Presentación de datos: dashboards y procesos ETL

Juan Francisco Mier Montoto, UO283319
Inteligencia de Negocio, primera entrega de teoría
GIITIN EPI Gijón, 23-24

Índice

Índice de contenidos

Índice de figuras

Bibliografía

- ¿Qué es un dashboard y para qué se usa? (2024)
- Dashboard y su significado estratégico | Kyocera
- Qué es Dashboard - Definición, significado y ejemplos
- Cuadro de mando - Wikipedia, la enciclopedia libre
- Conceptos de Data Warehouse: enfoque de Kimball vs. Inmon | Astera
- Dashboard (business) - Wikipedia
- https://books.google.es/books?hl=es&lr=&id=5nuYDwAAQBAJ&oi=fnd&pg=PR9&dq=types+of+dashboards&ots=AJRhsZWvlD&sig=-m8pVsbTE4sxHamYNmY8JUBr8u4#v=onepage&q=types%20of%20dashboards&f=false
- Software Engineering Dashboards: Types, Risks, and Future | SpringerLink
- Data Warehouse Design Process - Scaler Topics
- ¿Qué son los procesos ETL?
- *GPT 4 para generación de ideas, refinamiento de texto y corrección de errores.*
- *Github Copilot X para generación de ideas.*
- *Transparencias de la asignatura (Tema 2, Dashboards)*

Todos los enlaces, accedidos el 5 de noviembre de 2023.

Introducción

En la actualidad, la capacidad de una organización para tomar decisiones informadas y bien optimizadas se ha convertido en una piedra angular del objetivo de cualquier empresa para tratar de conseguir una ventaja competitiva. El volumen y la variedad de datos que se pueden obtener y procesar ha ido aumentando exponencialmente con el desarrollo de la tecnología, convirtiendo la gestión de esta información en una tarea compleja y crítica al mismo tiempo.

La capacidad de presentar datos de manera visual y comprensible se ha convertido en una herramienta fundamental para la toma de decisiones informadas. Los dashboards son una herramienta que permite presentar datos de manera visual, y que se utilizan para tomar decisiones informadas sobre un proceso o negocio. Estos dashboards suelen mostrar datos en tiempo real, y permiten a los directivos tomar decisiones informadas sobre el futuro de la empresa.

En el contexto de esta asignatura, y por lo tanto de este trabajo, se pretende explicar y expandir los conceptos "cuadro de mando" (dashboard) y "procesos de extracción, transformación y carga" (ETL), así como su importancia en el sector. Además de dar ejemplos de herramientas y procesos que se utilizan hoy en día, se realizarán unas breves demostraciones prácticas de algunas de estas herramientas.

Conceptos

¿Qué es un dashboard?

La palabra "dashboard", que traducido de manera literal es "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 nuestro ámbito, los dashboards suelen reflejar en tiempo real el rendimiento de una actividad o negocio, y se utilizan para tomar decisiones informadas sobre el mismo. Por ejemplo, un dashboard puede mostrar el número de ventas de un producto, el número de clientes, el número de productos defectuosos, etc. En el caso de una empresa, un dashboard puede mostrar el rendimiento de la misma en tiempo real, y permitir a los directivos tomar decisiones informadas sobre el futuro de la empresa.

Figure 1. Ejemplo de dashboard (Plecto)

Características

- **Visualización de datos:** es la característica fundamental de cualquier cuadro de mandos, y aquella que determina su utilidad. La visualización de datos es la ciencia de presentar los datos de manera que se pueda extraer información útil y realizar decisiones informadas sobre ellos. Un buen dashboard cuenta con gráficas, tablas, indicadores, etc. que permitan al usuario entender la información que se está presentando.
- **Interactividad y personalización:** un dashboard "moderno" debe permitir al usuario interactuar con los datos (filtrarlos, ordenarlos, profundizar en ellos...) y ajustar la información que se muestra a cada proceso o negocio que se esté evaluando. Esta capacidad asegura que el dashboard se adapte tanto a las necesidades actuales como a las evoluciones futuras de lo que se esté analizando.
- **Accesibilidad:** puesto que estamos hablando de tecnologías modernas con una gran cantidad de herramientas disponibles, se da por hecho que un dashboard debe ser accesible desde una variedad de situaciones y dispositivos, manteniendo su funcionalidad y forma. Aunque normalmente los dashboards se analizan en pantallas grandes, es importante que también se puedan consultar en otras circunstancias, como dispositivos móviles.

Terminología clave

Por lo general, los dashboards como concepto van siempre unidos a otros conceptos representados por sus siglas en inglés:

- **KPI** (*Key Performance Indicator*): valor cuantificable que permite medir el rendimiento de un proceso o actividad. Los valores de los KPIs son los números que se muestran en los dashboards, que se pueden mostrar de manera individual o en conjunto con otros KPIs, incluso comparándolos con mediciones históricas, en multitud de formatos.
- **BI** (*Business Intelligence*): conjunto de estrategias y herramientas que permiten transformar los datos en información útil para la toma de decisiones.
- **ETL** (*Extract, Transform, Load*): esta definición se expandirá en el siguiente apartado, pero básicamente se trata de un proceso que permite extraer datos de una fuente, transformarlos y cargarlos en otra fuente.

