PySpark Pipelines Develop

BigData - UBO2022

Sebastian Ulloa Quezada

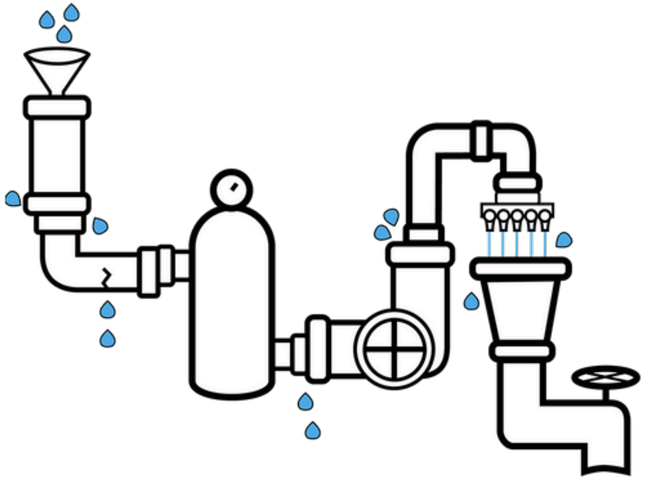
Que es un pipeline?

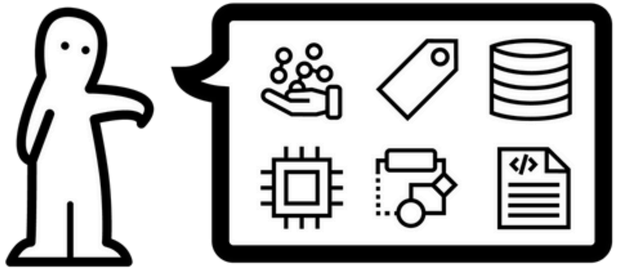
Una pipeline de datos es una construcción lógica que representa un proceso dividido en fases. Las pipelines de datos se caracterizan por definir el conjunto de pasos o fases y las tecnologías involucradas en un proceso de movimiento o procesamiento de datos.

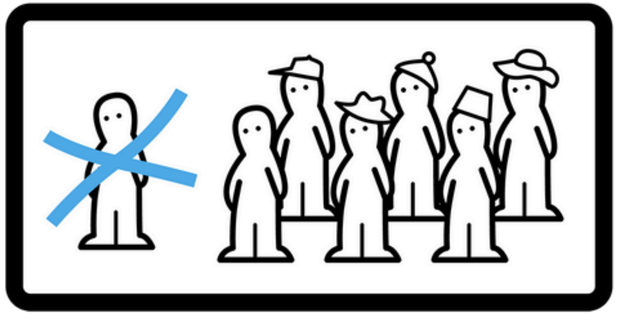
Las pipelines de datos son necesarias ya que no debemos analizar los datos en los mismos sistemas donde se crean. El proceso de analítica es costoso computacionalmente, por lo que se separa para evitar perjudicar el rendimiento del servicio.

Los movimientos de datos entre estos sistemas forman pipelines de datos y son procesos que no debemos obviar, y que involucran varias fases, estrategias y metodologías.

