

Practica 2: Azure ML

Integrantes del proyecto:

- ❖ Mateuz Roman Kolakowski Dziewic
- ❖ Fang, Yencheng
- ❖ Hernández López, Carlos

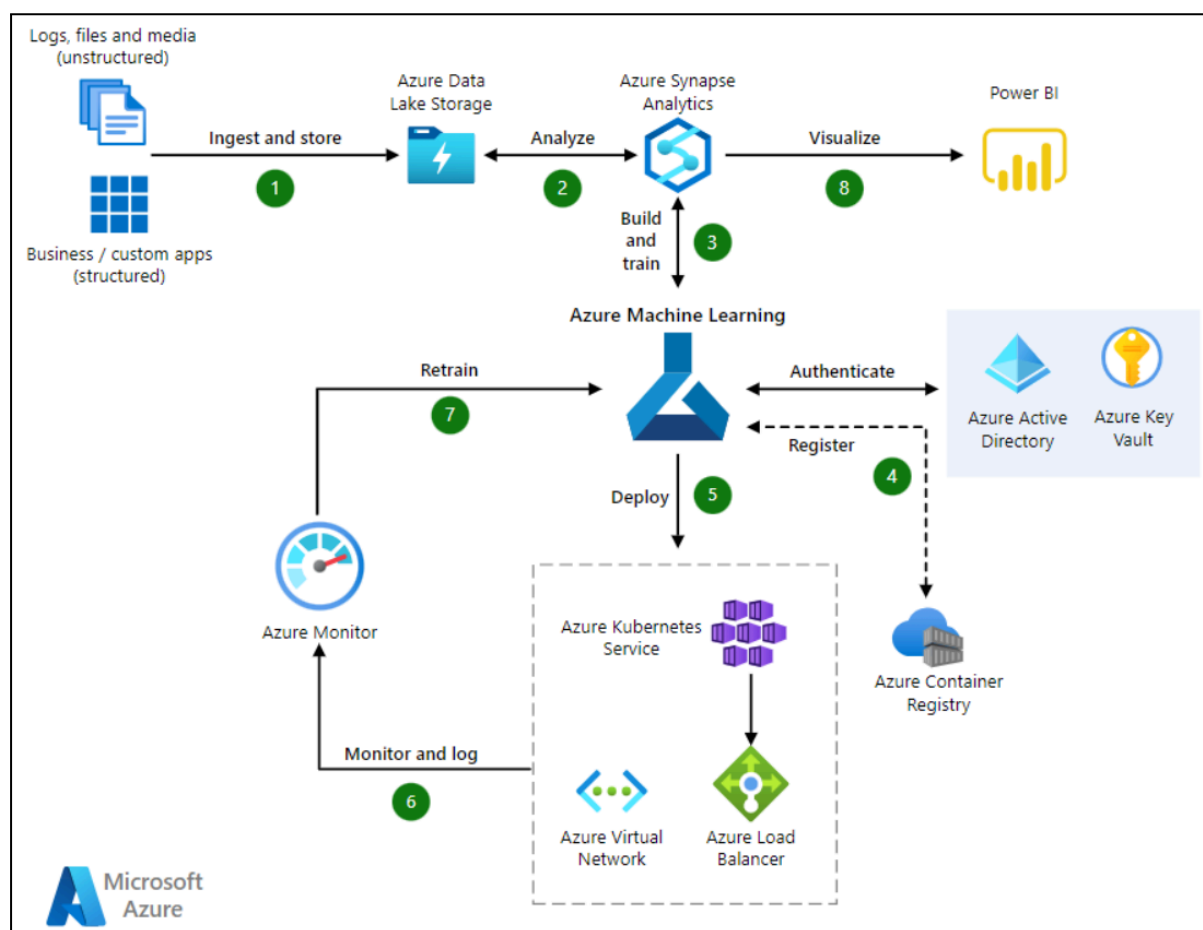
Máster universitario en aprendizaje automático y datos masivos.

Asignatura/s: Optimización de grandes volúmenes de datos

Marzo 2024

Flujo de datos y objetivos

El proyecto se desarrollará en Azure, aprovechando sus servicios para gestionar datos, realizar análisis, entrenar modelos de Machine Learning y visualizar resultados. El enfoque se centrará en el uso de Edge Computing para optimizar el procesamiento de datos y la implementación de modelos en contenedores.



1. Preparación de datos

- **Recolección de datos:** Se utilizará el dataset seleccionado, que será almacenado en Azure Data Lake Storage Gen2. Este dataset puede contener datos estructurados, no estructurados y semiestructurados.
- **Limpieza y transformación:** Se empleará Apache Spark en Azure Synapse Analytics para realizar tareas de limpieza, transformación y análisis exploratorio de datos (EDA).

2. Modelado de Machine Learning

- **Entrenamiento de modelos:** Utilizando Azure Machine Learning, se construirán y entrenarán modelos de Machine Learning utilizando los datos procesados. Se explorarán modelos de aprendizaje supervisado para resolver problemas específicos identificados en el dataset.

3. Gestión y seguridad

- **Control de acceso y autenticación:** Se implementará Microsoft Azure Active Directory (Azure AD) para controlar el acceso a los datos y el área de trabajo de Machine Learning.
- **Seguridad de datos:** Se utilizará Azure Key Vault para gestionar y proteger claves, contraseñas y otros secretos utilizados en el proyecto.
- **Gestión de contenedores:** Los contenedores de los modelos de Machine Learning se administrarán con Azure Container Registry, permitiendo su distribución y escalabilidad.

4. Implementación y evaluación del modelo

- **Implementación del modelo:** Se desplegarán los modelos entrenados en contenedores utilizando Azure Kubernetes Service (AKS), garantizando una implementación segura y escalable.
- **Evaluación del rendimiento:** Se utilizarán métricas de registro y supervisión de Azure Monitor para evaluar el rendimiento de los modelos desplegados en producción.

5. Valor añadido

- **Visualización de datos:** Se empleará Power BI para crear visualizaciones interactivas a partir de los datos procesados y los resultados de los modelos de Machine Learning. Se generarán informes y paneles para facilitar la comprensión de los datos y los resultados del análisis.

Datos

Los datos utilizados en este proyecto fueron extraídos de Kaggle y se encuentran disponibles en el siguiente enlace: [Students Performance in Exams](#).

Repositorio en GitHub

Los datos han sido subidos a un repositorio en GitHub, facilitando su acceso y uso. El enlace al archivo CSV es el siguiente: [exams.csv](#).

