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 figuras

1. Introducción

1.1. Antecedentes

1.1.1. Big data y evolución

En la actualidad, la cantidad de datos que se generan y almacenan es cada vez mayor [1], una tendencia que por supuesto se traduce a las empresas. Estos datos provienen de múltiples fuentes y en múltiples formatos, lo que dificulta su análisis y explotación. A esta característica de la información se le conoce como *heterogeneidad*[2].

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* (ver *2.1 Data lake*), que permiten almacenar grandes cantidades de datos de diferentes tipos y formatos, para poder analizarlos y explotarlos de forma más eficiente.

1.1.2. Visualización de datos y DIKW

La visualización de datos es una técnica que permite representar la información de manera visual, para facilitar su análisis y comprensión. La visualización de datos es una parte importante del proceso de análisis de datos, ya que permite identificar patrones, tendencias y anomalías en los datos de forma más rápida y sencilla.

La evolución de la visualización de datos ha ido de la mano de la evolución de la tecnología, y actualmente existen múltiples herramientas y técnicas que permiten visualizar datos de forma más eficiente y efectiva. Una de estas técnicas es el modelo DIKW (ver *2.2 Modelo DIKW*), que describe el proceso de transformación de los datos en información, la información en conocimiento y el conocimiento en sabiduría.

[1] `https://www.statista.com/statistics/871513/worldwide-data-created/`
[2] `https://www.sciencedirect.com/topics/computer-science/data-heterogeneity`

1.2. Motivación

El proyecto surge de la necesidad de la empresa (ver *1.4 La empresa*) de recoger y analizar datos heterogéneos de todas las fuentes de las que se disponen, tanto internas (e.g. bases de datos, archivos de registros, APIs, entre otros), como externas (e.g. APIs o datos de webs de terceros, datos de fuentes públicas...).

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. Este tipo de dashboards son diferentes a los dashboards de monitorización antes mencionados, ya que permiten al usuario final la consulta de datos de negocio, y no de infraestructura.

1.3. Finalidad del proyecto

El objetivo de este sistema es centralizar y unificar las fuentes de datos heterogéneas cuya consulta se realiza de manera manual, con la finalidad de analizar los datos de forma más eficiente.

El cumplimiento de este objetivo permitirá a la empresa obtener una serie de beneficios:

- 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 mediante 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.

La empresa tienen su suede principal en el Parque Tecnológico de Gijón, aunque cuenta con un número de sedes creciente en varios países, como Francia, Portugal o, más recientemente, México. En esta oficina principal se encuentran los departamentos de ventas y marketing, así como el equipo de desarrollo y soporte.

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. Este crecimiento se refleja en la recepción de un gran número de galardones y reconocimientos [3] [4] [5] [6]

La parte principal del negocio es el núcleo del software como servicio (Software as a Service en inglés, en adelante *SaaS*), es decir, la aplicación completa tanto para administradores como para empleados. Este SaaS se oferta a empresas de cualquier tamaño, cuyo precio final varía en función del número de usuarios, las características e integraciones que requiera la empresa cliente y el soporte que se ofrezca.

Recientemente se han añadido nuevas propuestas a la cartera de servicios ofertada por Okticket, como la OKTCard - una tarjeta inteligente que gestiona automáticamente los gastos, así como la inclusión de nuevos "módulos" de gestión de gastos y viajes.

[3] Okticket en el especial startups 2023 de Forbes (LinkedIn)
[4] Arcelor y Okticket, premios nacional de Ingeniería Informática (EL COMERCIO)
[5] Okticket recibe el sello Pyme Innovadora (okticket.es)
[6] Okticket, empresa emergente certificada (okticket.es)

2. Fundamento teórico

2.1. Data lake

Los *data lakes*[1] son almacenes de datos que almacenan grandes cantidades de datos de manera no estructurada [1].

A diferencia de los *data warehouses*, los *data lakes* no tienen un esquema definido, lo que permite almacenar datos *heterogéneos*. Esto permite almacenar grandes cantidades de información sin tener que definir un esquema de antemano, lo que puede ser útil en aquellos casos en los que no se conoce la estructura de los datos que se van a almacenar.

2.2. Modelo DIKW

La pirámide DIKW[2] es un modelo que describe la relación entre los datos, la información, el conocimiento y la sabiduría. Según este modelo, los datos son la materia prima de la información, que a su vez es la materia prima del conocimiento, que a su vez es la materia prima de la sabiduría.

La visualización de datos es una técnica que permite representar los datos de forma visual, para facilitar su análisis y explotación. Los dashboards (ver *2.4 Dashboards*) son una herramienta que permite visualizar los datos de forma sencilla y eficiente, para poder tomar decisiones informadas sobre los mismos.

2.3. 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 destino común.

Los procesos ETL deben de adaptarse a la estructura de los datos de la fuente de origen, ya que dichas fuentes pueden tener diferentes estructuras y tener tipos de datos diferentes (la

[1] `https://aws.amazon.com/es/what-is/data-lake/`
[2] `https://en.wikipedia.org/wiki/DIKW_pyramid`

heterogeneidad que ya se ha mencionado).

Una de las características clave de los procesos ETL es que sean 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. En ocasiones, los procesos ETL se pueden realizar en *streaming*, lo que significa que los datos se procesan en tiempo real a medida que se generan.