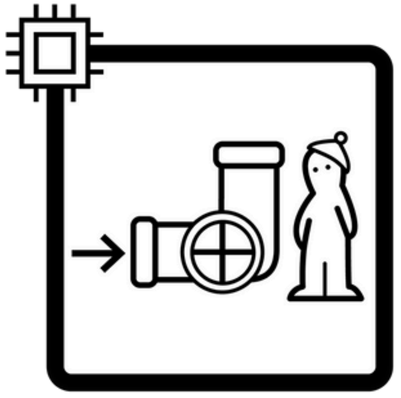
DATA PIPELINE



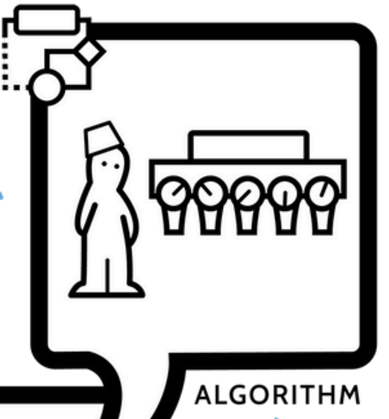




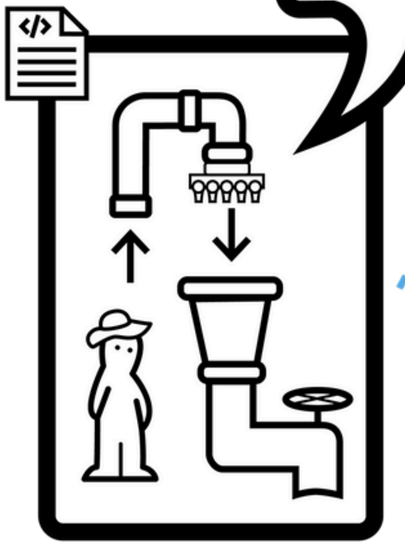
DATA COLLECTION



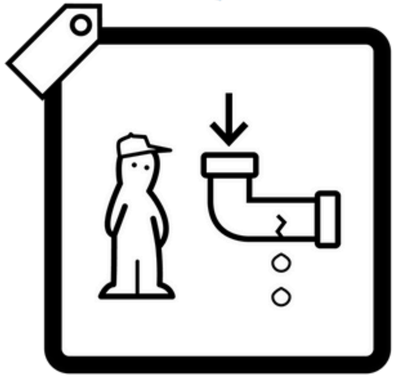
PROCESSING SUBSTRATE



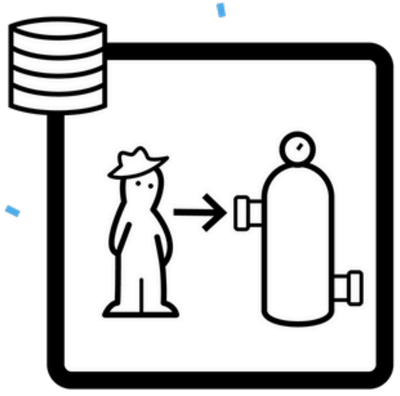
ALGORITHM



CODE



METADATA & CLEANING

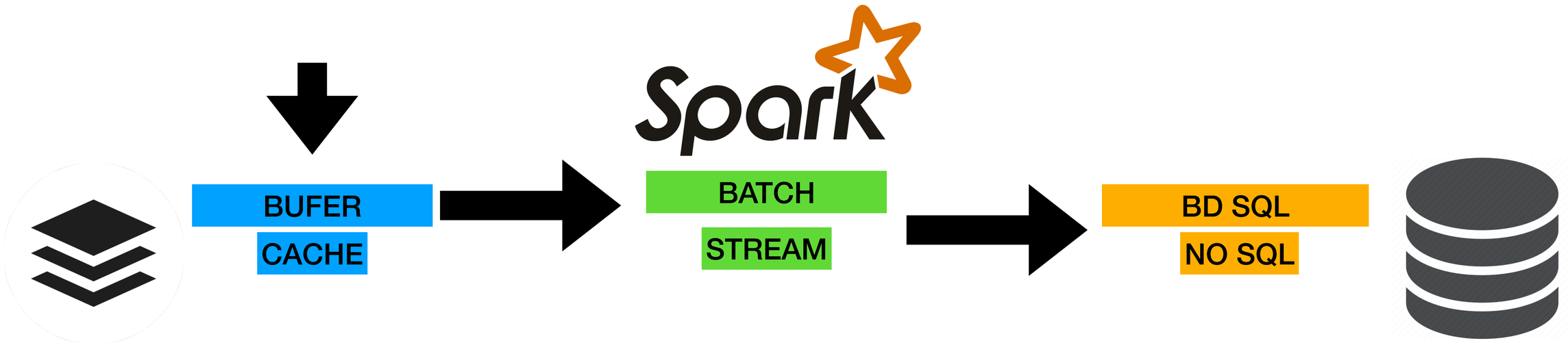


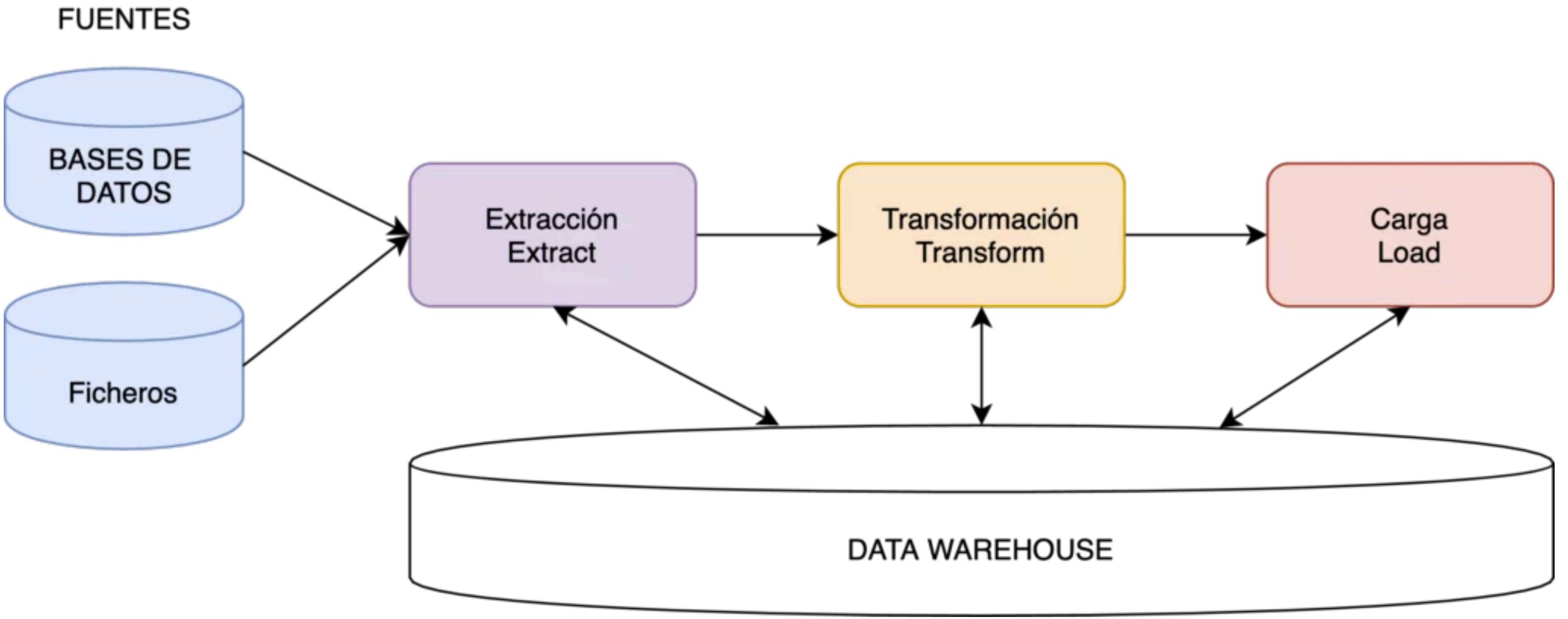
STORAGE & RETRIEVAL ARCHITECTURE

OUTPUT

Ejemplo

- Como ejemplo, podemos pensar en las APIs de ingesta para obtener los datos.Esta data es almacenada en un buffer.
- Después, una tecnología de procesamiento, que puede ser streaming o batch, leerá los datos de nuestro buffer. Por ejemplo, Apache Spark realizará analítica sobre estos datos.
- Por último, la pipeline termina con el resultado almacenado de forma persistente en una base de datos.
- Una vez que nuestros datos están persistidos se encuentran listos para ser usados. Podríamos implementar una aplicación web que muestra estos datos en un dashboard.




FUENTES
BASES DE DATOS
Ficheros
Extracción
Extract
Transformación
Transform
Carga
Load
DATA WAREHOUSE

Y que pasa con machine learning?

Un proyecto de aprendizaje automático tiene muchos componentes móviles que deben estar unidos antes de que podamos ejecutarlo con éxito.

La capacidad de saber cómo construir una tubería de aprendizaje automático de extremo a extremo es un activo valioso.

Actividad

Construir un modelo de segmentación con Mllib

1. Construir un pipeline que nos permita generar segmentación

Spark Streaming

BigData - UBO2022

Sebastian Ulloa Quezada

Que es Spark Streaming?

Spark Streaming es un método para analizar información "sin límite", a veces conocida como información online. Esto se logra dividiéndolo en micro-batches y permitiendo la ejecución en lotes.