Formato RAW

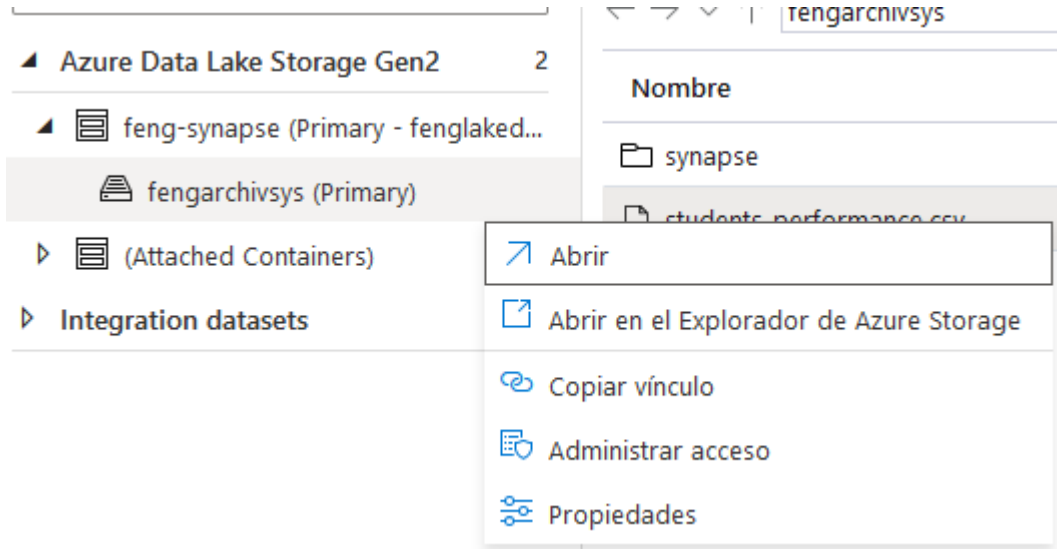
Los datos han sido importados en formato RAW desde GitHub para su procesamiento y análisis posterior. Este formato permite una fácil manipulación de los datos y su integración en diversas herramientas y plataformas de análisis.

Con esta estructura, se proporciona una manera clara y accesible para acceder a los datos y utilizarlos en el análisis correspondiente.

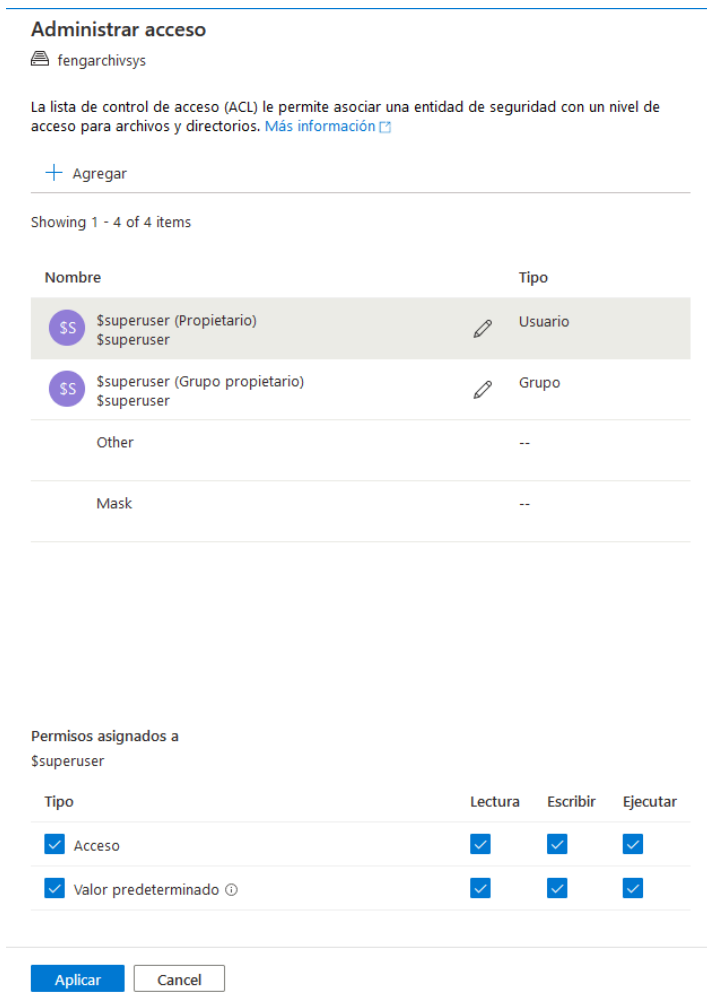
1. Ingesta de datos en el Data Lake

Se ha decidido utilizar un dataset de Kaggle.

2. Análisis de datos usando synapse analytics



The screenshot shows the Azure Data Lake Storage Gen2 interface. On the left, a sidebar lists storage containers: 'Azure Data Lake Storage Gen2', 'feng-synapse (Primary - fenglaked...', 'fengarchivsys (Primary)', '(Attached Containers)', and 'Integration datasets'. The main pane shows the 'fengarchivsys' container with a file named 'students_performance.csv'. A context menu is open over the file, showing options: 'Abrir', 'Abrir en el Explorador de Azure Storage', 'Copiar vínculo', 'Administrar acceso', and 'Propiedades'.



The screenshot shows the 'Administrar acceso' (Manage Access) dialog for the 'fengarchivsys' container. It explains that the ACL allows associating a security entity with a level of access for files and directories. Below this, there is a table showing the current ACL entries:

Nombre	Tipo
\$superuser (Propietario) \$superuser	Usuario
\$superuser (Grupo propietario) \$superuser	Grupo
Other	--
Mask	--

