

Trabajo fin de grado

Plataforma web para el análisis de mensajería instantánea



Juan Bautista Menchero Amigo

Escuela Politécnica Superior
Universidad Autónoma de Madrid
C/ Francisco Tomás y Valiente nº 11

UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR



Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

**Plataforma web para el análisis de mensajería
instantánea**

Autor: Juan Bautista Menchero Amigo
Tutor: Esther Guerra Sánchez

junio 2022

Todos los derechos reservados.

Queda prohibida, salvo excepción prevista en la Ley, cualquier forma de reproducción, distribución comunicación pública y transformación de esta obra sin contar con la autorización de los titulares de la propiedad intelectual.

La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (*arts. 270 y sgts. del Código Penal*).

DERECHOS RESERVADOS

© 24 de Mayo de 2022 por UNIVERSIDAD AUTÓNOMA DE MADRID

Francisco Tomás y Valiente, n.º 1

Madrid, 28049

Spain

Juan Bautista Menchero Amigo

Plataforma web para el análisis de mensajería instantánea

Juan Bautista Menchero Amigo

C\ Francisco Tomás y Valiente N.º 11

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

A mi madre.

Este título es tan tuyo como mío.

*Lo peor es cuando has terminado un capítulo
y la máquina de escribir no aplaude.*

Orson Welles

PREFACIO

Este estilo de $\text{\LaTeX} 2_{\epsilon}$ ha sido diseñado con dos propósitos. El primer propósito es el de facilitar en lo posible la escritura de trabajos de fin de grado y de máster y de tesis doctorales. En ese sentido se han diseñado un conjunto de comandos que simplifican la escritura y diseño de estos trabajos pero que reducen en cierta forma las capacidades de los paquetes de \LaTeX utilizados. Sin embargo, dado que los paquetes están incluidos en esta clase, pueden utilizarse directamente y hacer diseños más complejos pero si se hace esto se recomienda mantener una estética coherente con el resto del documento.

El segundo de los propósitos es que estos documentos mantengan una estética uniforme en la Universidad Autónoma de Madrid y fomentar una imagen corporativa en documentos tan relevantes como los trabajos de fin de grado o de máster y las tesis doctorales. Por ese motivo se recomienda mantener una coherencia estética en todo momento. El diseño facilita esa coherencia pero es posible salirse del diseño si se mantiene dicha coherencia.

Como creador de este estilo espero fervientemente que al usar este estilo te sientas cómodo y te facilite la escritura de un documento que es muy relevante en esta etapa de tu vida. Para facilitártela aún más, el código fuente de este documento también está disponible en tu ordenador o en overleaf para que te sirva a modo de ejemplo.

Eloy Anguiano Rey

AGRADECIMIENTOS

Principalmente quiero agradecer a mi madre, Maria Jesus Amigo Rodriguez, la unica persona que se ha esforzado en que yo llegara hasta aqui mas que yo mismo. Te lo debo todo, gracias por transmitirme los valores del amor, la bondad, el trabajo, la fuerza, la flexibilidad,... Eres la persona que mas admiro.

A mi padre, Juan Bautista Menchero Garcia, por haberme inculcado la importancia del conocimiento y la constancia. Por despertar mi interes por todas las cosas, como fuente inagotable de felicidad.

A mi hermano, Hugo Jimenez Amigo, por mostrarme la vida que queria seguir y no para la que estaba predestinado, y haberme allanado y limpiado el camino que era desconocido para ambos, por animarme y guiarme a como corregir mis errores y ser la mejor version de mi que puedo ser.

A Aitana Rickert Llacer, por haberme acompa ado en cada linea de codigo y cada pagina de la carrera con tu luz infinita. Por haberme cuidado y dado todo lo que necesitaba cuando yo no era capaz de darlo de vuelta, pero sobre todo por haberme hecho mejor persona, aunque no estes para verlo.

A Alvaro Martinez Morales (yo si que se escribir tus apellidos), por ser lo mejor que me ha dado esta etapa, no podia sonar nada que supere haberme llevado alguien mas en mi familia.

RESUMEN

En este trabajo se ha diseñado e implementado tanto el software como la infraestructura necesarias para ofrecer una aplicación web accesible por el público para la generación de reportes sobre su uso de aplicaciones de mensajería instantánea. El objetivo es contribuir a la democratización de los datos, así como concienciar a los usuarios de la información que se puede extraer de ellos.

PALABRAS CLAVE

Computación en la nube, análisis de datos, procesamiento de datos, AWS, Amazon Web Services, Terraform, Redshift, Vue

ABSTRACT

En este trabajo se ha diseñado e implementado tanto el software como la infraestructura necesarias para ofrecer una aplicación web accesible por el público para la generación de reportes sobre su uso de aplicaciones de mensajería instantánea. El objetivo es contribuir a la democratización de los datos, así como concienciar a los usuarios de la información que se puede extraer de ellos.