2.4. 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.

Para el sistema que se describe, se plantean dos tipos de dashboards diferentes:

- **Dashboards internos**: que reflejan el rendimiento de la plataforma en tiempo real.
- **Dashboards externos**: que reflejan el rendimiento de las ventas y permiten a los clientes tomar decisiones informadas sobre su negocio.

2.5. Componentes del sistema

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

2.5.1. Tópico

Un tópico es una categoría a la que se envían los mensajes a la que los consumidores están *suscritos*. Los consumidores pueden estar suscritos a uno o varios tópicos, y los productores pueden enviar mensajes a uno o varios tópicos. Los tópicos son la unidad básica de organización de los mensajes en cualquier sistema de mensajería de publicación/suscripción.

2.5.2. Productor

El productor es el componente responsable de crear y enviar mensajes al cluster de Kafka. Está separado del resto de los componentes y produce mensajes de manera asíncrona y rápida.

2.5.3. Consumidor

El consumidor es el componente responsable de leer los mensajes producidos por el productor. Está suscrito a un Tópico a través del broker y consume los mensajes.

2.5.4. Broker

El broker es el componente responsable de recibir los mensajes producidos por el productor y enviarlos a los consumidores. Es el intermediario entre los productores y los consumidores.

2.5.5. Zookeeper

Zookeeper es un servicio de coordinación distribuida que se utiliza para gestionar y coordinar los brokers de Kafka. Se encarga de mantener la información de los brokers y de los tópicos.

3. Planificación

La planificación de un proyecto es fundamental para su correcto funcionamiento y desarrollo, dentro de los plazos y costes establecidos. En este apartado se detallan las tareas que se llevarán a cabo en el proyecto, así como los recursos necesarios y los plazos de ejecución.

3.1. Planificación inicial

3.1.1. Partes interesadas (stakeholders)

Las partes interesadas en el proyecto son aquellas personas o entidades que tienen un interés en el mismo, ya sea porque se ven afectadas por el resultado del proyecto, o porque tienen algún tipo de interés en el mismo. Las partes interesadas en este proyecto son las siguientes:

1. **Okticket**: la empresa es la principal parte interesada en el proyecto, ya que es la que se beneficiará directamente de los resultados del mismo, así como de las oportunidades de negocio que se abren con la explotación de los datos.
    - **Equipo de desarrollo**: el equipo de desarrollo es otra parte interesada en el proyecto, ya que son los encargados de llevar a cabo la implementación del sistema y de garantizar su correcto funcionamiento.
    - **Equipo de soporte**: el sistema planteado ahorraría tiempo al equipo de soporte, ya que les permitiría analizar los datos de forma más eficiente e identificar problemas antes de que tener que resolver las peticiones de los clientes afectados.
2. **Clientes**: los clientes de la empresa también son partes interesadas, puesto que se beneficiarán de los nuevos servicios que se ofrecen, como los dashboards de negocio que se han descrito anteriormente.
3. **Investigador y desarrollador:** el desarrollador del proyecto tiene la oportunidad de aplicar los conocimientos adquiridos en el desarrollo de un proyecto real, y de adquirir nuevos conocimientos en el proceso.

3.1.2. Metodología

3.1.2.1. Scrum

Para la planificación del proyecto se ha escogido la metodología *Scrum*, que es una metodología ágil que se basa en la realización de iteraciones cortas y en la adaptación a los cambios. La metodología *Scrum* se basa en la realización de *sprints* (iteraciones cortas de una duración fija), en las que se llevan a cabo una serie de tareas que se han planificado previamente.

El primer paso de la metodología *Scrum* es la creación de un *product backlog*, una lista ordenada de las tareas a realizar durante el desarrollo del producto, a partir de los requisitos del sistema, que a su vez son una versión refinada de los requisitos iniciales del proyecto. A partir de este *product backlog* se planifican las tareas que se llevarán a cabo en cada *sprint*, de manera que sea posible cumplir con los objetivos del proyecto en el tiempo establecido.

3.1.2.2. Visualización de la planificación

Para la visualización de la planificación se ha utilizado la herramienta de gestión de proyectos de *GitHub*, que permite múltiples visualizaciones de tareas e *issues* en tableros separados.

- Se utiliza un tablero de *requisitos* al estilo *Kanban* para visualizar los requisitos del proyecto y su estado, siguiendo con la metodología *Scrum*.
- Se utiliza un *roadmap* de tareas, donde se visualizan las tareas y los hitos del proyecto, así como su estado y sus fechas límite.

Figura 3.1: Roadmap de tareas

3.1.2.3. Comunicación

La comunicación con los tutores y con el equipo de desarrollo se considera fundamental para el correcto desarrollo del proyecto. Puesto que el trabajo se desarrolla de manera presencial en la oficina de la empresa, la comunicación con el equipo de desarrollo se realiza de manera frecuente y directa, mientras que la comunicación con los tutores se realiza de manera remota pero igual de frecuente, manteniendo el contacto mediante correo electrónico y Teams para pedir revisiones e informar sobre el estado del trabajo en todo momento.

3.2. Presupuesto inicial

4. Análisis

4.1. Definición del sistema

4.2. Análisis de alternativas

5. Diseño del sistema

5.1. Arquitectura del sistema

5.2. Modelo de datos

6. Implementación

7. Resultados

8. 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.