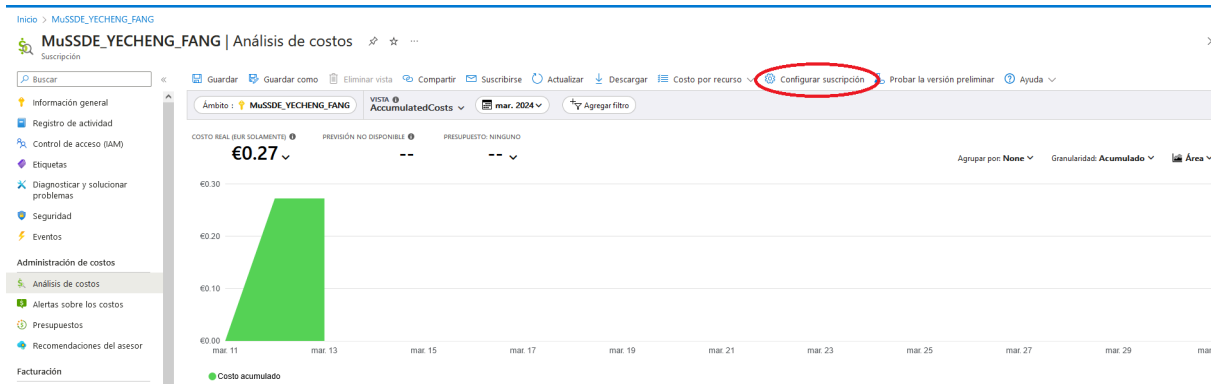
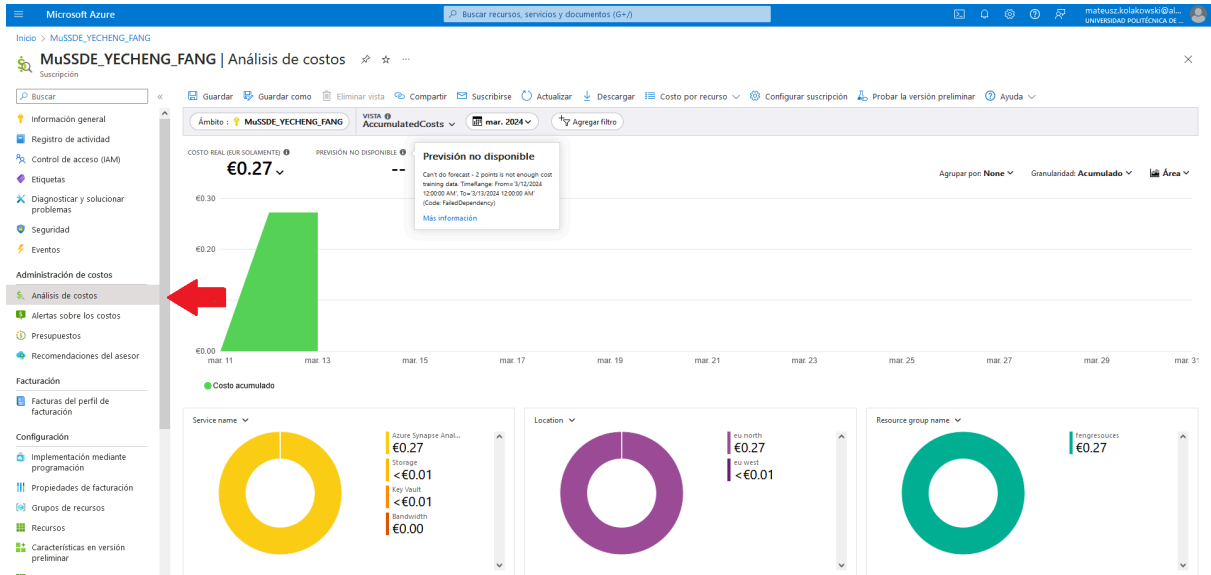
Below the table, it shows the permissions assigned to the '\$superuser' entity:

Tipo	Lectura	Escribir	Ejecutar
<input checked="" type="checkbox"/> Acceso	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/> Valor predeterminado ⓘ	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

At the bottom, there are 'Aplicar' and 'Cancel' buttons.

3. FinOps pre arranque del proyecto

3.1. Establecer Presupuesto de gasto



Microsoft Azure

Inicio > MuSSDE_YECHENG_FANG | Análisis de costos >

Configuración

Scope: MuSSDE_YECHENG_FANG (change)

[Rename](#) [Change directory](#) [Cancel subscription](#) [Probar la versión preliminar](#) [Abrir en Cost Management](#) [Ayuda](#)

Tag inheritance Disabled

Use tag inheritance to group resource costs.
[Editar](#)

Manage your cost

Alert rules

Configure email alerts for anomalies or low reservation utilization.

Budgets

Get notified or trigger custom actions before you exceed your budget.

Exports

Schedule automated exports of your data on a daily, weekly or monthly basis.

Connectors for AWS

View and manage your Azure and AWS costs in one place.

Manage your subscription

Resource groups

Manage applications and related resources deployed to resource groups.

Resources

Manage individual resources to apply tags for reporting or resize to reduce costs.

Properties

Additional details about this subscription.

Microsoft Azure

Inicio >

Presupuestos

Agregar Actualizar [Ayuda](#)

Ámbito: MuSSDE_YECHENG_FANG Todos los periodos

Nombre	↑↓	Ámbito	↑↓	Periodo de rest... ↑↓	Fecha de creaci... ↑↓	Fecha de expira... ↑↓	Presupuesto	↑↓	Previsión	↑↓	Gasto evaluado	↑↓	Progreso
No tiene ningún presupuesto.													

 Microsoft Azure[Inicio](#) > [Presupuestos](#) >

Crear presupuesto



Presupuesto

Ámbito del presupuesto

El presupuesto que cree se asignará al ámbito seleccionado. Use filtros adicionales, como grupos de recursos, para supervisar su presupuesto con mayor granularidad si lo necesita.


Ámbito  MuSSDE_YECHENG_FANG

Filtros

ResourceGroupName : **fengresouces**  Agregar filtro

Detalles del presupuesto

Asigne un nombre único a su presupuesto. Seleccione el período de tiempo que analiza durante cada período de evaluación, la fecha de expiración y la cantidad.

* Nombre * Período de restablecimiento  * Fecha de creación    * Fecha de expiración    

Importe del presupuesto

Indique un umbral de importe para el presupuesto.

Importe * [Anterior](#)[Siguiete >](#)

Microsoft Azure

Inicio > Presupuestos >

Crear presupuesto

Presupuesto

☒ Crear un presupuesto
 ☒ Establecer alertas

Configure condiciones de alerta y envíe notificaciones por correo electrónico en función de sus gastos.

*** Condiciones de alerta**

Tipo	Porcentaje del pr...	Importe	Grupo de acciones
Previsión	Escriba el porcenta...	-	Application Insi...
Seleccionar tipo	Escriba el porcenta...	-	Ninguno

[Administrar grupo de acciones](#)

*** Destinatarios de la alerta (correo electrónico)**

Destinatarios de la alerta (correo electrónico)

☒

Se recomienda agregar azure-noreply@microsoft.com a la lista de permitidos de correo electrónico para asegurarse de que los correos de alerta no vayan a la carpeta de correo no deseado.

Preferencia de idioma

Seleccione el idioma en el que quiera recibir el correo electrónico de alerta para todos los

Luego de crear un Budgets personalizado podrás ver el resultado en pantalla.

