ESCUELA POLITÉCNICA DE INGENIERÍA DE GIJÓN

GRADO EN INGENIERÍA INFORMÁTICA EN TECNOLOGÍAS DE LA INFORMACIÓN

Lenguajes y Sistemas Informáticos

TRABAJO FIN DE GRADO/MÁSTER Nº ???

Explotación, integración y visualización de múltiples fuentes de datos mediante un Data Lake

Mier Montoto, Juan Francisco

TUTORES:
D. Augusto Alonso, Cristian
D. Morán Barbón, Jesús
D. Vázquez Faes, Eduardo

FECHA: junio 2024

Índice de contenido

Índice de contenido 1

Índice de figuras 2

1. Introducción 3
1.1. Antecedentes . . . 3
1.2. Motivación . . . 3
1.3. Finalidad del proyecto . . . 4
1.4. La empresa . . . 4

2. Fundamento teórico 6
2.1. Definición de Data Lake . . . 6
2.2. Procesos ETL . . . 6
2.3. Dashboards . . . 6
2.4. Componentes del stack . . . 6
2.4.1. Productor . . . 7
2.4.2. Consumidor . . . 7
2.4.3. Tópico . . . 7
2.4.4. Broker . . . 7
2.4.5. Zookeeper . . . 7

3. Planificación 8
3.1. Planificación inicial . . . 8
3.2. Presupuesto inicial . . . 8

4. Diseño del sistema 9

5. Implementación del sistema 10

6. Resultados 11

7. Conclusiones y trabajo futuro 12

Bibliografía 13

Índice de figuras

1. Introducción

1.1. Antecedentes

En la actualidad, la cantidad de datos que se generan y almacenan en las empresas es cada vez mayor. Estos datos provienen de múltiples fuentes y en múltiples formatos, lo que dificulta su análisis y explotación.

A su vez, el progreso tecnológico ha permitido la creación de nuevas herramientas y técnicas que facilitan la recogida, almacenamiento y análisis de estos datos. Una de estas técnicas son los *data lakes*, que permiten almacenar grandes cantidades de datos de diferentes tipos y formatos, para poder analizarlos y explotarlos de forma más eficiente.

De manera paralela, el progreso en la *ciencia de datos* ha supuesto una mejora en la visualización de estos, lo que permite un mejor análisis. La visualización de datos suele significar la existencia de uno o varios *dashboards*, que permiten al usuario final la consulta rápida y sencilla de los datos sin necesidad de conocimientos técnicos del sistema.

1.2. Motivación

El proyecto surge de la necesidad de *1.4 La empresa* de recoger y analizar datos de múltiples características de todas las fuentes de las que se disponen, tanto internas (como bases de datos, logs, la propia API, etc.) como externas (como APIs de otras empresas, datos web de terceros, etc.).

En la actualidad, la empresa dispone de una gran cantidad de datos que se encuentran en diferentes formatos y en diferentes ubicaciones, lo que dificulta su análisis y explotación. Por otra parte, se depende de la consulta manual o de servicios de terceros (como dashboards en NewRelic o AWS CloudWatch) para poder analizar estos datos, lo que supone un coste adicional.

Además del uso interno, la empresa también quiere ofrecer a sus clientes la posibilidad de consultar estos datos de forma visual y sencilla, para que puedan analizarlos y explotarlos de forma autónoma, lo que supondría un valor añadido para los mismos.

1.3. Finalidad del proyecto

El objetivo de este sistema es centralizar y unificar todas estas fuentes que se encuentran dispersas, en diferentes formatos y que en la actualidad se tienen que consultar manualmente, para poder realizar análisis de los datos de forma más eficiente.

Como se comenta anteriormente, el cumplimiento de este objetivo permitirá:

- una eliminación del tiempo invertido en la consulta manual de los datos.
- una reducción de los costes de las plataformas de terceros.
- una mejora en la toma de decisiones, al poder analizar los datos de forma más eficiente.
- una mejora en la calidad de los servicios ofrecidos a los clientes, al poder ofrecerles la posibilidad de consultar los datos de forma visual y sencilla.
- el beneficio económico que supondría la venta de este servicio a los clientes.