Aspectos técnicos

Aunque los dashboards se pueden crear de muchas maneras, en la actualidad se suelen crear utilizando herramientas de software que permiten crearlos de manera visual, sin necesidad de programar. Estas herramientas suelen tener una interfaz gráfica que permite al usuario arrastrar y soltar elementos para crear el dashboard, y suelen tener una gran variedad de elementos disponibles para crearlos. En apartados posteriores se profundizará en algunas de estas herramientas.

Puesto que los dashboards muestran información crítica sobre procesos o negocios a menudo vitales, es importante que los dashboards cuenten con una serie de protecciones y medidas de seguridad para evitar que los datos sean accesible públicamente o se puedan vulnerar de alguna manera, tanto los datos como el acceso al dashboard en sí.

Además, un control de mando debe ser capaz de conectar y funcionar, en cada caso, con las bases de datos y plataformas software que se estén utilizando.

¿Qué son los procesos ETL?

Los procesos ETL 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 dashboard muestra datos diferentes, 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.

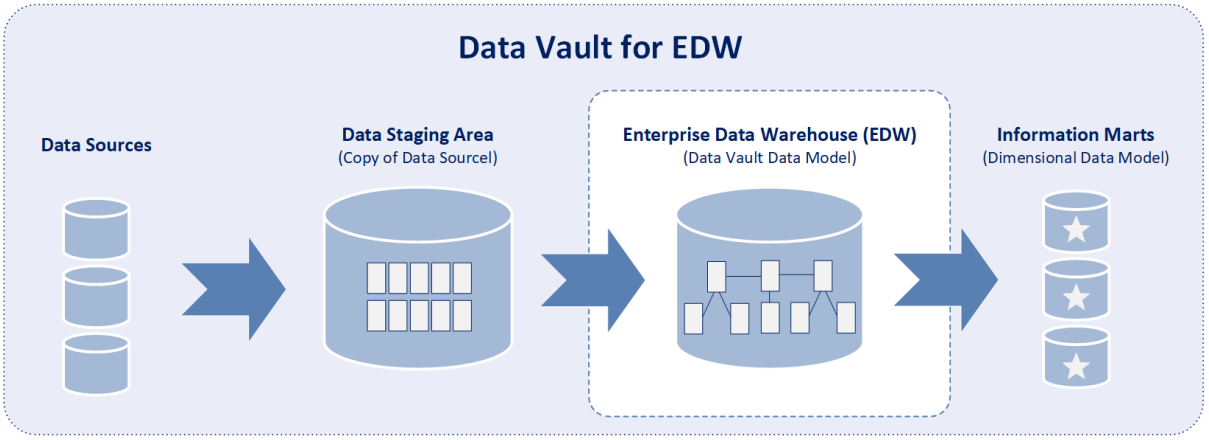

Figure 2. Flujo de información generado por un proceso ETL

Relevancia

Todos estos "datos" de los que hablamos no vienen de la misma fuente, ni en el mismo formato, ni siquiera representan todos lo mismo. Por lo general, los datos que se utilizan en los dashboards provienen de diferentes fuentes, como bases de datos, logs, APIs, etc. Además, estos datos pueden estar en diferentes formatos, como CSV, JSON, XML, etc. Por último, estos datos pueden representar diferentes cosas, como ventas, clientes, productos, etc.

Es por este motivo por lo que es relevante la existencia de procesos ETL, ya que permiten extraer los datos de las diferentes fuentes, transformarlos en un formato común y cargarlos en el dashboard para que se muestren de manera visual.

Evolución

Origen

Los procesos ETL surgieron con la aparición de las bases de datos relacionales, que permitían almacenar grandes cantidades de datos de manera estructurada. En un principio,

se solía almacenar los datos en bruto en las bases de datos, lo cuál no facilitaba la extracción de información útil. Para solucionar este problema, las primeras herramientas ETL convertían los datos en datos *relacionales*, es decir, los transformaban en tablas que se podían relacionar entre sí. De esta manera, se podía extraer información útil de los datos almacenados en las bases de datos.

Evolución

Con la evolución de estas herramientas, las fuentes y los tipos de datos fueron aumentando de manera *exponencial*, lo que llevó a la creación de herramientas más complejas que permitían extraer datos de diferentes fuentes, transformarlos en diferentes formatos y cargarlos en diferentes destinos.