Inicio >

Presupuestos

Ámbito: mussa_yecheng_fang (Suscripción)

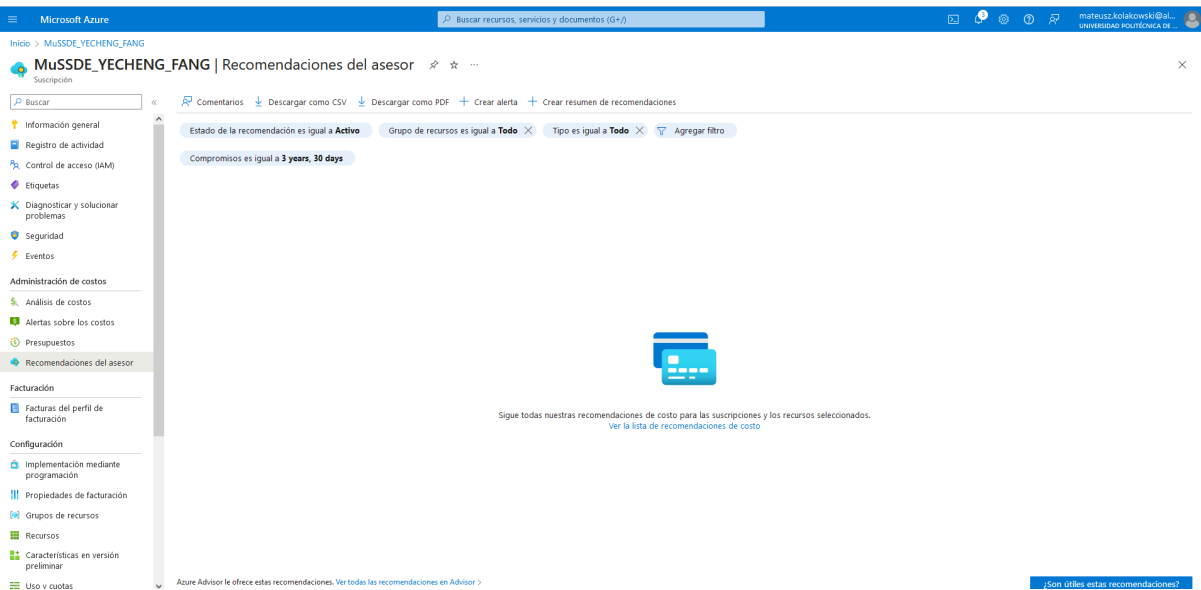
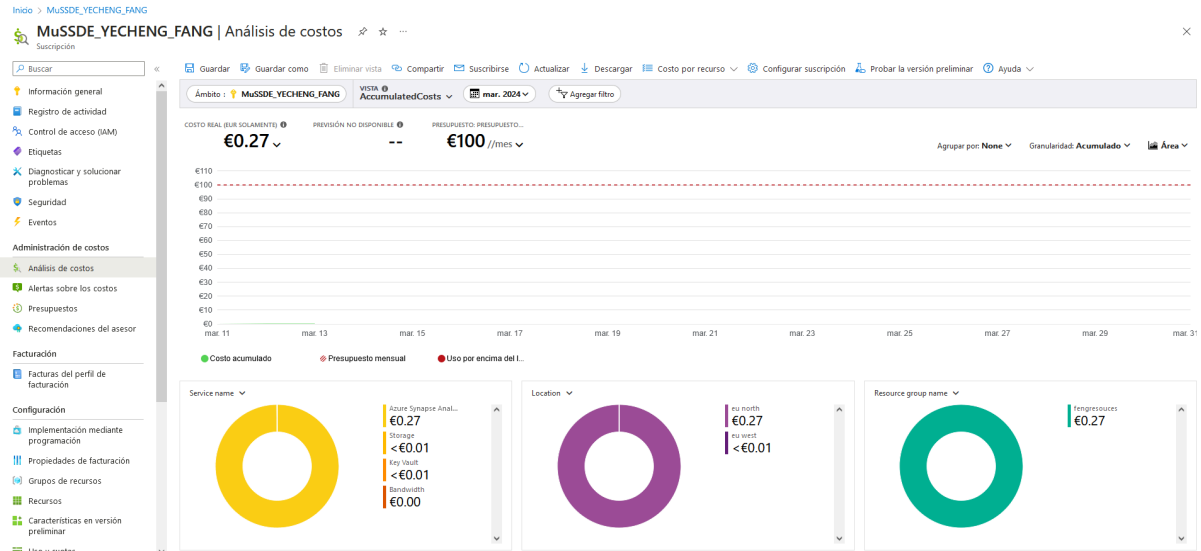
Ámbito: Microsoft, YECHEUNG, JANG

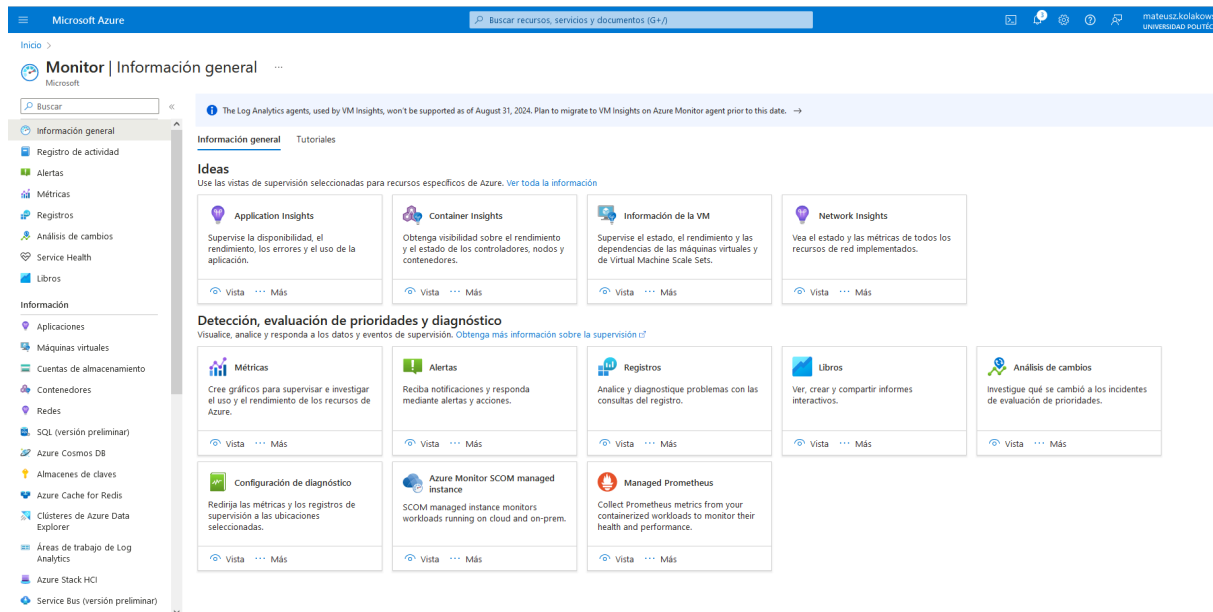
Nombre	Ámbito	Periodo de rest...	Fecha de creaci...	Fecha de expira...	Presupuesto	Previsión	Gasto evaluado	Progreso
Presupuestomensual	8b657da4-bd82-4401-8106-7...	Monthly	1/3/2024	28/2/2026	100,00 €	0	€0.00	0,00%

3.2 Analizar regularmente el comportamiento de los servicios

Realizar un análisis periódico del comportamiento de tus servicios en la nube es fundamental para identificar oportunidades de optimización. Herramientas como Análisis de Costos, Recomendación del asesor y Azure Monitor pueden ayudar en este proceso.

Examinar detenidamente en qué estamos gastando y buscar posibles áreas de mejora. Este análisis permitirá tomar decisiones informadas y reducir los costos innecesarios.