La interfaz de transmisión de Spark es un módulo de aplicación de la API de Spark. Python, Scala y Java son compatibles. Le permite manejar flujos de datos reales de una manera tolerante a fallos y flexible.

El Spark Engine toma los lotes de datos y produce el flujo de resultados finales en lotes.

Pipeline Streaming?

Es una tecnología que permite que los datos se muevan de forma fluida y automática de una ubicación a otra. Esta tecnología elimina muchos de los problemas típicos, como la fuga de información, los cuellos de botella, los conflictos de datos múltiples y la creación repetida de entradas.

Los pipelines en streaming de data son arquitecturas de pipelinesde datos que procesan miles de entradas en tiempo real y escaladles.

Como resultado, podrás recopilar, analizar y retener muchos datos. Esta funcionalidad permite aplicaciones, monitoreo e informes en tiempo real.

Arquitectura de Streaming en Spark

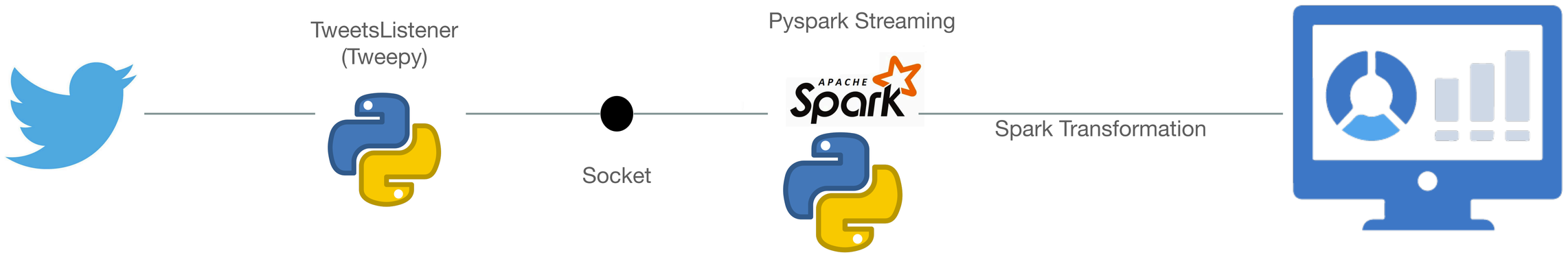
Que es Spark Streaming?

La estructura principal de Spark Streaming es la transmisión de tiempo discreto lote por lote.

Los microbatches se asignan y analizan constantemente, en lugar de viajar a través del Pipeline de procesamiento de un elemento a la vez.

Como resultado, los datos se distribuyen a los workers en función de los recursos y la ubicación accesibles.

Cuando se reciben los datos, el receptor los divide en divisiones RDD.



https://developer.twitter.com/en

https://raw.githubusercontent.com/lawlesst/vivo-sample-data/master/data/csv/people.csv

Actividades

Streaming Data Con Pyspark

1. Construir y probar de manera offline el funcionamiento de un streaming con pyspark
2. Construir un extractor de twitters usando Tweepy
3. Construir un monitor de hashtag con tweepy y pyspark

Cloud computing

BigData - UBO2022

Sebastian Ulloa Quezada

Que es el cloud computing?

Cloud computing es acceso bajo demanda, a través de Internet, a recursos informáticos, como aplicaciones, servidores (servidores físicos y servidores virtuales), almacenamiento de datos, herramientas de desarrollo, funcionalidades de red y más, alojados en un centro de datos remoto gestionado por un proveedor de servicios cloud (o CSP). El CSP ofrece estos recursos por una cuota de suscripción mensual o los factura según el uso.

En comparación con la TI en local tradicional, y en función de los servicios cloud que seleccione, cloud computing ayuda a:

- **Reducir los costes de TI:** el cloud le permite descargar algunos o la mayoría de los costes y el esfuerzo de comprar, instalar, configurar y gestionar su propia infraestructura en local.

- **Mejorar la agilidad y acelerar la generación de valor:** con cloud, su organización puede empezar a utilizar aplicaciones empresariales en minutos, en lugar de esperar semanas o meses para que la TI responda a una solicitud, adquiera y configure hardware de soporte e instale software. Cloud también le permite capacitar a determinados usuarios, más concretamente a desarrolladores y científicos de datos, para ayudarse a sí mismos con la infraestructura de software y soporte.

