Engine con Ubuntu y Python instalado a través de la que nos conectaremos a la API de Twitter. En ella extraeremos tweets en base a palabras clave que se mencionen. Esta máquina permanecerá activa hasta que se haya extraído una muestra significativa de tweets. También podrá levantarse periódicamente por medio de Google Cloud Scheduler + Cloud Functions.
3. Volcar los tweets extraídos, en forma de documentos json, a un bucket de Google Storage.
4. El cluster de hadoop permanecerá levantado mientras se continúen extrayendo tweets, es decir, mientras el extractor de tweets (2) esté activo.
5. Lanzar un job de hadoop que ingiere los datos del bucket y lleva a cabo las transformaciones necesarias de los datos para ejecutar un análisis del sentimiento, generando archivos con los resultados que se almacenarán en el propio Hadoop.
6. Conectar Hadoop con elasticsearch, volcando los resultados del análisis en un índice que permita construir un dashboard en Kibana.
7. Conceder accesos a las visualizaciones de Kibana.

Una vez llevado a cabo el análisis, desarrollar una plataforma propia que satisfaga las necesidades identificadas y repetir el proceso, sobre nuestra propia plataforma, para continuar mejorando el servicio ofrecido y valorar cómo ha sido acogido por nuestros consumidores.

### Desarrollo de la plataforma DAaaS.

*Construcción iterativa de todas las capacidades de la plataforma, incluido el diseño, desarrollo e integración,* ***pruebas****, carga de datos, metadatos y población de catálogos, y despliegue.*

* Solicitar la “Developer Account” de Twitter que nos permita conectarnos a su API: <https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-the-twitter-api>

Link a Diagrama:

<https://docs.google.com/drawings/d/1DHwJt28S3WGa7yuG3pKHAADJN72mhk0_OGuFGJ7hR7o/edit?usp=sharing>

Diagrama

Descripción generada automáticamente