Sistemas de Big Data

Elena Tejadillos

Analista de Datos

Elena.Tejadillos@iesabastos.com

Cloud

Qué veremos hoy

- Cloud Computing
- Azure
- Laboratorios

Cloud Computing

”You don't generate your own electricity. Why generate your own computing?”
Jeff Bezos, CEO, Amazon, 2008

Cloud Computing

La Computación en Nube es un modelo para permitir el **acceso ubicuo**, conveniente y **bajo demanda** mediante red a un conjunto compartido de **recursos de cómputo configurables** (i.e., redes, servidores, almacenamiento, aplicaciones y servicios) que pueden ser **rápidamente aprovisionados y liberados** con **mínimo esfuerzo** de gestión o interacción con el proveedor del servicio.
National Institute of Standards and Technology (NIST)

Cloud Computing (Computación en Nube) es un paradigma que permite ofrecer servicios (cómputo, almacenamiento, etc.) a través de Internet.

- Computación / almacenamiento ofrecido como servicio por parte de un proveedor a múltiples clientes.
- Aprovechamiento de las economías de escala de grandes proveedores para ofrecer ahorro de costes a los usuarios.
- Pago por uso, sin inversiones iniciales.

Tipos de recursos a aprovisionar

- Cómputo
  Máquinas virtuales con unas determinadas características (CPU, RAM, disco, etc.) y configuración (SO).
- Almacenamiento
  Permite almacenar y acceder a datos guardados en el Cloud, generalmente en forma de ficheros

Ejemplo

Una start-up desarrolla un sistema de Data Analytics para un cierto sector y se ofrece como aplicación en la nube. Requiere capacidad de cómputo y almacenamiento de datos para su puesta en producción.

- **Opción A**: Adquirir servidores y realizar el housing y el hosting de la aplicación en la propia infraestructura de la empresa (In-House).
- **Opción B**: Aprovisionar los recursos necesarios de un proveedor Cloud.

Opción A: In-House

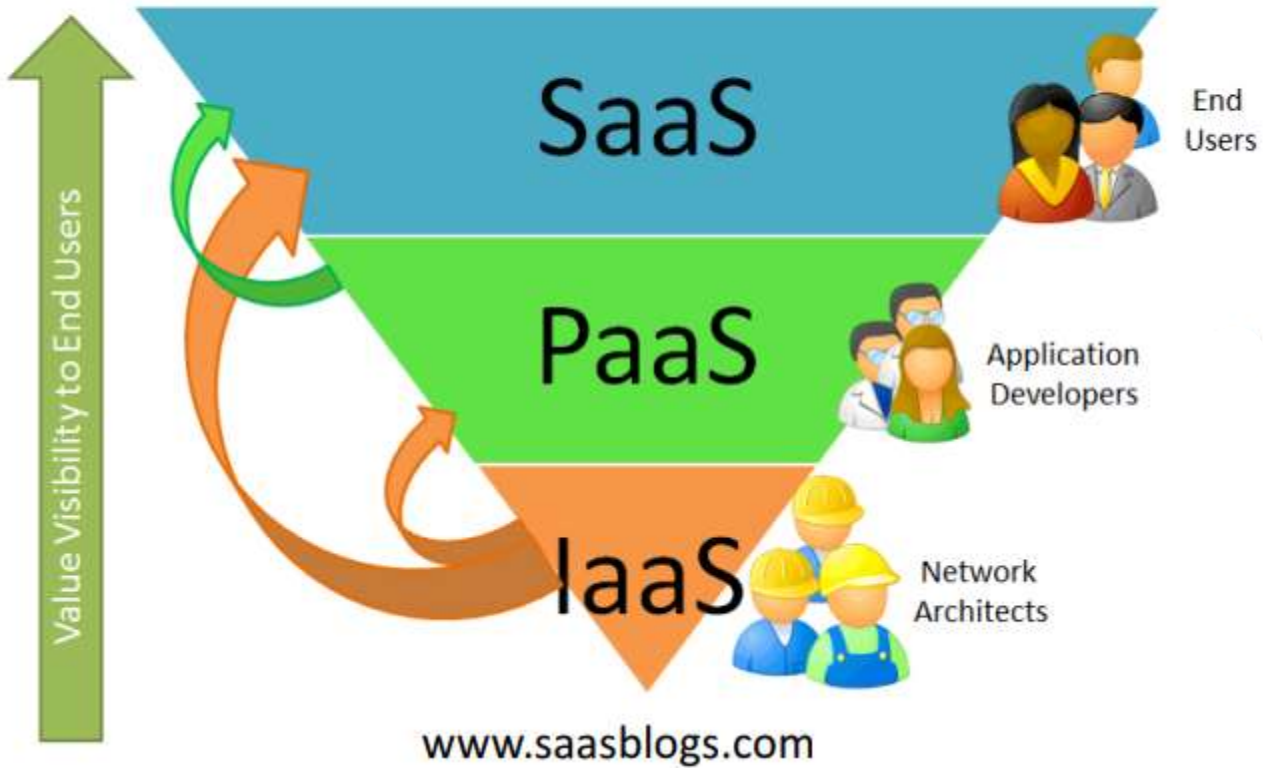
- Alquilar y acondicionar un local (refrigeración, cableado,SAIs, etc.).
- Dimensionar y adquirir hardware para cómputo y almacenamiento, actualizarlo periódicamente.
- Configurar los recursos, actualizarlos.

Opción B: Cloud computing

- Aprovisionar los recursos de cómputo iniciales necesarios para la puesta en producción.
- Configurar la aplicación para que autoaprovisione y libere dinámicamente nuevos recursos (cómputo, almacenamiento) dependiendo de la carga de trabajo / usuarios de la misma.

Modelos de servicio

- SaaS vs PaaS vs IaaS



IaaS: Infractructure as a Service

Infrastructure as a Service (IaaS) -> Sysadmin

- El proveedor ofrece el acceso inmediato a recursos de cómputo (máquinas virtuales) y de almacenamiento bajo demanda y, típicamente, mediante un modelo de pago por uso.
- El usuario debe configurar las máquinas virtuales para configurar la infraestructura.
  - Ejemplos: Amazon EC2, GoGrid, Rackspace, ElasticHosts, etc.
  - Acceso root a los recursos.

PaaS: Platform as a Service

Platform as a Service (PaaS)-> Desarrollador

- El proveedor proporciona al desarrollador un conjunto de herramientas (APIs) de plataforma para el desarrollo de aplicaciones que se ejecutan sobre la plataforma del proveedor.
- El usuario no gestiona el hardware ni el SO subyacente.
- Ejemplos: Google App Engine, Microsoft Azure*, entre otros.