- **Escalar de forma más fácil y rentable:** cloud proporciona elasticidad; en lugar de adquirir una capacidad excesiva que se queda sin utilizar durante períodos de poco trabajo, puede aumentar o disminuir la capacidad en respuesta a picos y caídas en el tráfico. También puede aprovechar la red global de su proveedor de cloud para acercar sus aplicaciones a los usuarios de todo el mundo.

El término "cloud computing" también se refiere a la tecnología que habilita el funcionamiento del cloud. Esto incluye algún tipo de *infraestructura de TI virtualizada*, servidores, software de sistema operativo, red y otra infraestructura que se abstrae, utilizando software especial, de modo que la TI se pueda agrupar y dividir independientemente de los límites físicos del hardware. Por ejemplo, un único servidor de hardware se puede dividir en varios servidores virtuales.

1. La virtualización permite a los proveedores de cloud maximizar uso de sus recursos del centro de datos. No es de extrañar que muchas empresas hayan adoptado el modelo de entrega en cloud para su infraestructura en local para conseguir la máxima utilización y un ahorro de costes, en comparación con la infraestructura de TI tradicional y ofrecer el mismo autoservicio y agilidad a sus usuarios finales.

SaaS

PaaS

IaaS

Hosted applications

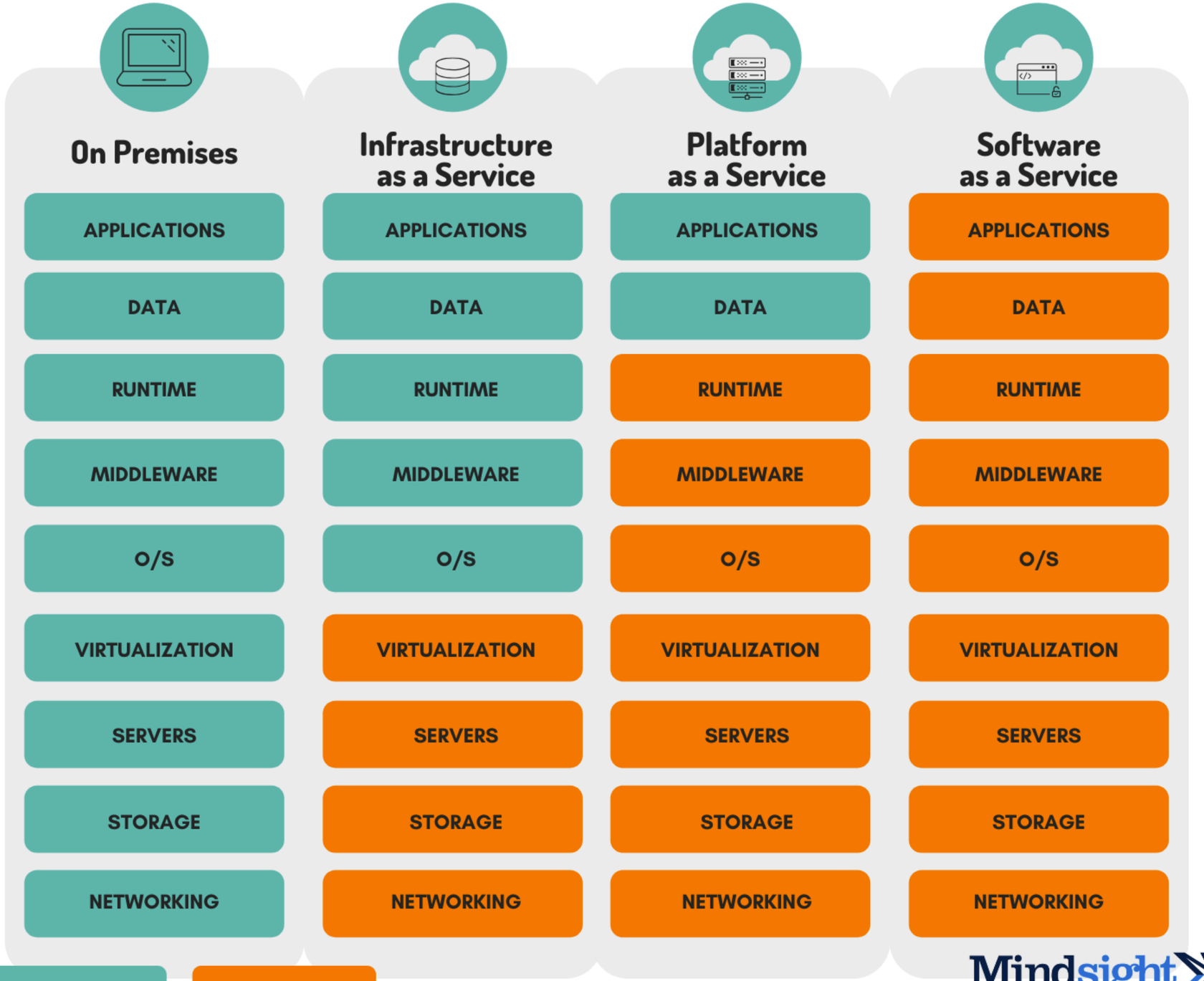
Development tools, database management, business analytics