A continuación se listan algunas de las nuevas tecnologías o conceptos que se han ido desarrollando con la evolución de los procesos ETL:

- Los *data sinks* (o **sumideros de datos**), son capaces de recibir datos de múltiples fuentes y disponen de la elasticidad característica de los servicios en la nube que permiten escalar de manera automática según las necesidades de cada momento, en este caso según el volumen de datos que se esté procesando. Además, los *data sinks* suelen tener un coste muy bajo, ya que se paga por el volumen de datos que se procesa, y no por el volumen de datos que se almacena.
- Los **almacenes de datos** o *data warehouses* son bases de datos que almacenan grandes cantidades de datos de manera estructurada, y que se utilizan para realizar análisis de negocio. Estos almacenes de datos suelen ser el destino final de los procesos ETL, ya que permiten almacenar grandes cantidades de datos de manera estructurada y optimizada para realizar análisis de negocio.
- Los **lagos de datos** o *data lakes* son almacenes de datos que almacenan grandes cantidades de datos de manera no estructurada. A diferencia de los *data warehouses*, los *data lakes* no tienen un esquema definido, lo que permite almacenar datos de cualquier tipo y formato. Esto permite almacenar grandes cantidades de datos sin tener que definir un esquema de antemano, lo que puede ser útil en algunos casos. Sin embargo, esto también puede ser un inconveniente, ya que no se puede realizar un análisis de negocio de los datos almacenados en un *data lake* sin antes transformarlos en un formato estructurado.

El almacenamiento de datos es un tema muy amplio y complejo, con muchas alternativas y posibilidades (metodologías *Inmon* y *Kimball*[1], enfoques *top-down* y *bottom-up*[2], etc.), por lo que no se profundizará más en este tema.

[1] Aprovechando la mención de Kimball en las transparencias de teoría: https://www.astera.com/es/type/blog/data-warehouse-concepts/
[2] https://www.scaler.com/topics/data-warehouse-design-process/

Funcionamiento

Como su propio nombre indica, los procesos ETL se dividen en tres partes:

Extracción

En este proceso se extraen los datos de las fuentes de datos, que pueden ser bases de datos, logs, APIs, etc. En este paso, se pueden aplicar filtros para extraer solo los datos que se necesiten, y se pueden extraer datos de múltiples fuentes.

Frecuentemente, los datos brutos se almacenan temporalmente en una zona de almacenamiento intermedia llamada *staging area* (que es estrictamente *transitoria*).

En algunos casos, los datos se pueden extraer de manera incremental, es decir, solo se extraen los datos que han cambiado desde la última extracción. Esto puede ser útil para reducir el tiempo de procesamiento y el volumen de datos que se almacenan.

En otros casos, los datos se pueden extraer de manera continua, es decir, se extraen los datos en tiempo real según se van generando. Esto puede ser útil para procesar datos que se generan en tiempo real, como logs o datos de sensores.

Transformación

En este proceso se transforman los datos extraídos en un formato común, normalmente tablas relacionales.

Una transformación básica de datos es la limpieza, revisión y corrección de los datos extraídos, para asegurar que los datos que se almacenan son correctos y consistentes. Otras operaciones más complejas pueden ser la agregación de datos, la conversión de formatos, la normalización de datos, el cifrado, etc.

Carga

En este proceso se cargan los datos transformados en el destino final.

Frecuentemente, los datos se almacenan en una *data warehouse* o *data lake* para su posterior análisis.

En algunos casos, los datos se pueden cargar de manera incremental, es decir, solo se cargan los datos que han cambiado desde la última carga. Esto puede ser útil para reducir el tiempo de procesamiento y el volumen de datos que se almacenan. En otros casos, los datos se pueden cargar de manera continua, es decir, se cargan los datos en tiempo real según se van generando. Esto puede ser útil para procesar datos que se generan en tiempo real, como logs o datos de sensores.

Alternativas

Aunque lo más común es el flujo anteriormente explicado de *Extracción → Transformación → Carga*, existen algunos flujos y procesos alternativos que evitan algunos de estos pasos, normalmente en casos específicos que se beneficien del cambio:

- **Virtualización:** en lugar de extraer los datos de las fuentes, se crea una capa virtual que permite acceder a los datos de las fuentes sin necesidad de extraerlos. Esto permite ahorrar espacio de almacenamiento y tiempo de procesamiento, pero puede ser menos eficiente en algunos casos.
- **Proceso ELT:** en lugar de transformar los datos antes de cargarlos en el destino, se cargan los datos en bruto y se transforman en el destino. Funciona bien para grandes conjuntos de datos sin estructura que requieran una carga (o recarga) contínua, aunque, al igual que la virtualización, puede ser menos eficiente en algunos casos.