SaaS: Software as a Service

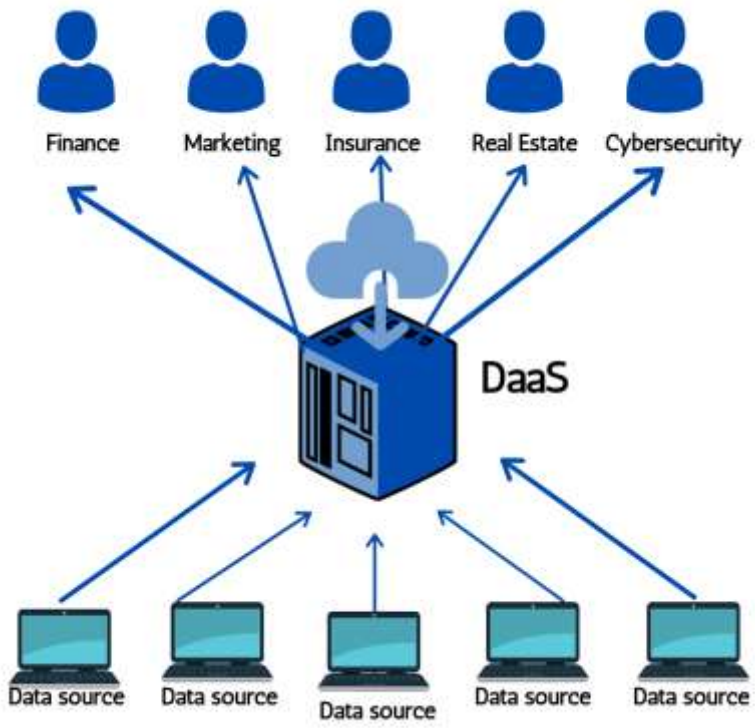
Software as a Service (SaaS) -> Usuario

- El usuario accede a las aplicaciones, típicamente con un navegador web, que se ejecutan en la infraestructura (Cloud) de un proveedor externo.
- Para el usuario, las aplicaciones y sus datos son ubicuas y siempre disponibles.

XaaS: Anything as a Service

DaaS: Data as a Service

- Data Hub
- Data Exchange



Data Centers

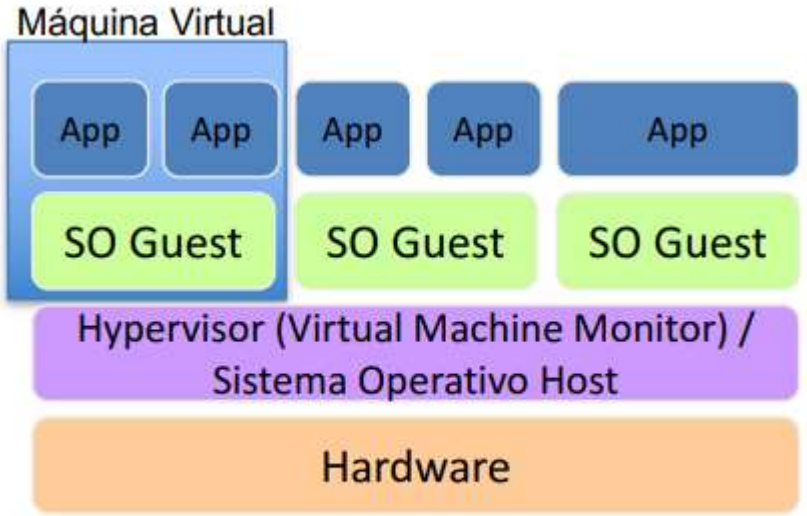
Google, Amazon, Microsoft, etc. tienen grandes centros de datos por todo el mundo.
Centros de Datos de Google [Localizaciones]
http://www.google.com/about/datacenters

Microsoft data center de Dublin, con más de 60.000 m^2 construidos.

Virtualización

La Virtualización permite crear un (o varios) entorno simulado (Máquina Virtual, MV) que ejecuta un SO guest (invitado). Todo ello, corriendo sobre un SO host (anfitrión) con ayuda de un hipervisor (o Virtual Machine Monitor).

Plataforma tradicional



Plataforma Virtualizada

Modelos de despliegue

- **Cloud Privado** (Private cloud, on-premises Cloud).
  - Infraestructura de uso exclusivo para una institución.
- **Cloud de Comunidad** (Community cloud).
  - Federación de recursos de diferentes organizaciones.
- **Cloud Público** (Public cloud).
  - Disponible para todo el mundo mediante pago por uso.
- **Cloud Híbrido** (Hybrid cloud).
  - Combinación de Cloud privado (o recursos on-premise) con extensión a Cloud público.

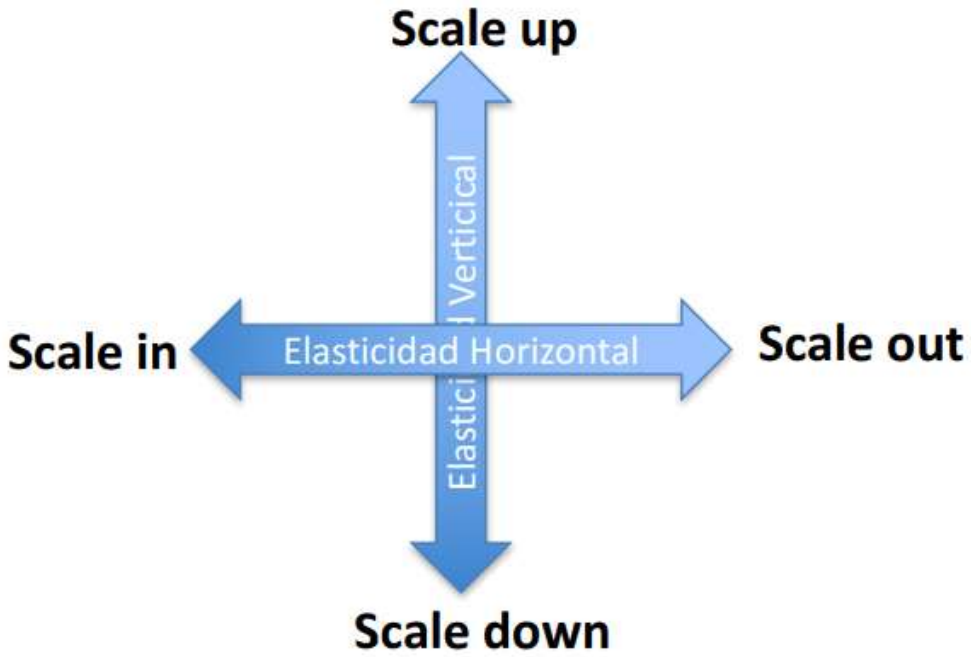
Cloud Privado vs Cloud Publico

- El Cloud privado da soporte a una organización con sus recursos.
- El Cloud público ofrece un servicio de cómputo mediante pago por uso.
  - Amazon Web Services
  - Azure
  - Google Cloud
  - …

  Ventajas del Cloud público:
  – *Nula inversión en infraestructura*
  – *Infraestructura just-in-time*: Sin sobredimensionar o malgastar recursos
  – *Utilización de recursos más eficiente:* Adaptando la infraestructura virtual al consumo de recursos.
  – *Coste basado en el uso*: Mejoras en la aplicación pueden reducir el coste de la factura mensual.
  – *Ilusión de capacidad infinita*: Elasticidad en el aprovisionamiento de recursos bajo demanda.

Elasticidad

La elasticidad es una de las características diferenciadora de una plataforma Cloud.



Elasticidad

- Escalado Vertical (Scale Up/Scale Down)

- Escalado Horizontal (*Scale Out/Scale In*)

Tolerancia a Fallos

La tolerancia a fallos debe conseguirse mediante:

- **Replicación**

– Servicios que toman el control de la carga de trabajo de los servicios fallidos.

- **Distribución**

– Servicios geográficamente distribuidos para reducir la exposición ante fallos en una región.