3.3. Aprovechar la automatización

La automatización es una aliada poderosa para optimizar costos en la nube. Una práctica común es apagar el cómputo durante las noches y los fines de semana, lo que puede representar un ahorro de hasta el 75% en los costos. Automatizando esta tarea utilizando herramientas y scripts que te permitan programar el encendido y apagado de instancias de manera eficiente. Esto asegurará que solo estés utilizando recursos cuando realmente los necesitas.

En nuestro caso, dado lo pequeño que es el proyecto y el budget tan bajo prácticamente no se ofrecen posibilidades.

4. Entrenamiento del modelo

En nuestro caso, se ha considerado interesante realizar el entrenamiento y despliegue del modelo mediante pipelines. Azure ofrece la posibilidad de entrenar los modelos mediante código, pero el uso de pipelines para el entrenamiento de modelos en lugar de código directo ofrece diversas ventajas organizativas y prácticas. En primer lugar, la automatización del flujo de trabajo es esencial para garantizar la consistencia y eficiencia en cada etapa, desde la preparación de datos hasta el despliegue del modelo en producción. La gestión de dependencias también es simplificada, asegurando que cada paso se complete antes de avanzar, evitando errores y garantizando la coherencia en todo el proceso.

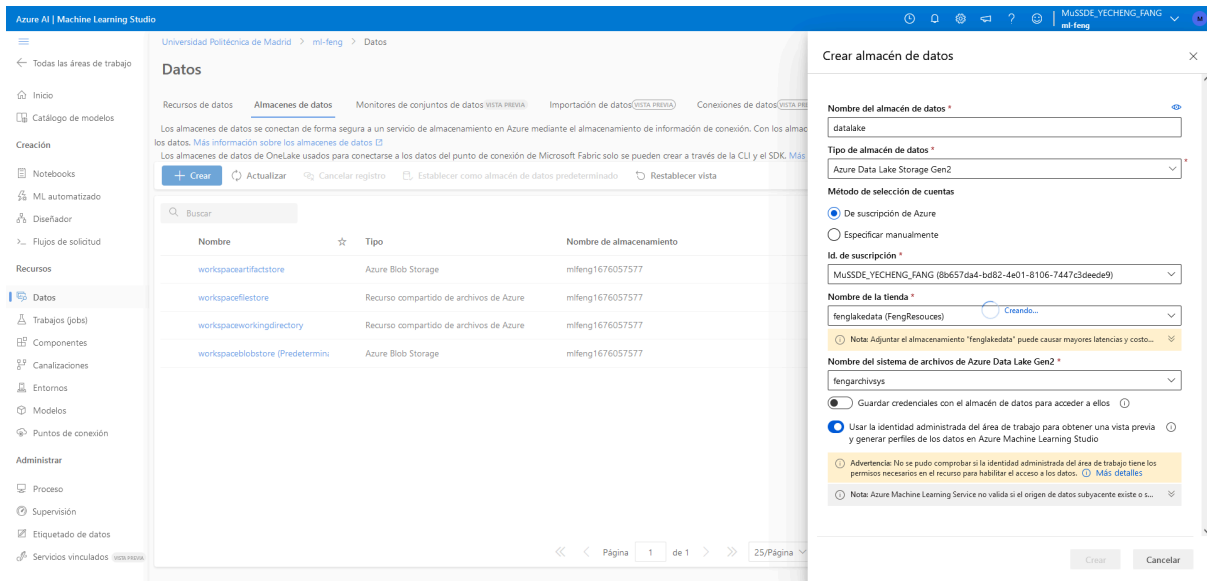
Además, los pipelines facilitan la reproducibilidad de los experimentos y resultados. Cada ejecución se registra y puede replicarse fácilmente, lo que simplifica la auditoría, la depuración y la mejora continua del modelo. La escalabilidad y paralelización de los pipelines aceleran el tiempo de entrenamiento y procesamiento, especialmente en tareas computacionalmente intensivas o con grandes conjuntos de datos.

La gestión centralizada de la configuración es otra ventaja, permitiendo la adaptación y ajuste de parámetros sin necesidad de modificar el código Python directo. Integrar pipelines con prácticas de integración continua/despliegue continuo (CI/CD) facilita el despliegue rápido y eficiente de modelos en producción, crucial en entornos ágiles y de desarrollo rápido.

Los pipelines también fomentan la colaboración entre equipos al proporcionar una descripción clara y separada del código Python específico, mejorando la documentación automática del proceso. Asimismo, permiten una gestión efectiva de versiones tanto de datos como de código, crucial para la reproducibilidad y el seguimiento de cambios a lo largo del tiempo.

En resumen, aunque el código Python directo puede ser útil para experimentar y prototipar modelos, los pipelines ofrecen un enfoque más estructurado y gestionado para el desarrollo de modelos a escala, proporcionando beneficios significativos en términos de eficiencia y mantenimiento a medida que los proyectos crecen en complejidad.

Para acceder a los datos, se ha realizado una conexión del recurso de ML con el Datalake de Gen 2 que utiliza Synapse



The screenshot shows the Azure Machine Learning Studio interface. On the left, the 'Datos' (Data) section is selected. The main area displays a table of data stores. On the right, the 'Crear almacén de datos' (Create data store) dialog is open, showing the configuration for a new data store named 'datalake'.

Nombre	Tipo	Nombre de almacenamiento
workspaceartifactstore	Azure Blob Storage	mlfeng1676057577
workspacefilestore	Recurso compartido de archivos de Azure	mlfeng1676057577
workspaceworkingdirectory	Recurso compartido de archivos de Azure	mlfeng1676057577
workspaceblobstore (Predeterminado)	Azure Blob Storage	mlfeng1676057577

Crear almacén de datos

Nombre del almacén de datos:

Tipo de almacén de datos:

Método de selección de cuentas:

- ☒ De suscripción de Azure
- ☐ Especificar manualmente

Id. de suscripción:

Nombre de la tienda:

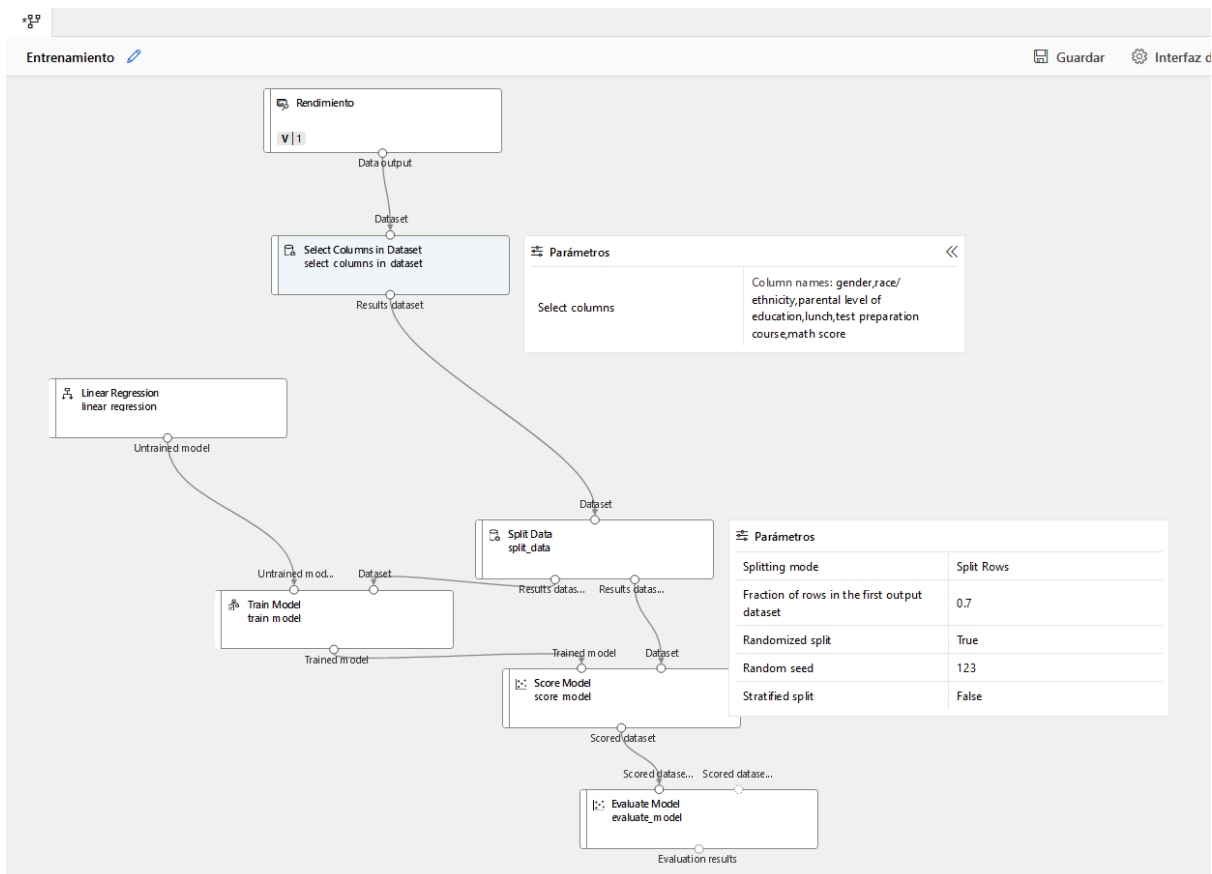
Nombre del sistema de archivos de Azure Data Lake Gen2:

Usar la identidad administrada del área de trabajo para obtener una vista previa y generar perfiles de los datos en Azure Machine Learning Studio

Advertencia: No se pudo comprobar si la identidad administrada del área de trabajo tiene los permisos necesarios en el recurso para habilitar el acceso a los datos. [Más detalles](#)

Nota: Azure Machine Learning Service no valida si el origen de datos subyacente existe o no.

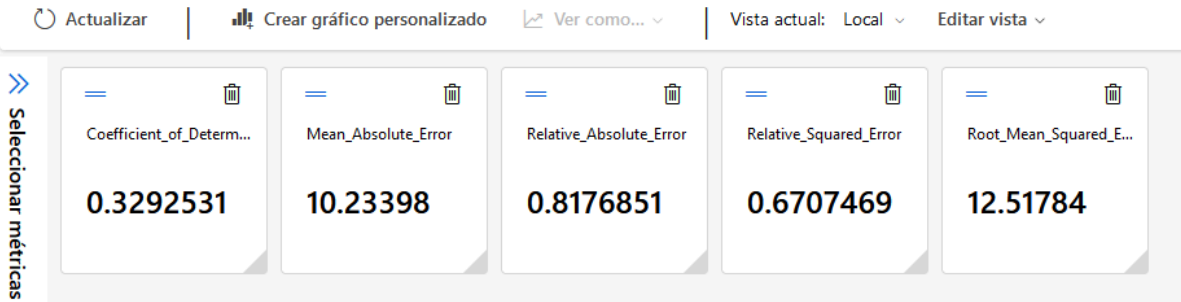
Para el entrenamiento se creó el siguiente pipeline:



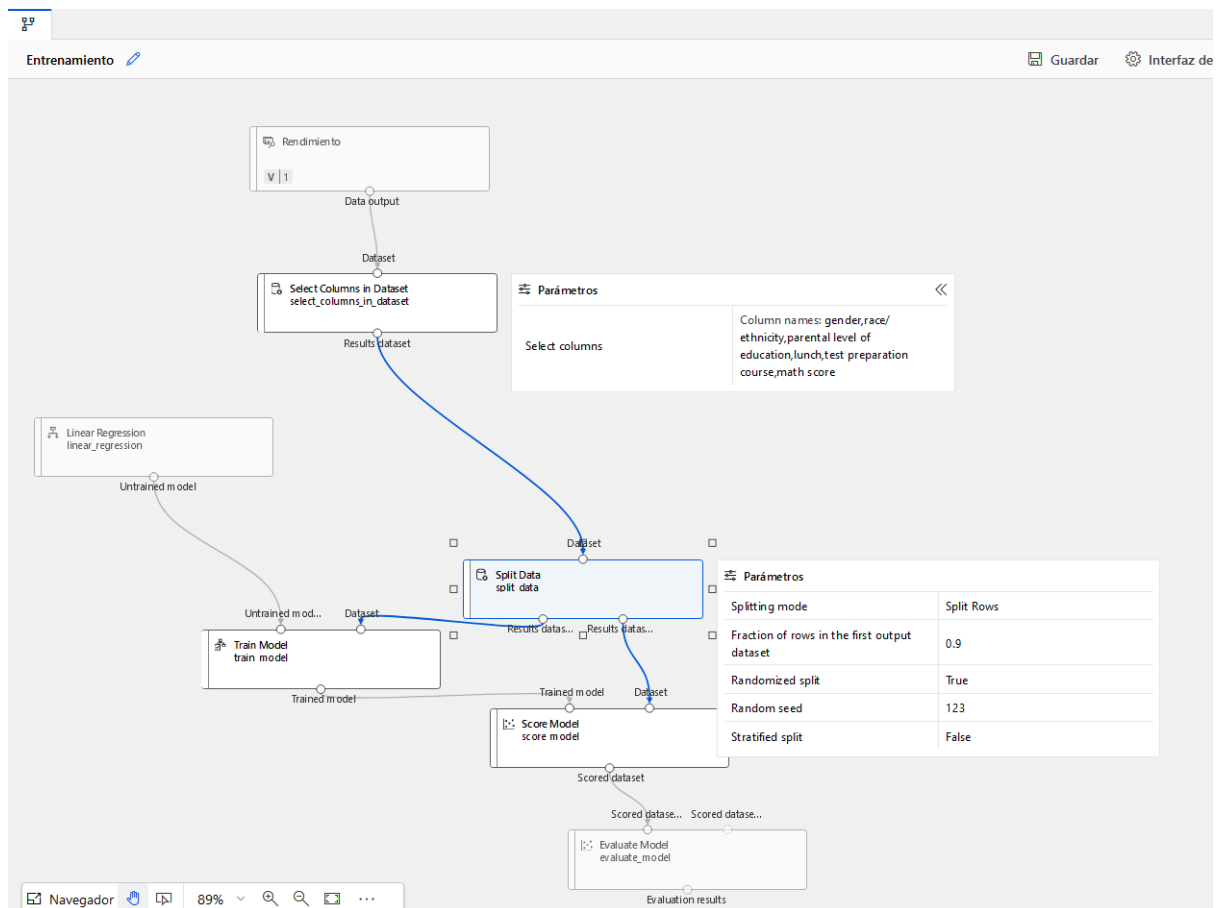
Inicialmente se utilizó únicamente el 70% del dataset. Dando lugar a los siguientes resultados:

Evaluate Model

Información general Parámetros Resultados y registros **Métricas** Trabajos secundarios Imágenes Código ...

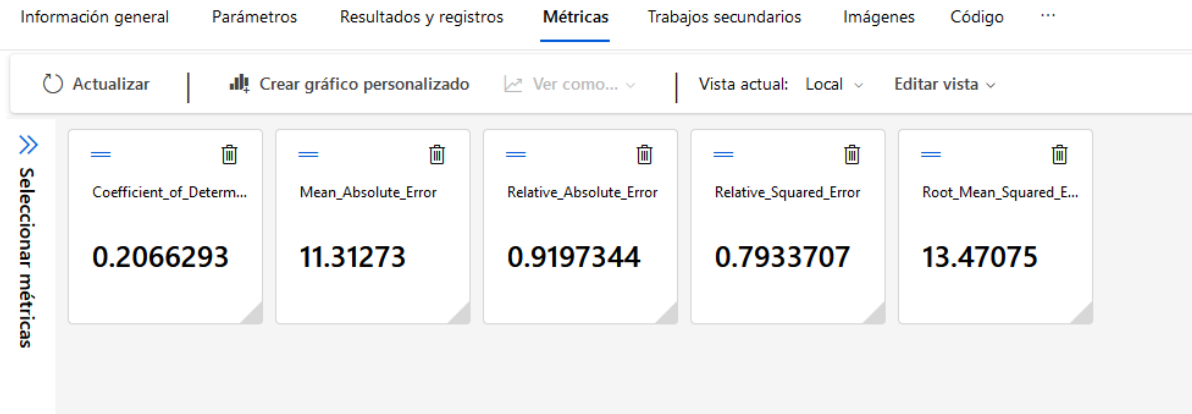


Considerando que los resultados se podían mejorar. Se tomó la decisión de aumentar el dataset de entrenamiento a 0.9:



Es por ello que:

Evaluate Model



Podemos comprobar como aumenta el Error Relativo Absoluto y el Error Cuadrático Medio. Aun así, entendemos que

Monitorización de datos y recursos

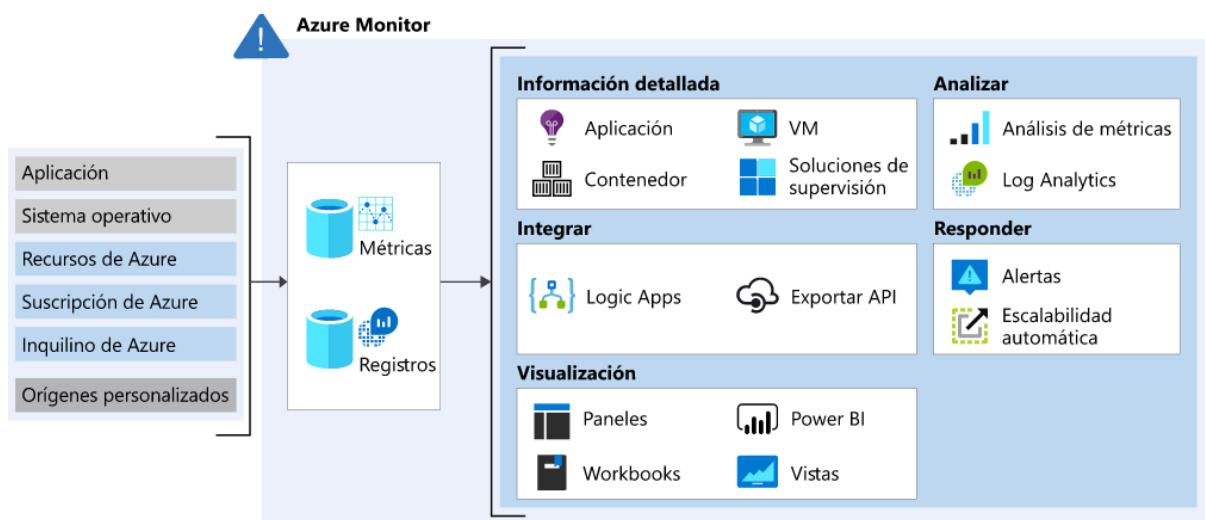
La monitorización de datos y recursos es fundamental para garantizar el rendimiento óptimo de los servicios y aplicaciones desplegados en Azure. Azure Monitor es una herramienta esencial en este proceso, ya que permite recopilar información de supervisión y diagnóstico sobre el estado de los servicios en la nube. A través de Azure Monitor, se pueden obtener métricas y registros detallados que proporcionan información valiosa sobre el funcionamiento de los sistemas y aplicaciones.

Funcionalidades clave de Azure Monitor:

- **Recopilación de datos:** Azure Monitor recopila datos de telemetría, como métricas y registros, de una amplia variedad de recursos en Azure, incluidas las máquinas virtuales, los servicios de aplicaciones, los contenedores y más.
- **Visualización y análisis:** Los datos recopilados por Azure Monitor se pueden visualizar y analizar a través de paneles personalizables y consultas avanzadas. Esto permite identificar tendencias, detectar anomalías y tomar medidas correctivas rápidamente.
- **Alertas y acciones automáticas:** Azure Monitor permite configurar alertas basadas en umbrales predefinidos o reglas personalizadas. Además, se pueden configurar acciones automáticas, como el inicio de procesos de autoscale o la ejecución de scripts de solución de problemas.
- **Integración con otras herramientas:** Azure Monitor se integra con otras herramientas de Azure, como Azure Log Analytics y Azure Application Insights, proporcionando una visión unificada del entorno de Azure y facilitando el análisis de extremo a extremo.