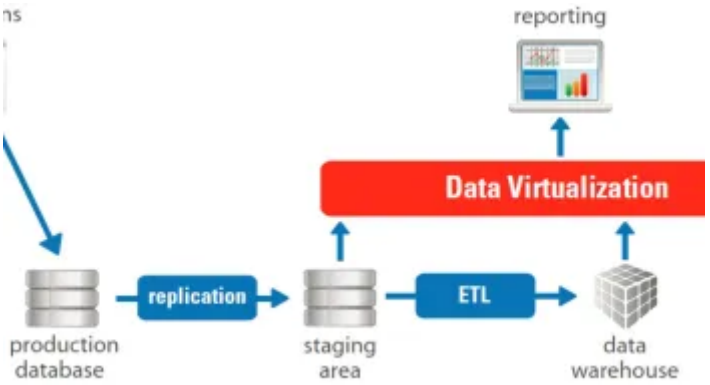

Figure 3. Ejemplo de virtualización (en lugar de procesos ETL)

Diseño de dashboards

Principios de diseño y visualización de datos

Aunque los dashboards se pueden crear de muchas maneras, en la actualidad se suelen crear utilizando herramientas de software que permiten crearlos de manera visual, sin necesidad de programar. Estas herramientas suelen tener una interfaz gráfica que permite al usuario arrastrar y soltar elementos para crear el dashboard, y suelen tener una gran variedad de elementos disponibles para crearlos. En apartados posteriores se profundizará en algunas de estas herramientas.

Pese a que la creación de dashboards sea intuitiva y visual, es importante tener en cuenta principios básicos a la hora de crear o diseñar una herramienta de visualización de datos importante como lo son los dashboards: simplicidad, consistencia, relevancia o interactividad pueden parecer palabras sueltas y carentes de significado, pero son buenos

atributos a tener en cuenta respecto a nuestro dashboard para mejorar su utilidad y posibilidad de mejora en el futuro.

Herramientas

Como se ha mencionado repetidas veces en este trabajo, existen multitud de herramientas que facilitan la creación (y el mantenimiento) de dashboards de manera interactiva y visual, sin necesidad de tener un conocimiento profundo de programación. A continuación se explicarán algunas de las herramientas más utilizadas en la actualidad.

Cada herramienta, aparte de tener sus ventajas y desventajas, está enfocada a distintos tipos de dashboards, por lo que es importante elegir la herramienta adecuada para cada caso.

Algunas de las herramientas más populares (incluyendo también las que se mencionan en las transparencias de teoría) son:

- Tableau
- Grafana
- Power BI
- QlikView/Qlik Sense
- NewRelic

En las transparencias de teoría se encuentra un buen resumen de populares herramientas, incluyendo las ventajas y las desventajas específicas a cada una.

Demostraciones prácticas

Demostraciones con Python y Jupyter Notebook

Como exigido en el enunciado de la práctica, se incluyen un ejemplo de uso ligeramente más elaborado que el presentado en clases de teoría[3].

ipywidgets

El notebook está realizado utilizando la librería *ipywidgets*, que genera una ventana interactiva, en este caso mostrando información sobre el clásico *dataset* de *Iris*:

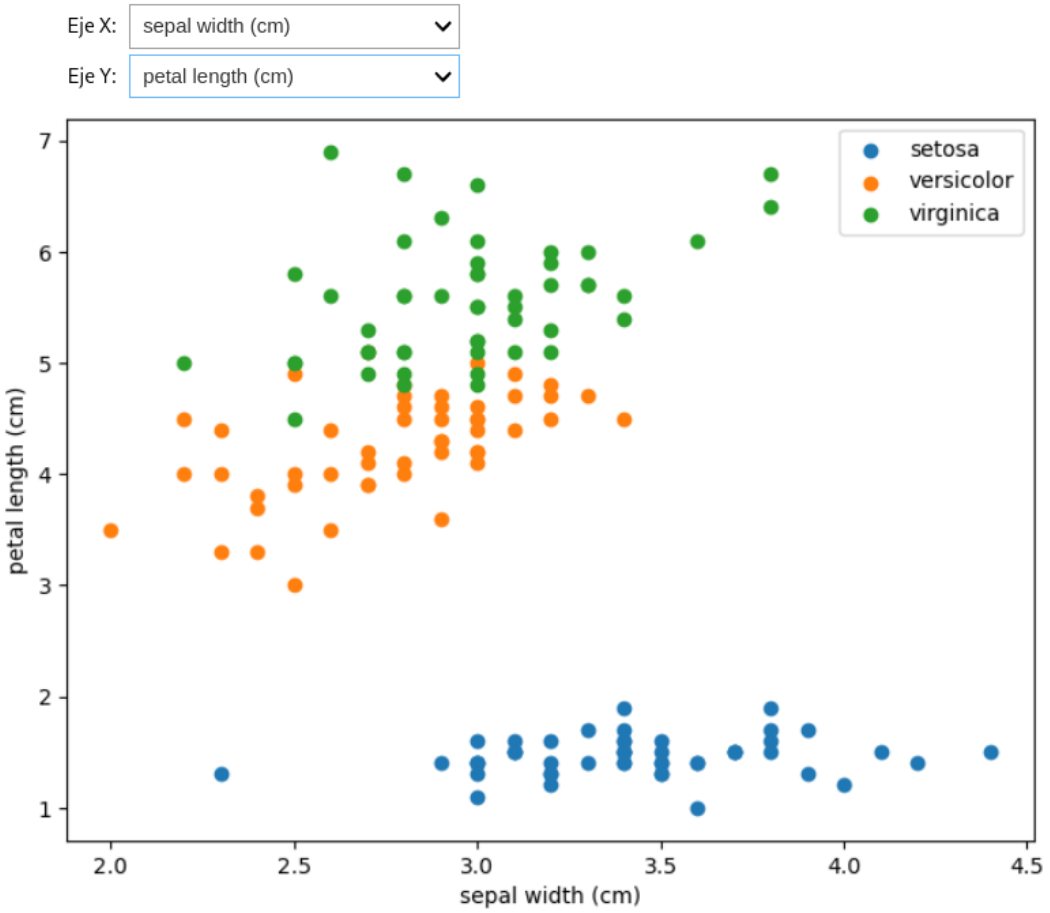

Figure 4. Resultado de ejecución del notebook de Ipywidgets

Las instrucciones de uso se encuentran dentro del Notebook entregado.

[3] Extraídos del siguiente enlace: https://txtify.it/https://towardsdatascience.com/4-python-packages-to-create-interactive-dashboards-d50861d1117e

voila

Haciendo uso de otra librería, *voila*, se puede convertir exactamente el mismo notebook a formato web, algo más “profesional”:

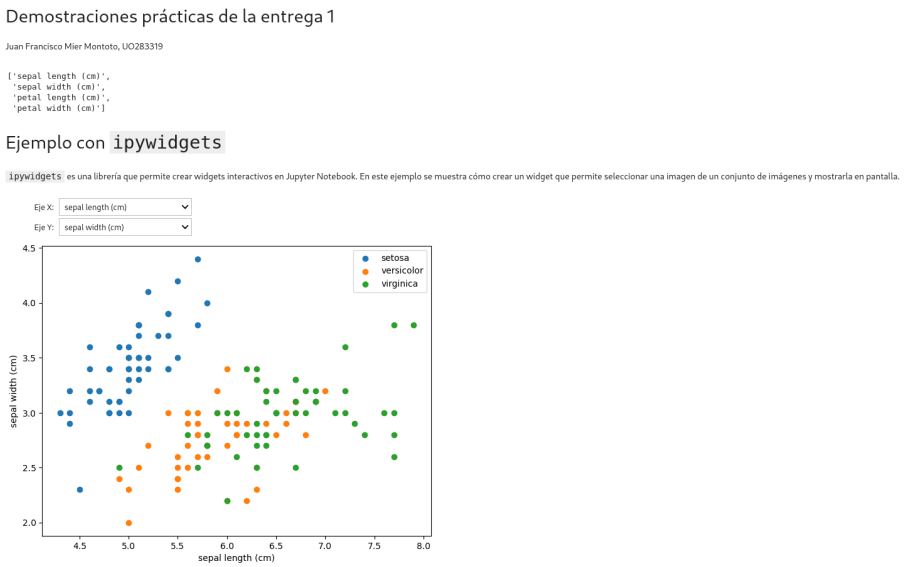

Figure 5. Resultado de ejecución del servidor de voila

Por supuesto, esta versión web del notebook es también interactiva, por lo que se puede hacer uso de los desplegables, botones, etc., con los que cuente el *widget* original. Para ejecutar esta versión web, tan solo hay que ejecutar el siguiente comando tras instalar la librería:

```
voila dashboards-widgets.ipynb
```

Dash

Fuera de los notebooks de Jupyter, existen otras librerías que permiten la generación de dashboards interactivos a través del navegador. Una de ellas es Dash, que se ejecuta como si de un script normal de Python se tratara.

En el siguiente ejemplo, vemos una representación de ejemplo de la expectativa de vida con respecto al producto interior bruto de todos los países asiáticos gracias al dataset *Gapminder*.

```
python dashboards-dash.py
```