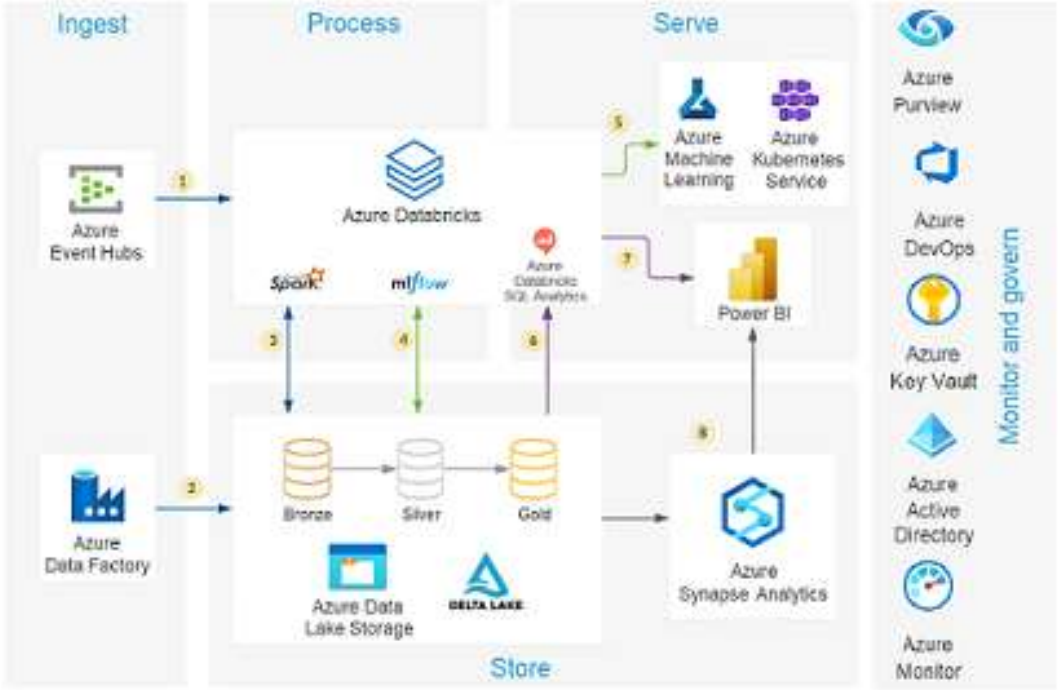
Opiniones críticas sobre el cloud

Se pierde el control sobre los datos en el Cloud
– Los datos se transfieren a recursos con un nivel de confianza poco claro.
– El acceso, modificación o manipulación encubierta de los datos almacenados es potencialmente posible.
– La operación y acceso a las máquinas virtuales está totalmente bajo el control del proveedor.

Aspectos éticos y legales dudosos
Ejemplo: Una empresa con sede en EEUU está obligada a proporcionar al gobierno datos de sus clientes si hay sospecha de delito grave. • Aunque los datos estén en Europa, no hay garantías de que el gobierno no pudiera exigir dicha información.

Data Lake

- Un Data Lake es un repositorio de almacenamiento que contiene una gran cantidad de datos en su formato nativo y sin procesar.
- Los Data Lakes están optimizados para escalar a terabytes y petabytes de datos. Los datos provienen típicamente de múltiples orígenes heterogéneos y pueden ser estructurados, semiestructurados o no estructurados.
- La idea con los Data Lakes es almacenar todo en su estado original, no transformado



Ventajas de los Data Lakes

- Los datos nunca se desechan, porque se almacenan en su formato sin procesar. Esto es especialmente útil en un entorno de macrodatos, cuando es posible que no se conozca de antemano qué información está disponible a partir de los datos.

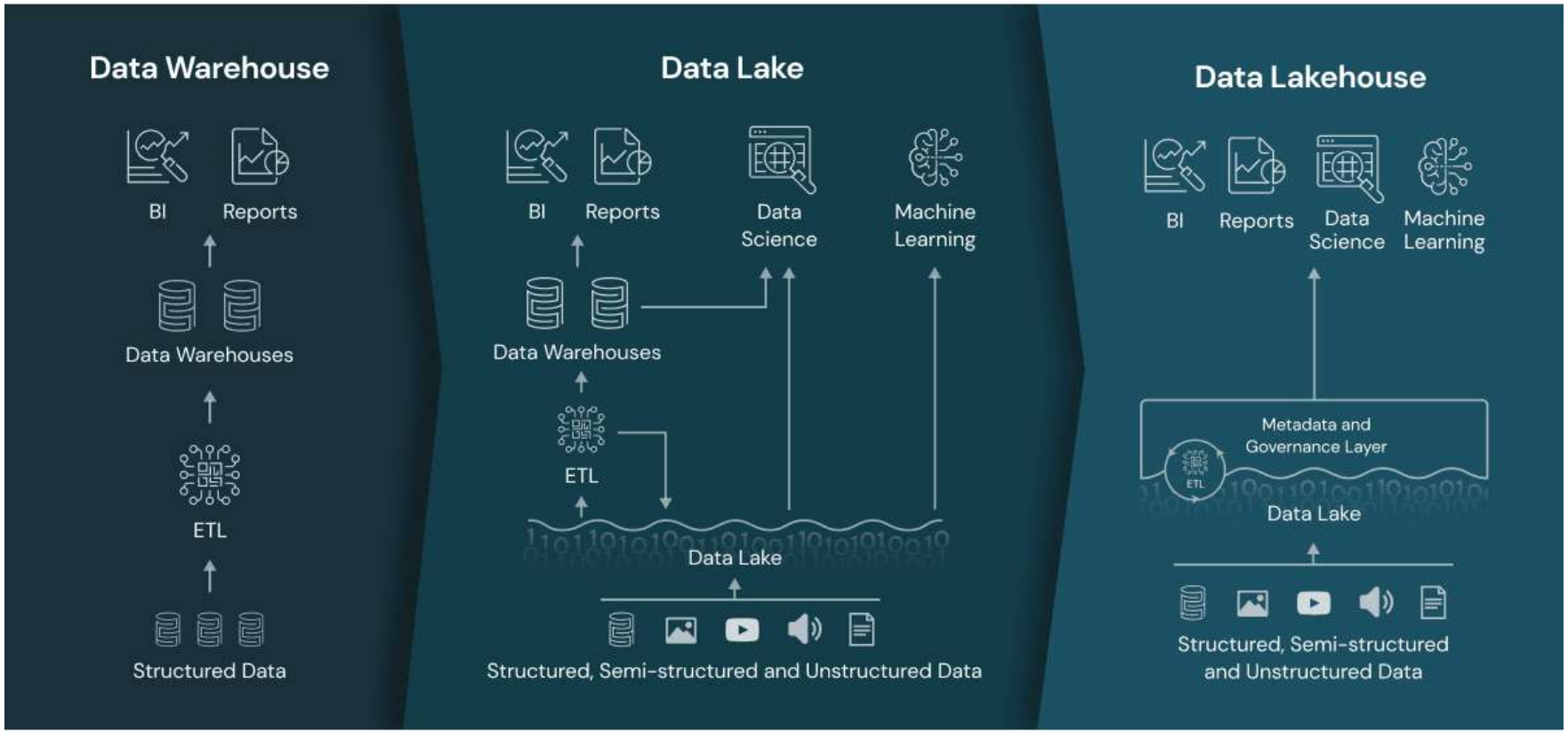
- Los usuarios pueden explorar los datos y crear sus propias consultas.

- Puede ser más rápido que las herramientas de extracción, transformación y carga de datos tradicionales.