KEYWORDS

Computación en la nube, análisis de datos, procesamiento de datos, AWS, Amazon Web Services, Terraform, Redshift, Vue

ÍNDICE

1	Introduccion	1
1.1	Motivacion	1
1.2	Objetivos	1
1.3	Organizacion de la memoria	1
2	Estado del arte	3
2.1	Derecho de acceso	3
2.2	Informacion destilada de libre acceso	3
2.3	Computacion en la nube	4
2.4	Frameworks frontend	4
2.5	Bases de datos	4
3	Analisis de requisitos	5
3.1	Reportes individuales	5
3.2	Reportes comparativos	5
4	Disenyo	7
4.1	Experiencia de usuario	7
4.2	Infraestructura	7
4.3	Tecnologias	7
4.4	Planificacion	8
4.5	Implementacion	8
5	Pruebas	9
6	Conclusiones y trabajo futuro	11
6.1	Conclusiones	11
6.2	Trabajo futuro	11
	Bibliografía	11

LISTAS

Lista de algoritmos

Lista de códigos

Lista de cuadros

Lista de ecuaciones

Lista de figuras

Lista de tablas

Lista de cuadros

INTRODUCCION

1.1. Motivacion

Actualmente generamos una cantidad inmensa de información sobre nosotros mismos todos los días a la que solo tienen acceso las plataformas que nos prestan los servicios que consumimos, sin ni siquiera ser conscientes de las conclusiones que son capaces de sacar sobre nuestros datos.

1.2. Objetivos

La intencion es concienciar (o por lo menos entretener) a la gente con la información que se puede obtener de ellos, y tener la oportunidad de extraer conclusiones creativas a partir de lenguaje natural y metadatos, así como explorar tecnologías de cloud provistas por AWS para la ingesta y procesado de los datos.

1.3. Organizacion de la memoria

En este trabajo presento una aplicación web donde los usuarios pueden analizar sus copias de seguridad de aplicaciones de mensajería instantánea (como Telegram o What's App) y visualizar varios reportes con información que se puede destilar en base a esos históricos.

ESTADO DEL ARTE

2.1. Derecho de acceso

El 14 de abril de 2016 se aprobo en el Parlamento Europeo el Reglamento General de Proteccion de Datos, entrando en vigor el 24 de Mayo de 2016 y concediendo un periodo de aplicacion de dos anyos hasta el 24 de Mayo. A partir del 25 de Mayo de 2018 todas las empresas, organizaciones, organismos o instituciones dentro del marco europeo comenzaron a tener la obligacion, bajo multa por incumplimiento de hasta 20 millones de euros, de ofrecer a los usuarios de una manera facilmente accesible y legible una copia de sus datos almacenados.

Gracias a esta ley podemos solicitar ante cualquier plataforma que almacene o trate nuestros datos, una copia de seguridad de toda nuestra informacion que tienen disponible, desde Google, Facebook, Twitter, Spotify, Tinder, o, en el caso de la informacion a analizar en este trabajo, a aplicaciones de mensajeria instantanea como What's App o Telegram. Permittiendonos tener acceso a una fuente de datos de manera sencilla y facilmente ingerible por el sistema propuesto.

2.2. Informacion destilada de libre acceso

Spotify Wrapped:

Desde 2016 Spotify lanza anualmente una campana de marketing que permite a sus usuarios visualizar una compilacion de sus datos de uso, comprandolos con el resto de la comunidad que utiliza la aplicacion. Desde resúmenes meramente estadísticos, así como análisis de emociones o intereses.

Esta camapana ha tenido un gran impacto social y cultural, llegando a formar parte de la cultura pop de las nuevas generaciones, y generando de una forma inconsciente inquietud e interes por los datos a la poblacion general, regularmente ajena y desinteresada de sus datos en la era digital.

Tinder Insights:

Creada en 2019 por Dora Szucs (Software Engineer) y Krisztina Szucs (UX Designer), da acceso

a informacion estadistica sobre el uso de la aplicacion Tinder, como la cantidad de mensajes enviados y recibidos, tiempo medio de chat, ...

Ejemplo practico de que el interes por la informacion sobre uno mismo puede ser suficiente para que un gran numero de poblacion comparta sus datos de uso con una aplicacion de terceros para analizarla.

<https://tinderinsights.com/contact> <https://who.is/whois/tinderinsights.com>

Chat Visualizer:

Una herramienta que permite analizar la actividad de un chat de What's App de manera estadistica (72K chats analizados en el momento de la redaccion de esta memoria).