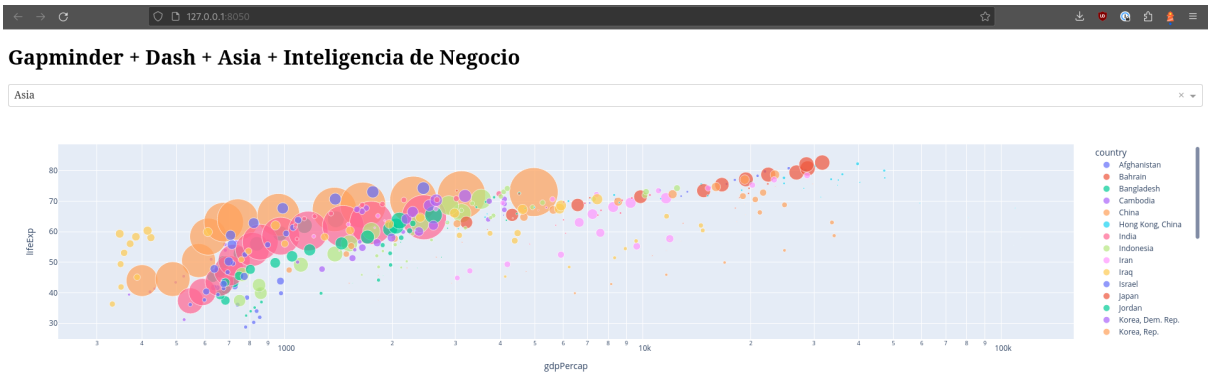

Figure 6. Visualización de Gapminder con Dash

Streamlit

Por último, vemos otro ejemplo de uso de la librería Streamlit, muy similar a Dash pero con una presentación más cercana a un dashboard real.

```
streamlit run dashboards-streamlit.py
```

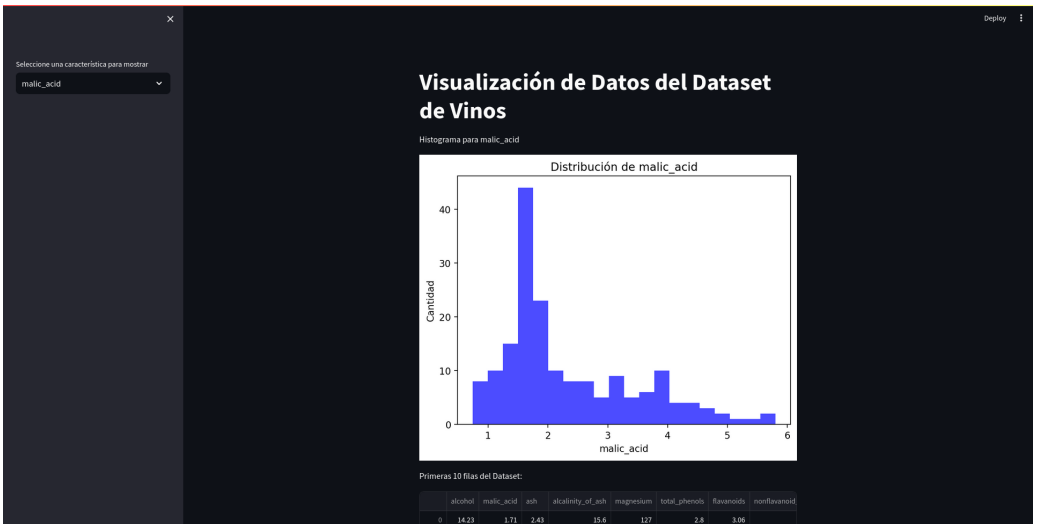

Figure 7. Visualización del dataset de Vinos con Streamlit

Demostración real de un dashboard funcional

Durante mis prácticas extracurriculares en verano, una de las tareas a desarrollar era la implementación de un dashboard bonito y funcional que mostrara métricas de uso, tiempo de respuesta y disponibilidad de las diferentes áreas y servicios de la empresa.

Puesto que toda la información relevante se encontraba en los balanceadores de carga de Amazon (ALB), hicimos uso de la herramienta *NewRelic*, que tiene plantillas de conexión que facilitan el *traslado* de datos de muchos servicios, entre ellos PHP y ALB.