- Es más flexible que un almacenamiento de datos, porque puede almacenar datos no estructurados y semiestructurados

Data Lakehouse

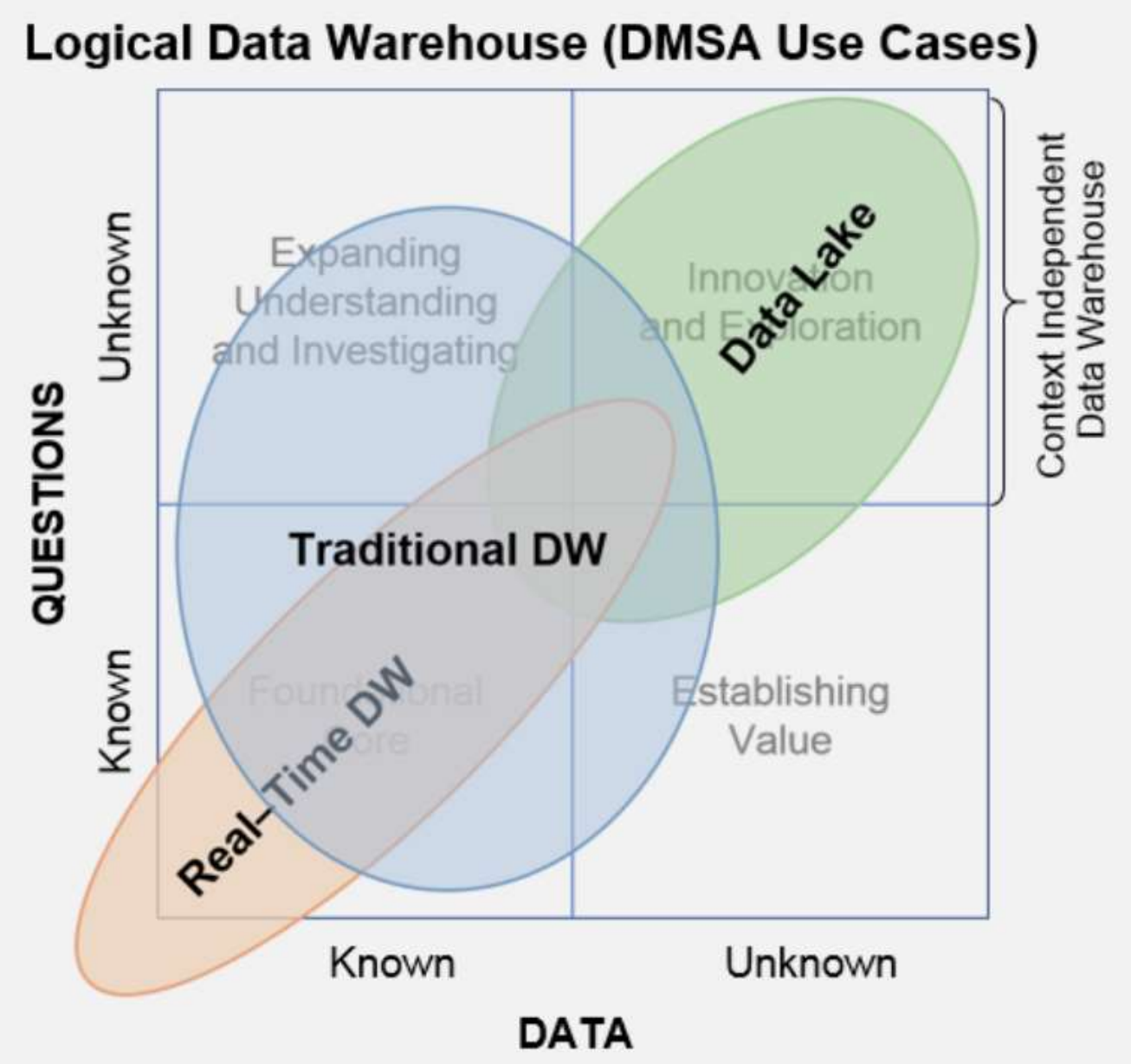
Un Data Lakehouse es una arquitectura abierta de gestion de datos que combina la flexibilidad, el coste eficiente y la escalabilidad de los DataLakes con la gestion de datos y ACID transaccional de los Data Warehouse, proporcionando la union de Business Intelligence (BI) y la Inteligencia Artificial (AI)



Data Warehouse vs Data Lake

DATA WAREHOUSE	vs.	DATA LAKE
structured, processed	DATA	structured / semi-structured / unstructured, raw
schema-on-write	PROCESSING	schema-on-read
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals	USERS	data scientists et. al.

Data Warehouse vs Data Lake



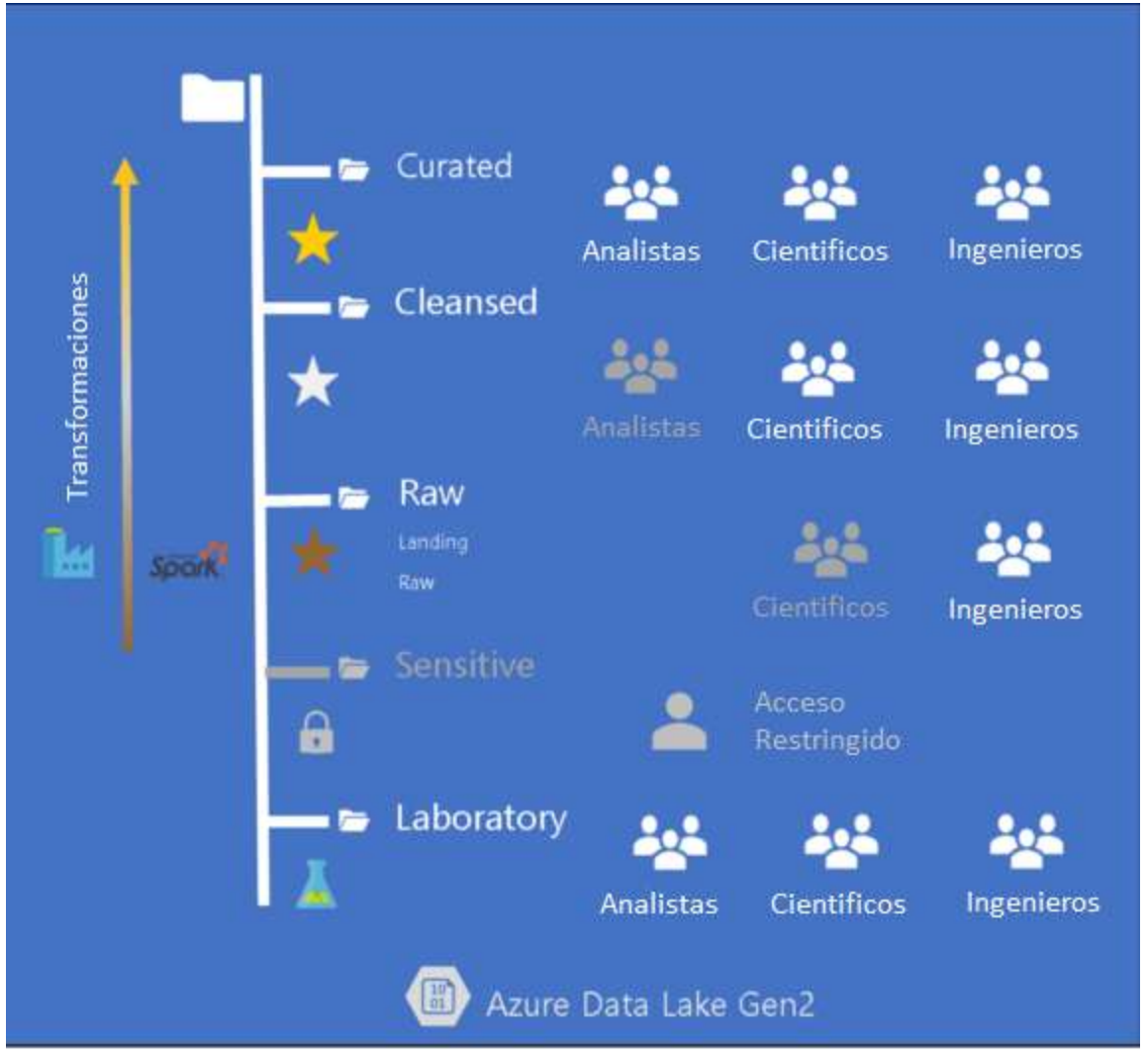
Data Lake, qué planificar

Hay varios temas que debe considerar al planificar un data lake:

- Necesidades de ingestión (push / pull, streaming o batch)
- Seguridad sobre el acceso a los datos
- Políticas de retención y archivo de datos
- Requisitos de encriptación
- Gobernanza
- Calidad de los datos
- Gestión de datos maestros
- Verificaciones de validez necesarias
- Gestión de metadatos
- Organización de datos para una recuperación de datos óptima
- Programación y gestión de jobs
- Registro y auditoría
- Necesidades de enriquecimiento, estandarización, limpieza y curación
- Enfoque modular del diseño general

Data Lake - Estructura

Raw / Bronze
Esta capa es un depósito que almacena datos en su estado original, sin filtrar ni limpiar.
Cleansed / Silver
Es una zona de filtración que elimina las impurezas, pero también puede implicar un enriquecimiento.
Curated / Gold
Esta capa está optimizada para análisis. Puede almacenar datos en data marts desnormalizados o esquemas en estrella
Laboratory
Capa donde ocurre la exploración y la experimentación. Los científicos y analistas de datos son libres de crear prototipos e innovar, combinando sus propios conjuntos de datos.



Data Lake: Nomenclatura de carpetas

- Una convención de nomenclatura legible por humanos, comprensible, coherente y autodocumentada
- Permisos suficientemente granulares, pero no a una profundidad que genere gastos y administración adicionales.
- Estrategias de particionamiento que pueden optimizar los patrones de acceso y los tamaños de archivo apropiados.
- Cada carpeta con archivos del mismo esquema y el mismo formato / tipo

Ejemplos.
\Raw\DataSource\Entity\YYYY\MM\DD\File.extension
Tediosa
\Raw\YYYY\MM\DD\DataSource\Entity\File.extension
Por seguridad
\Raw\General\DataSource\Entity\YYYY\MM\DD\File.extension
\Raw\Sensitive\ DataSource\Entity\YYYY\MM\DD\File.extension

Data Lake: Nomenclatura de carpetas

¿Cuántos data lakes, cuentas de almacenamiento y sistemas de archivos son necesarios?

El DL en sí mismo puede considerarse una entidad lógica única, pero podría estar formado por múltiples cuentas de almacenamiento en diferentes suscripciones en diferentes regiones, con administración y gobierno centralizados o descentralizados.

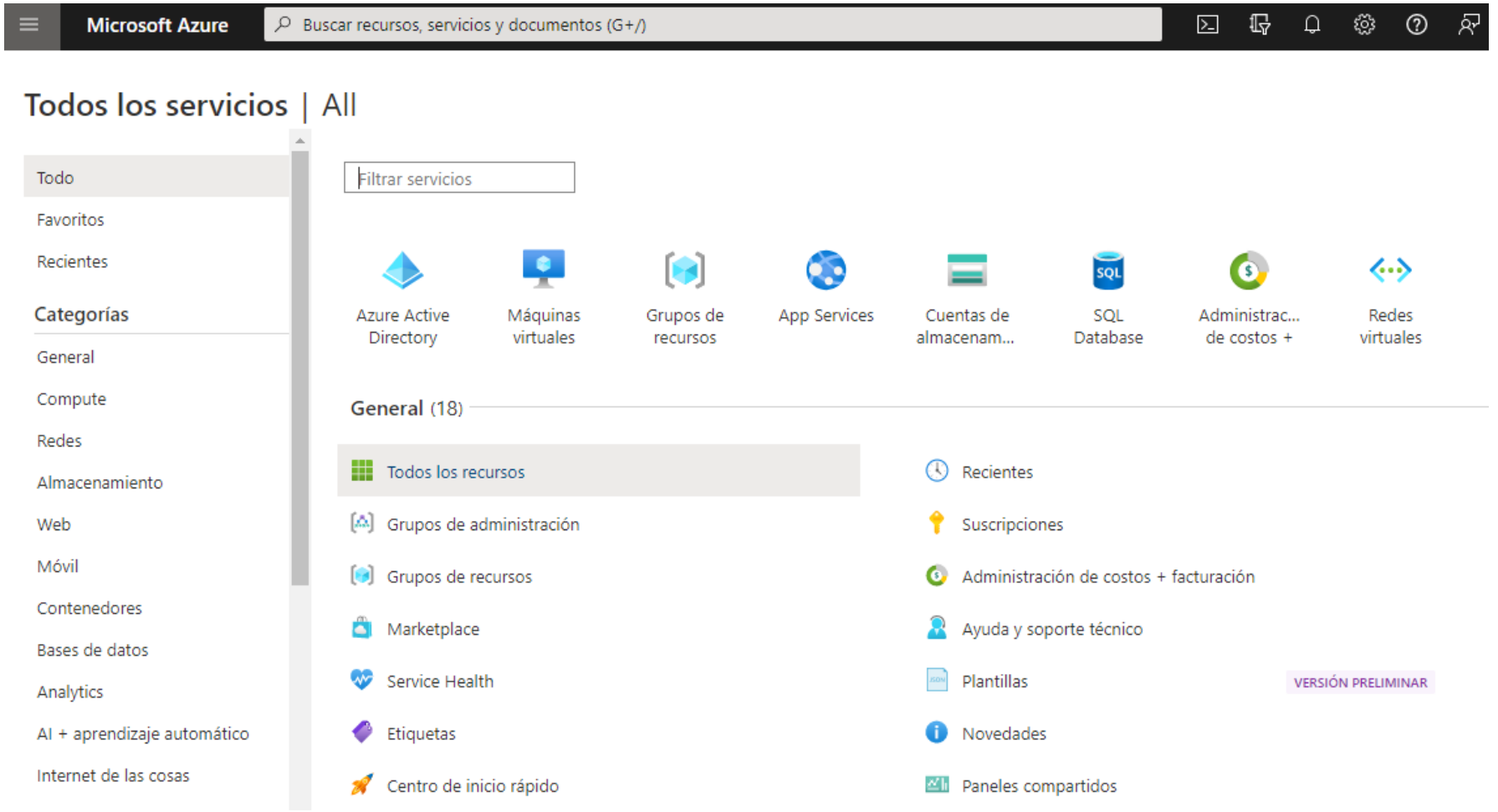
- La planificación de cargas de trabajo a gran escala
- DL regionales vs globales: Las empresas globales pueden tener varios DL regionales pero necesitan obtener una visión global de sus operaciones. Un DL centralizado podría recopilar y almacenar datos agregados regionalmente para ejecutar análisis y pronósticos en toda la empresa.
- Facturación y motivos organizativos.
- Aislamiento y previsibilidad del entorno.
- Características y funcionalidad a nivel de cuenta de almacenamiento.