Además de la mejora de los procesos ya existentes, la explotación de esta herramienta abrirá la puerta a nuevas posibilidades de análisis y explotación de los datos, como la detección de anomalías en la infraestructura o la predicción de patrones y eventos futuros.

1.4. La empresa

Okticket es una startup nacida en Gijón en 2017 cuyo producto principal es un servicio software que reduce los costes y el tiempo que invierten las empresas en contabilizar y manejar los gastos de viaje de los profesionales mediante el escaneo automático de tickets y notas de gastos.

Sus oficinas principales (incluyendo la zona de desarrollo) se encuentran en el Parque Tecnológico de Gijón, aunque cuenta con un número de sedes creciente en varios países: Francia, Portugal y, más recientemente, México.

Okticket es una de las empresas que más crecen tanto del sector como del propio Parque Tecnológico. Debido a este rápido crecimiento, el equipo está en constante desarrollo y cambio, tanto aquí en España como en el resto de sedes.

Pese a que la parte principal del negocio es el SaaS (Software as a Service en inglés), es

decir, la aplicación completa tanto para administradores como para empleados, recientemente se han añadido nuevas propuestas como la OKTCard, una tarjeta inteligente que gestiona automáticamente los gastos, entre otros proyectos.

2. Fundamento teórico

2.1. Definición de Data Lake

2.2. Procesos ETL

Los procesos ETL [1] son procesos que combinan datos de múltiples fuentes en un único destino, transformando los datos en un formato común. Estos procesos se utilizan para extraer datos de diferentes fuentes, transformarlos en un formato común y cargarlos en un dashboard para que se muestren de manera visual.

Estos procesos deben estar "personalizados" para cada caso, ya que cada sistema procesa diferentes tipos de datos, de diferentes fuentes y en diferentes formatos. Además, estos procesos deben ser escalables, ya que los datos que se muestran en los dashboards suelen ser datos que se generan de manera continua, y por lo tanto los procesos ETL deben ser capaces de procesar grandes cantidades de datos de manera eficiente.

2.3. Dashboards

La palabra *dashboard*, que traducido de manera literal significa *cuadro de mandos*, es un término que se utiliza para referirse a cualquier interfaz gráfica que muestre información relevante de manera visual sobre un proceso o negocio. Aunque el término se utiliza en muchos ámbitos (puede incluir indicadores comerciales, de producción, de marketing, de calidad, de recursos humanos...)

En el ámbito de deste proyecto, los dashboards reflejan en tiempo real el rendimiento de actividades o procesos de negocio, y se utilizan para tomar decisiones informadas sobre los mismos. En el caso de una empresa, un dashboard puede mostrar desde el rendimiento de la plataforma en tiempo real hasta un reflejo del de las ventas, y permitir a los directivos tomar decisiones informadas sobre el futuro de la empresa.

2.4. Componentes del stack

En el *stack* tecnológico escogido para el proyecto se manejan diferentes términos y conceptos que son necesarios desarrollar para entender el funcionamiento del sistema.

2.4.1. Productor

2.4.2. Consumidor

2.4.3. Tópico

2.4.4. Broker

2.4.5. Zookeeper

3. Planificación

3.1. Planificación inicial

3.2. Presupuesto inicial

4. Diseño del sistema

5. Implementación del sistema

6. Resultados

7. Conclusiones y trabajo futuro

Bibliografía

[1] J. Mier, "Presentación de datos: dashboards y procesos ETL." Primera entrega de teoría de la asignatura Inteligencia de Negocio, EPI Gijón, curso 23-24, 2023.

[2] J. Mier, "latexTemplate." `https://github.com/miermontoto/latexTemplate`, 2024. Plantilla de LaTeX personal para trabajos académicos.

[3] J. Mier, "Minería de anomalías." Segunda entrega de teoría de la asignatura Inteligencia de Negocio, EPI Gijón, curso 23-24, 2024.