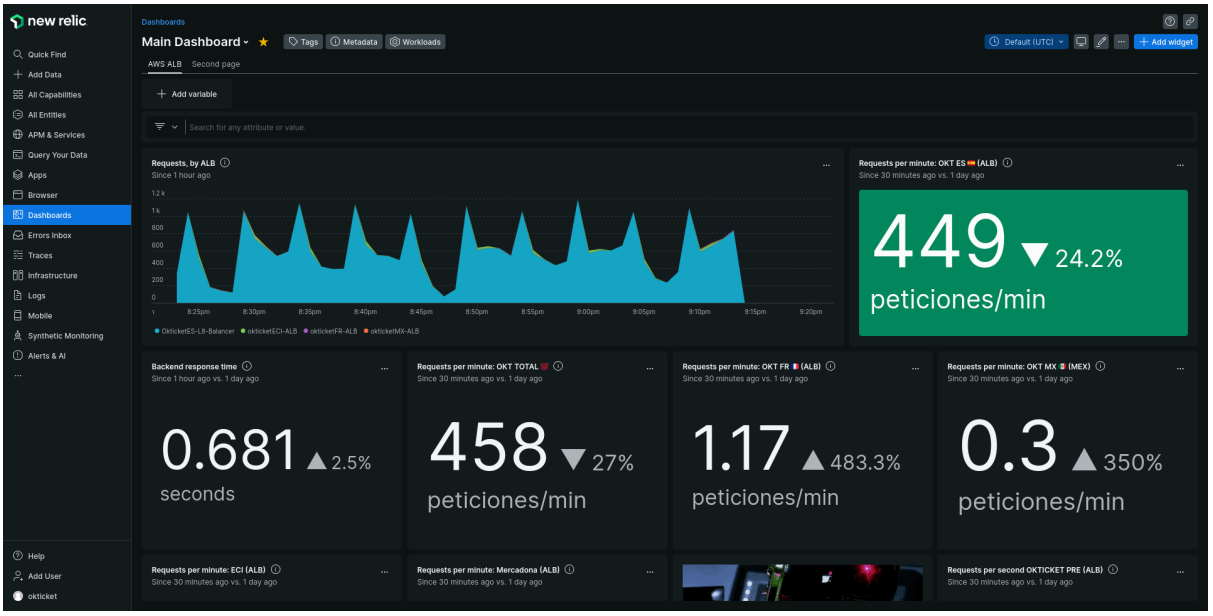

Figure 8. Dashboard principal en New Relic

Los datos de los balanceadores se manipulan para generar gráficas, mostrar tiempos de respuesta, comparar con históricos (normalmente del día anterior), y demás información. Dichos datos se obtienen mediante *queries* en *NRQL*, el lenguaje de SQL propietario de New Relic.

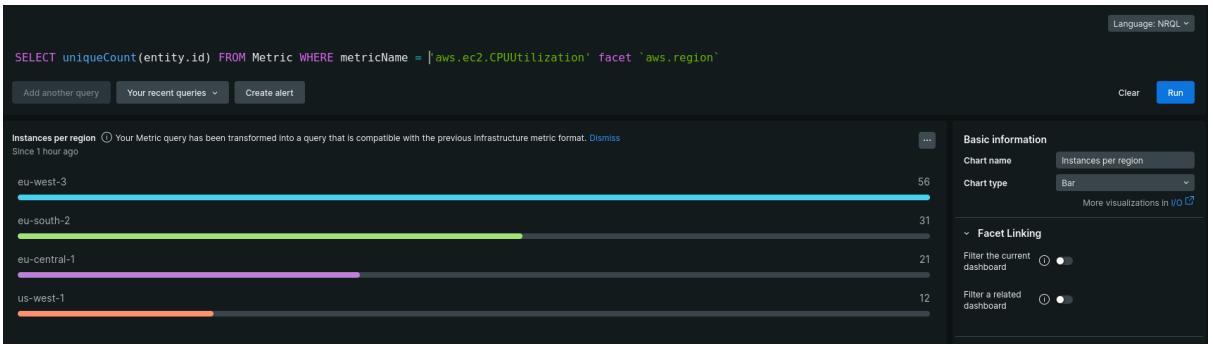

Figure 9. Ejemplo de query NRQL en New Relic

Gracias a la conexión directa con AWS y los balanceadores de carga, se pueden obtener todo tipo de datos de manera que los procesos ETL que funcionan por detrás son totalmente transparentes para nosotros, los usuarios.

Figure 10. Dashboard secundario con información genérica de AWS en New Relic

Demostración de juguete de un dashboard de mis propios servicios

Durante el curso pasado, desarrollé un *dashboard personal*, a partir de la herramienta *Uptime Kuma*[4], con la que poder comprobar la conectividad y disponibilidad de mis servicios y páginas web personales de manera continua.

Gracias a la sencillez de la aplicación base, que cuenta con menús muy interactivos e intuitivos (características clave de un dashboard como ya hemos visto), la modificación de los detalles de conexión con los servicios es increíblemente sencilla.