Data Lake: Tipos de ficheros

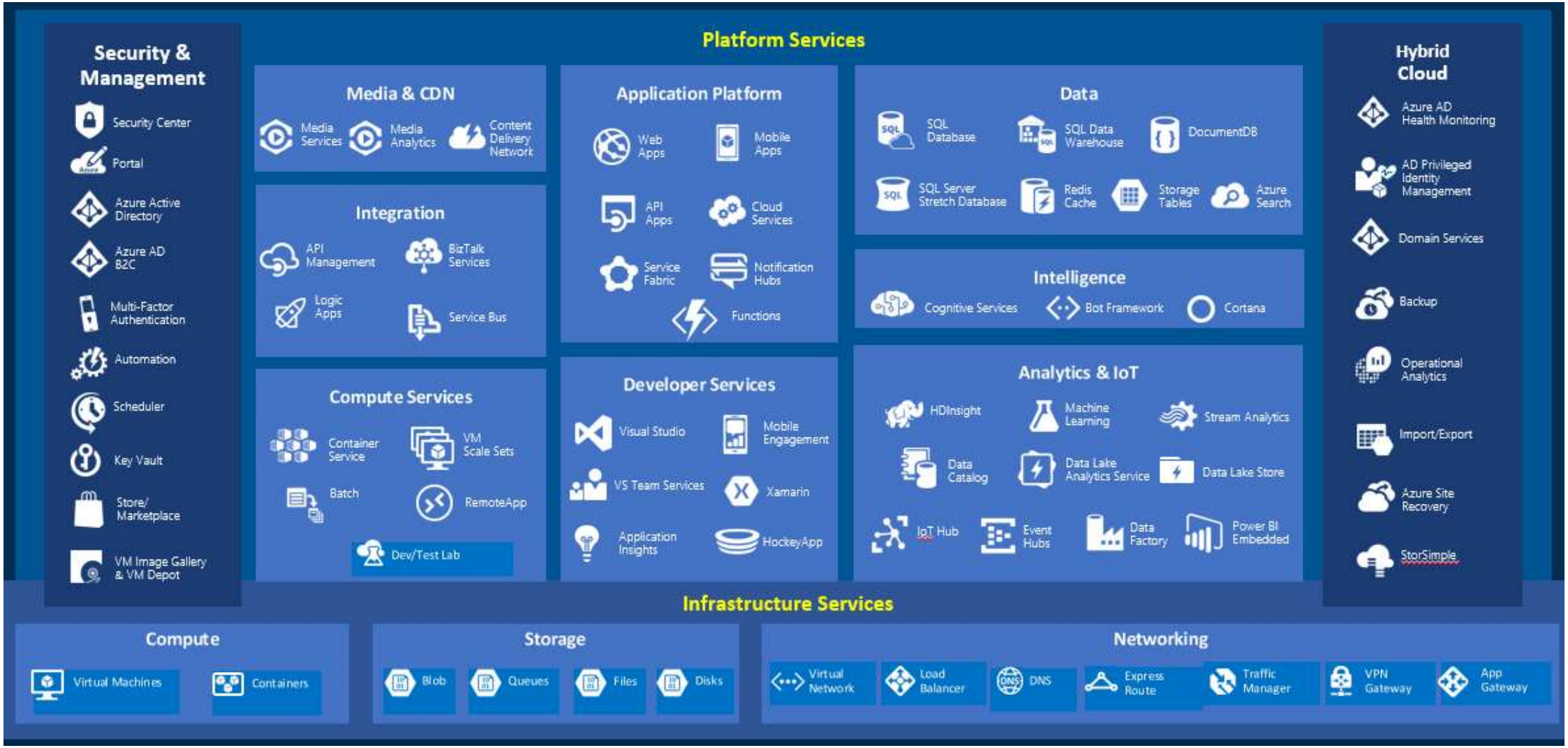
- Considere la posibilidad de escribir archivos en lotes y utilice formatos con una buena relación de compresión, como **Parquet**, o utilice un formato optimizado para escritura como **Avro**.
- Introduzca una **zona / capa de intermedia**, que periódicamente toma archivos sin comprimir y / o pequeños de raw y los compacta en archivos comprimidos más grandes en esta nueva capa.
- Utilice la **gestión del ciclo de vida para archivar datos raw** y reducir los costos de almacenamiento a largo plazo sin tener que eliminar datos.