Beneficios de la monitorización con Azure Monitor:

- **Mejora de la fiabilidad:** La monitorización proactiva con Azure Monitor ayuda a identificar y solucionar problemas antes de que afecten a los usuarios finales, lo que mejora la fiabilidad y la disponibilidad de los servicios.
- **Optimización de recursos:** Con Azure Monitor, se pueden identificar recursos subutilizados o mal configurados, lo que permite optimizar la asignación de recursos y reducir costos.
- **Mejora continua:** Mediante el análisis de datos históricos y el seguimiento del rendimiento a lo largo del tiempo, Azure Monitor facilita la identificación de áreas de mejora y la implementación de medidas correctivas para optimizar continuamente los servicios en la nube.



Tipos de Datos Gestionados por Azure Monitor

Registros

- **Descripción:** Los registros contienen información detallada sobre eventos y actividades dentro de los servicios de Azure. Los datos de Azure se organizan en registros con diferentes propiedades según su origen.
- **Contenido:** Los registros pueden incluir datos de texto, como descripciones de eventos, junto con valores numéricos, como métricas de Azure Monitor.

Nota: Son muy útiles para comprender el contexto de los problemas y para la identificación de causas raíz (muchas de las variables no son numéricas).

- **Almacenamiento:** Los datos de registro se pueden almacenar en Workspace de Log Analytics de Azure. Aquí se pueden analizar y consultar la información proporcionando insights sobre eventos y actividades en el entorno de Azure.

Métricas

- **Descripción:** Las métricas describen aspectos específicos de un sistema en un momento dado. Estas métricas se recopilan a intervalos regulares y son útiles para el monitoreo en tiempo casi real y la detección de problemas.

Nota: Al contrario los registros, las métricas son en gran mayoría valores numéricos, por lo que pueden ser comparadas y analizadas para identificar tendencias y anomalías temporales (así como alguna predicción supongo).

- **Contenido:** Las métricas capturan datos como el rendimiento del sistema, la utilización de recursos y otros indicadores clave. (hay que ver más).
- **Almacenamiento:** Las métricas se almacenan en una base de datos de serie temporal en Azure, eficaz para analizar datos con marcas de tiempo, facilitando la detección rápida de problemas y la generación de alertas en caso de que queramos configurar umbrales de aviso.

Kusto, lenguaje de consulta de Azure monitor

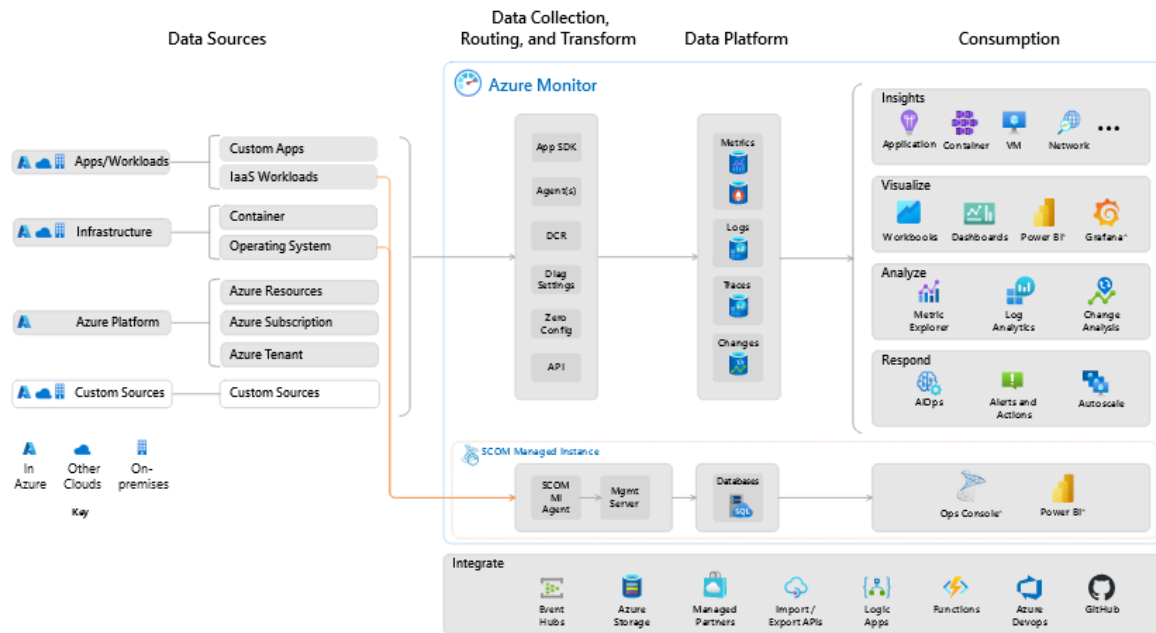
Para recuperar, consolidar y analizar los datos, puede especificar una consulta por medio de kusto mediante consultas de registro en Azure Portal

Nota: Me recuerda bastante a **Dax** pero orientado a consulta y no a manipulación de datos.

En Azure Portal lo relevante es que podemos crear **paneles personalizados**, que será donde están plasmados los recursos y los datos. Cada panel se puede crear mediante un conjunto de iconos, que no son más que artefactos para mostrar datos de recursos y métricas.

Nota: Me refiero como artefactos a cualquier tipo de tabla, report visual o resultado que nos ayude a desglosar información de manera interactiva, no a un Scrum Artifact.

Mediante los paneles de Azure, puede combinar varios tipos de datos, incluidos los registros y las métricas, en un solo panel en Azure Portal. En Azure Portal encontrará la herramienta **Log Analytics**, que sirve para ejecutar las consultas Kusto para Azure Monitor



Desplegar modelo de ML

Evaluate Model

Información general

Parámetros

Resultados y registros

Métricas

Trabajos secundarios

Imágenes

Código

...



Actualizar

Crear gráfico personalizado

Ver como...

Vista actual: Local

Editar vista

Seleccionar métricas



Coefficient_of_Determ...

0.3292531



Mean_Absolute_Error

10.23398



Relative_Absolute_Error

0.8176851



Relative_Squared_Error

0.6707469



Root_Mean_Squared_E...

12.51784

Registrar modelo desde una salida de trabajo

✓ Seleccionar trabajo

2 Seleccionar salida

✗ Configuración del modelo

4 Revisar

Seleccionar salida

Especifique la salida correspondiente para registrar el modelo.

Tipo de modelo *

Tipo no especificado

Salida del trabajo *

module_statistics

1 archivo seleccionado

Nombre de archivo

{ } module_statistics/error_info.json

Registrar modelo desde una salida de trabajo

✓ Seleccionar trabajo

✓ Seleccionar salida

✓ Configuración del modelo

4 Revisar

Revisar

Revise o realice cambios en las selecciones.

Seleccionar trabajo

Trabajo
train_model

Seleccionar salida

Tipo de modelo
Tipo no especificado
Salida del trabajo
module_statistics

Archivos
1 archivo seleccionado
{ } module_statistics/error_info.json

Configuración del modelo

Nombre
Students_perform_model

Descripción
--

Versión
--

Etiquetas

① Sin etiquetas