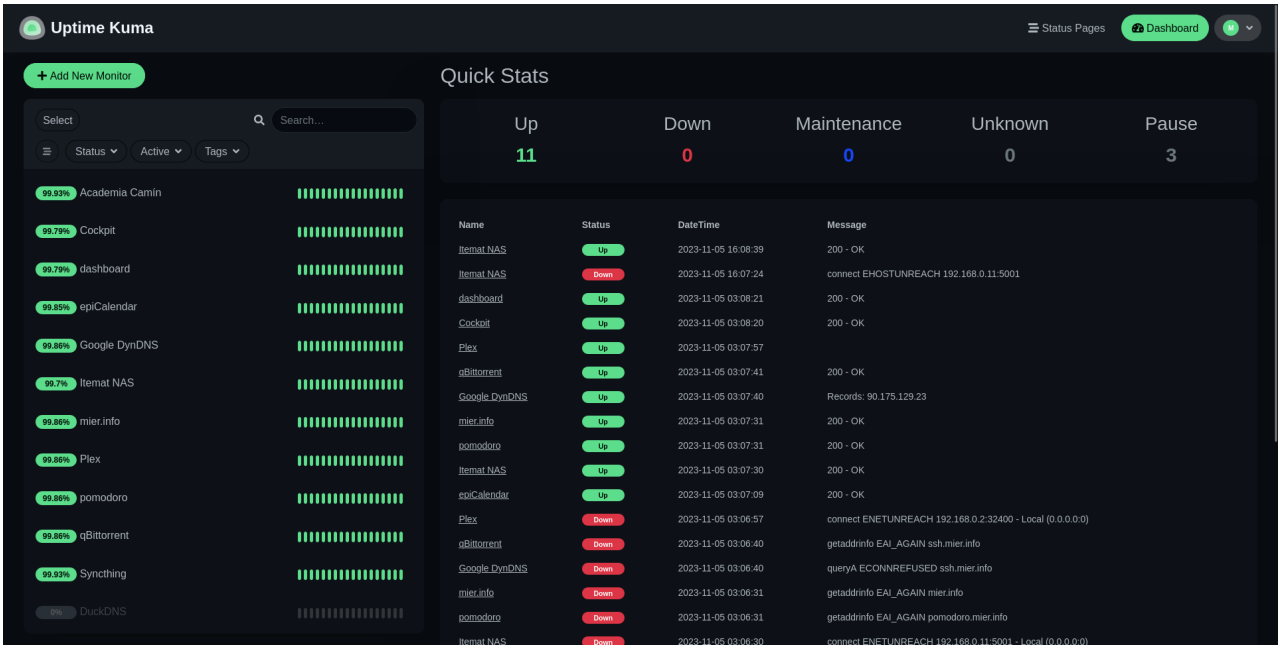

Figure 11. Menú de administrador (monitores) en Uptime Kuma

Además, el sistema está conectado con un *bot* de Telegram, que envía alertas de manera automática cuando se cae alguno de los servicios o dominios, lo que facilita tareas de mantenimiento.

El dashboard es accesible de manera pública, aunque obviamente cuenta con un sistema de inicio de sesión de administrador para editar los servicios o el estilo de la página.

[4] https://uptime.kuma.pet/

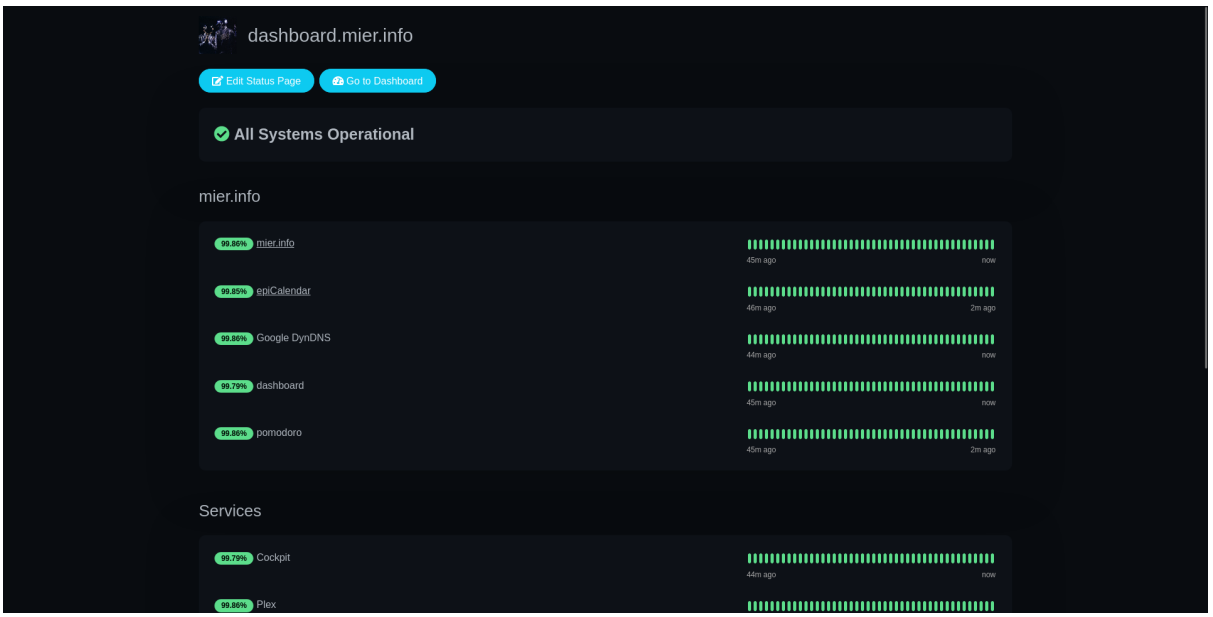

Figure 12. Dashboard principal personal en Uptime Kuma

La página personal cuenta con estilos personalizados gracias a CSS inyectado y está montado sobre computadores personales que funcionan las 24 horas del día, que a su vez apuntan hacia mi dominio personal.

Obviamente, el alcance de los dashboards, en especial de aquellos *enfocados al negocio* como se ve en esta asignatura, tienen una importancia y profundidad diferente a estos servicios “de juguete”, pero sirven como una primera aproximación excelente.