Azure

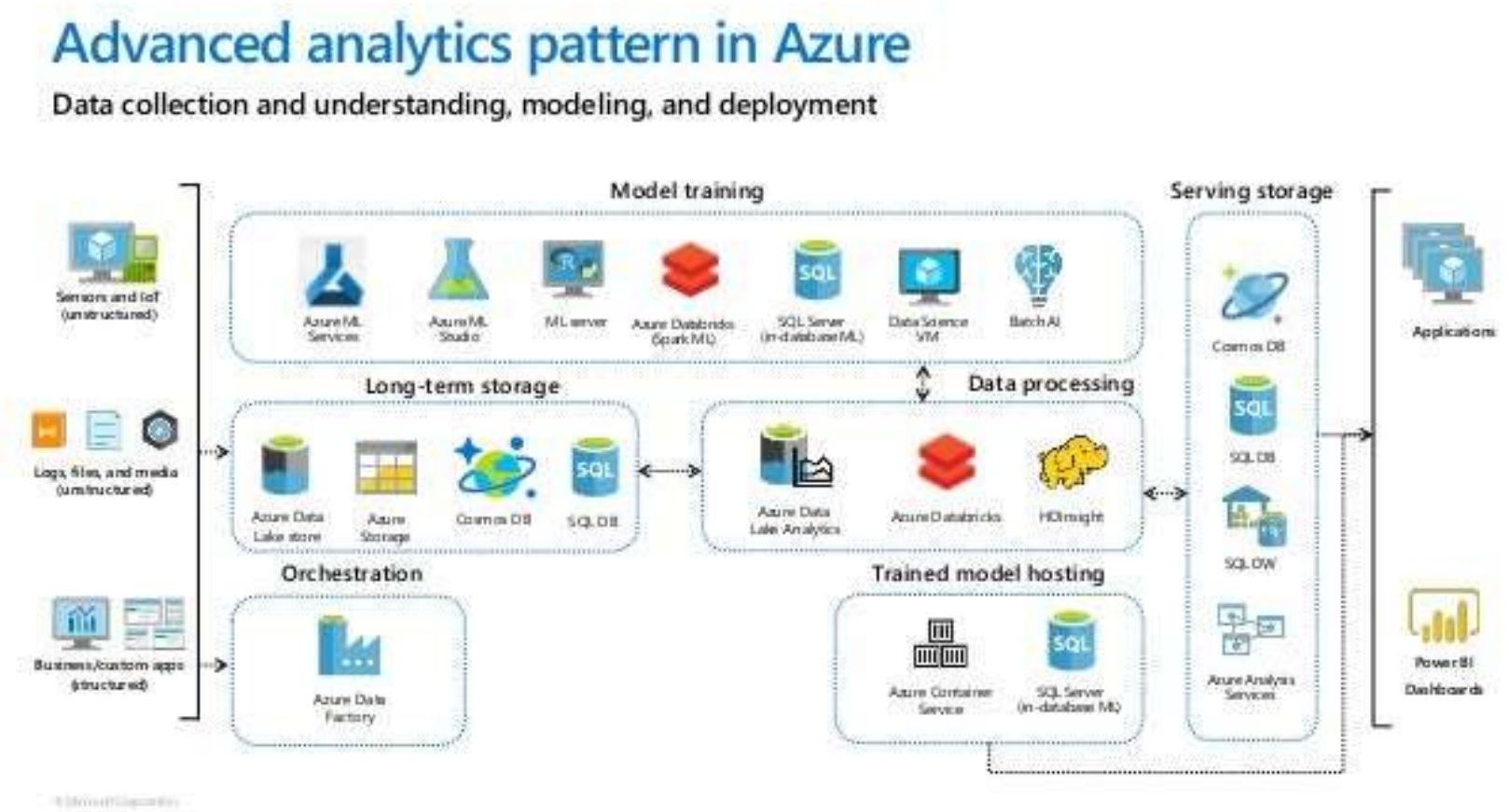
Portal de Azure



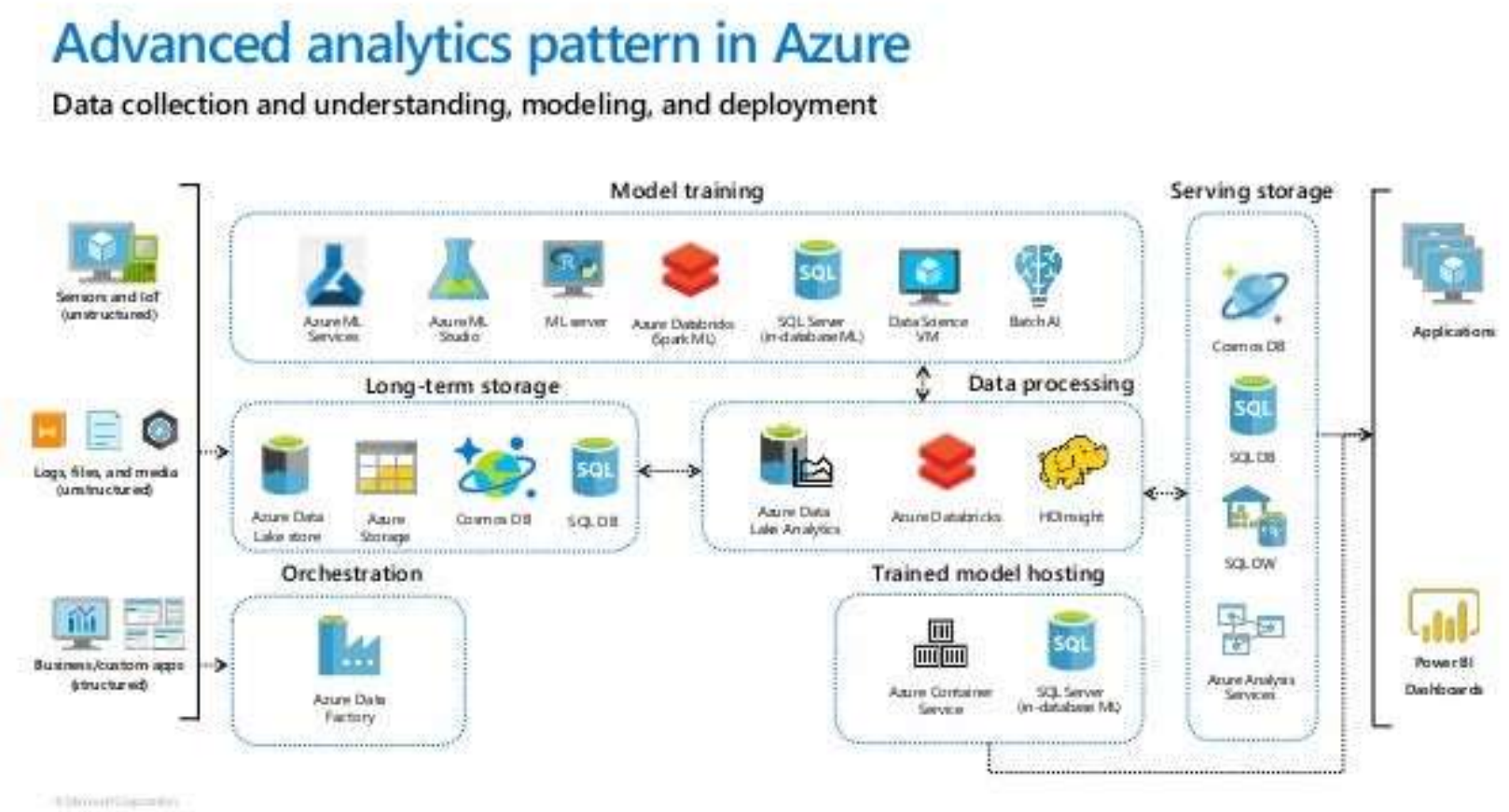
Tipos de Servicios de Azure



Data/Analytics services



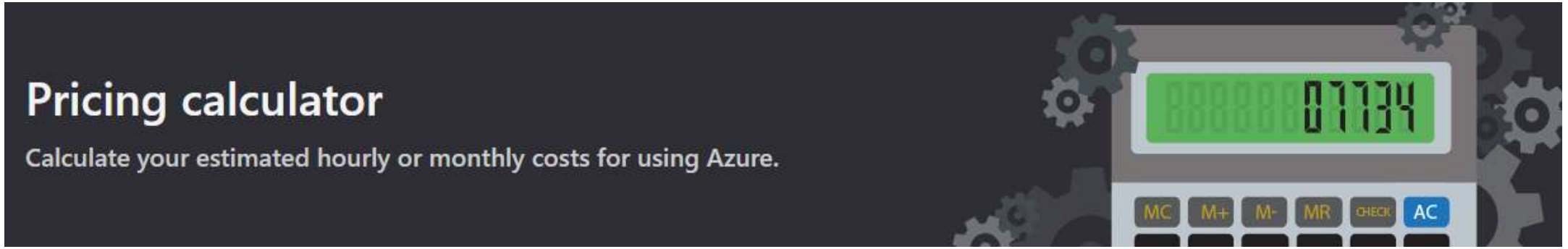
Data/Analytics services



Azure Pricing

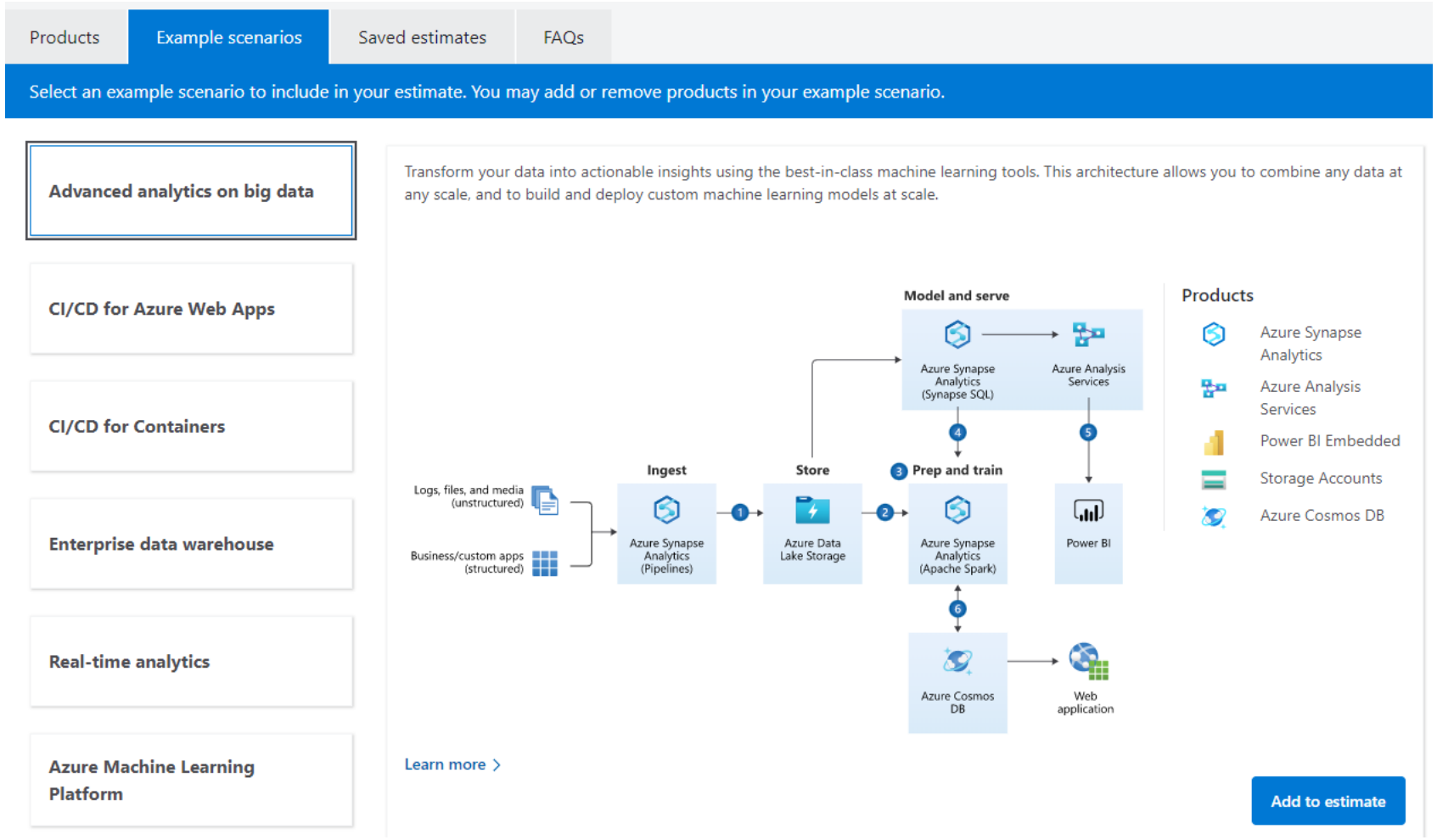
Azure Pricing

[Pricing Calculator | Microsoft Azure]



Azure Pricing

[Pricing Calculator | Microsoft Azure]

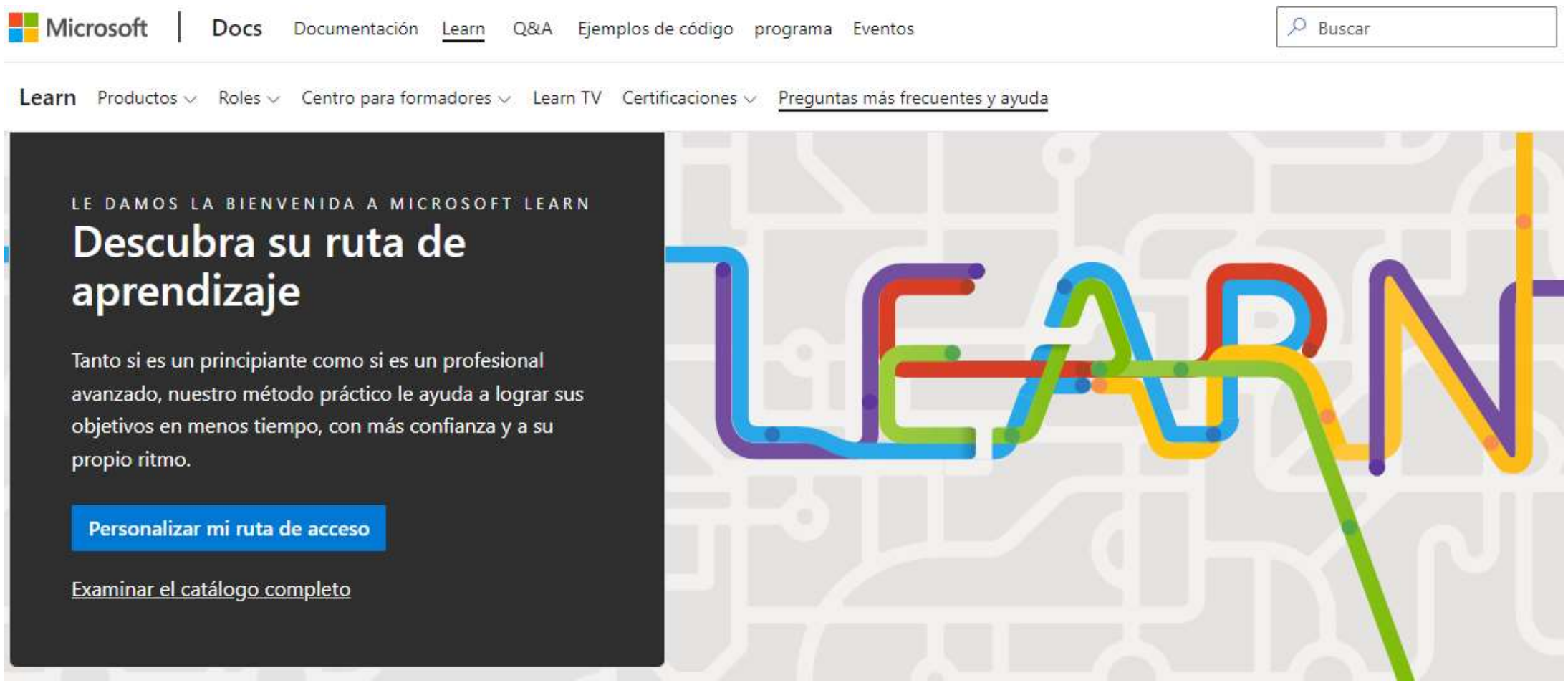


Labs

Microsoft Learn

Microsoft Learn

https://docs.microsoft.com/es-es/learn/



Labs

Crear cuenta de laboratorio

https://docs.microsoft.com/es-es/azure/lab-services/tutorial-setup-lab-account

Empiece con una prueba gratuita de Azure

Obtenga un crédito gratuito de 200 USD para los productos y servicios de Azure, más 12 meses de servicios gratuitos populares.

Inicio Más información

Azure storage

Azure Storage

https://docs.microsoft.com/es-es/learn/modules/create-azure-storage-account/

Creación de una cuenta de Azure Storage

30 min. • Módulo • 7 Unidades

Azure Data Factory

Azure Data Factory

https://docs.microsoft.com/es-es/learn/modules/intro-to-azure-data-factory/