Contras que este trabajo pretende subsanar:

- Solo permite analizar una conversacion
- Solo analiza informacion de What's App
- Procesa los datos en el lado del servidor, no garantizando la privacidad de esa informacion
- Solo ofrece un analisis estadistico de actividad

2.3. Computacion en la nube

2.4. Frameworks frontend

2.5. Bases de datos

Relacionales vs No Relacionales

Relacionales: Almacenan la informacion de manera estructurada, con una forma constante y una declaracion explicita de las relaciones entre los diversos elementos (MySQL, Oracle, SQL Server, PostgreSQL, ...) No Relacionales: Almacenan la informacion como documentos independientes, que no tienen por que mantener la misma forma ni tener relaciones explicitas (MongoDB, Redis, Elasticsearch, Cassandra, ...)

Columnares vs Orientadas a filas

Orientadas a filas: Almacenan en memoria contigua la informacion de cada entidad, permitiendo un acceso rapido a todos los datos de una misma entidad (SQLServer, PostgreSQL, ...) Columnares: Almacenan de manera contigua cada columna o atributo, para acelerar el analisis estadistico de toda una tabla (Redshift, BigQuery, ...)

ANALISIS DE REQUISITOS

3.1. Reportes individuales

- Horas a las que sueles responder - Contactos favoritos a los que respondes mas rapido - Tendencias de tu estado de humor segun los emoticonos que usas mas frecuentemente - Personas que han aparecido o desaparecido de tu vida - Tus temas e intereses favoritos - Horas de sueno o trabajo. - Personas con las que estas en mas grupos

- 1.- Ofrecer una interfaz grafica intuitiva para permitir el acceso a cualquier tipo de usuario
- 2.- Tiempos de respuesta rapidos para evitar que el tiempo de procesamiento sea un impedimento a la hora de utilizar la aplicacion
- 3.- Mostrar reportes interesantes para que los usuarios tengan una motivacion a la hora de utilizar la plataforma
- 4.- Aportar informacion que conciencie del compromiso a la intimidad que supone otorgar acceso a los historicos de chat
- 5.- Utilizar infograficos llamativos para favorecer que se compartan por redes sociales y publicitar la herramienta
- 6.- Evitar que la informacion del usuario salga de su ordenador para garantizar la privacidad total de los datos

3.2. Reportes comparativos

- Top emojis - Tiempo de uso - Longitud del mensaje medio

- 1.- Almacenar solo los minimos datos necesarios para los reportes comparativos, sin informacion privada y completamente anonimizados

DISENYO

4.1. Experiencia de usuario

UX:

1. Upload screen 2. Loading 3. First report 4. Second report 5. Link to share and keep (targeting first report link)

4.2. Infraestructura

Architecture:

- AWS account and console - IAM: Create access for GithubActions and Terraform - S3: Manually create state bucket - GitHub: Add AWS secrets to GitHub

UNDERSTANDING TERRAFORM REMOTE STATE

You might be wondering what's going on with our remote state? What even is remote state and why do we need it? So let's answer that question now: state is what Terraform uses to compare the current state (note the wording here) of your infrastructure against the desired state. You can either create this state locally (i.e Terraform writes to a file) or you can do it remotely.

We need to create our state remotely if we are to run it on Github Actions. Without remote state, Terraform generates a local file, but it wouldn't commit it to GitHub, so we'd lose the state data and end up in a sticky situation. With remote state we avoid this problem by keeping state out of our pipeline in separate persistent storage.

4.3. Tecnologias

- Redshift, columnar por ser analitica, poder usar data api y no tener que montar otra capa. posibilidad de escalado.
- DC2 porque es gratis <https://aws.amazon.com/es/redshift/pricing/>

- Acceder a redshift primero desde cli y luego desde js sdk <https://docs.aws.amazon.com/cli/latest/userguide/getting-started-install.html>

Working with the Amazon Redshift Data API

Before you use the Amazon Redshift Data API, review the following steps:

- 1.— Determine if you, as the caller of the Data API, are authorized.
- 2.— Determine if you plan to call the Data API with authentication credentials from Secrets Manager or temporary credentials.
- 3.— Set up a secret if you use Secrets Manager for authentication credentials.
- 4.— Review the considerations and limitations when calling the Data API.
- 5.— Call the Data API from the AWS Command Line Interface (AWS CLI), from your own code, or using the query editor in the Amazon Redshift console.

4.4. Planificacion

Sprints

- 1.— Frontend scaffolding with CD/CI and Infrastructure as a code with AWS
- 2.— Upload steps and easy reports entirely in frontend
- 3.— Save RAW data in S3 Warehouse
- 4.— Save reports in DB

4.5. Implementacion

PRUEBAS

CONCLUSIONES Y TRABAJO FUTURO

6.1. Conclusiones

6.2. Trabajo futuro

- 1.– Mejorar politicas de acceso IAM
- 2.– Migrar claves de acceso a GitHub secrets
- 3.– Contratar a un UX Designer para replantear el aspecto grafico de la plataforma
- 4.– Comprar un dominio y liberar el acceso a la aplicacion



Universidad Autónoma
de Madrid