Operating systems

Servers and storage

Networking firewalls and security

Data center Physical plant or building


On Premises
Infrastructure as a Service
Platform as a Service
Software as a Service
APPLICATIONS
DATA
RUNTIME
MIDDLEWARE
O/S
VIRTUALIZATION
SERVERS
STORAGE
NETWORKING
YOU MANAGE
OTHERS MANAGE
Mindsight
Technology. Transparency. Trust.

SaaS (Software como servicio)

SaaS, también conocido como software basado en cloud o aplicaciones cloud, es el software de aplicación que se aloja en el cloud y que se accede y utiliza a través de un navegador web, un cliente de escritorio dedicado o una API que se integra con el sistema operativo de su escritorio o móvil .

En la mayoría de los casos, los usuarios de SaaS pagan una cuota de suscripción mensual o anual; algunos pueden ofrecer precios de pago por uso en función de su uso real.

Además de los beneficios de ahorro de costes, tiempo de generación de valor y escalabilidad del cloud, SaaS ofrece lo siguiente:

- **Actualizaciones automáticas:** con SaaS, puede aprovechar las nuevas características en el mismo momento en que el proveedor las añade, sin tener que coordinar una actualización en local.

- **Protección frente a pérdida de datos:** debido a que los datos de aplicación residen en el cloud, con la aplicación, no perderá los datos si el dispositivo se bloquea o se rompe.

PaaS (Plataforma como servicio)

PaaS proporciona a los desarrolladores de software plataforma bajo demanda (hardware, pila de software completa, infraestructura e incluso herramientas de desarrollo) para ejecutar, desarrollar y gestionar aplicaciones sin el coste, la complejidad y la inflexibilidad de mantener la plataforma en local.

Con PaaS, el proveedor de cloud aloja todo (servidores, redes, almacenamiento, software de sistema operativo, middleware, bases de datos) en su centro de datos. Los desarrolladores simplemente escogen de un menú los servidores y entornos que necesitan para ejecutar, crear, probar, desplegar, mantener, actualizar y escalar aplicaciones.

Actualmente, PaaS a menudo se crea con base en contenedores, un modelo de computación virtualizado, un paso que se ha eliminado de los servidores virtuales. Los contenedores virtualizan el sistema operativo, lo que permite a los desarrolladores empaquetar la aplicación con solo los servicios del sistema operativo que necesita para ejecutarse en cualquier plataforma, sin modificación y sin requerir middleware.

IaaS (Infraestructura como servicio)

IaaS proporciona acceso bajo demanda a los recursos informáticos fundamentales (servidores físicos y virtuales, redes y almacenamiento) a través de Internet en una base de pago por uso. IaaS permite a los usuarios finales aumentar y reducir los recursos según sea necesario, reduciendo los elevados gastos de capital iniciales y evitando una infraestructura de propiedad o en local innecesaria o la compra excesiva de recursos para acomodar incrementos periódicos de uso.

En contraste con SaaS y PaaS (y los modelos informáticos de PaaS aún más nuevos como contenedores y sin servidor), IaaS proporciona a los usuarios el control de nivel inferior de los recursos informáticos en el cloud.

IaaS era el modelo de cloud computing más popular cuando surgió a principios de la década de 2010.Aunque sigue siendo el modelo de cloud para muchos tipos de cargas de trabajo, el uso de SaaS y PaaS está creciendo a un ritmo mucho más rápido.