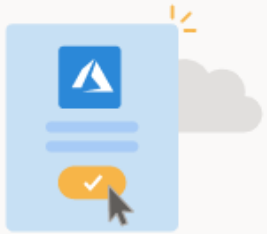
Azure Data Factory es un servicio de integración de datos y ETL basado en la nube que permite crear flujos de trabajo controlados por datos para lo siguiente:

- Organizar el movimiento de datos.
- Transformar los datos a escala.

Azure Labs

https://labs.azure.com/

1 Obtención de una suscripción a Azure



Use su cuenta personal de Microsoft (MSA) o una cuenta profesional o educativa para crear una suscripción de Azure.

Crear una cuenta de Azure gratis

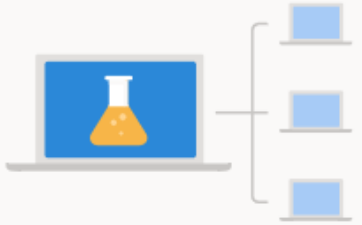
2 Creación de un plan de laboratorio (versión preliminar) o una cuenta de laboratorio

En el Azure Portal, use su suscripción de Azure para crear un plan de laboratorio (versión preliminar) o una cuenta de laboratorio para configurar y administrar sus laboratorios.

Vea aquí un tutorial detallado.

Configuración de un plan de laboratorio (versión preliminar) o una cuenta de laboratorio

3 Creación del primer laboratorio



Cree un laboratorio para su clase, formación, hackathon y mucho más.

Iniciar sesión para crear un laboratorio