BIG DATA for BUSINESS

2.10 Data Engineering en Servicios Cloud

Conecta Empleo

Contenido desarrollado por
Synergic Partners

Índice del módulo

2.10. DATA ENGINEERING EN SERVICIOS CLOUD

- ¿Qué es un servicio en la nube (Cloud)?
- Modelos de Servicios Cloud
- Tipos de Implementación
- Oferta de Servicios en AWS, Microsoft Azure y Google Cloud
- Talleres Prácticos - Data Engineering en Google Cloud Platform

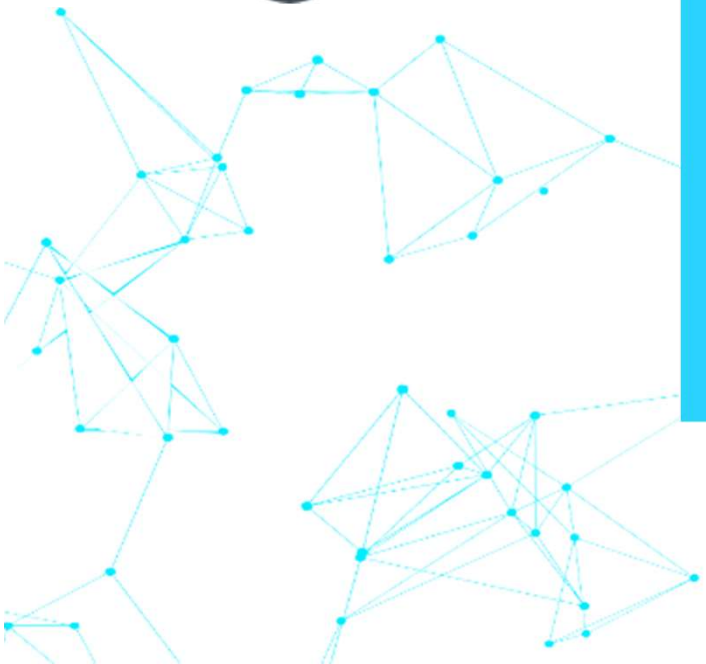
¿Qué es un Servicio en la nube (Cloud)?

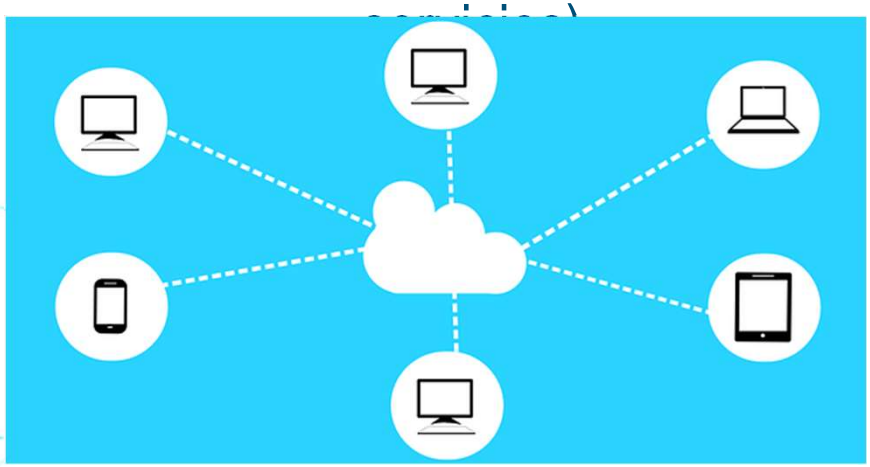
¿Qué es un servicio en la nube (Cloud)?

“No existe tal cosa como la nube, es simplemente un ordenador en otra parte”

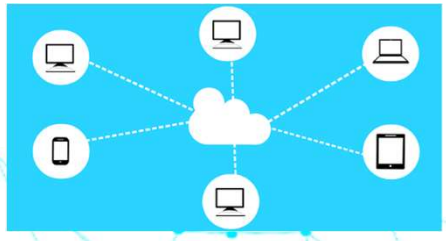
¿Qué es un servicio en la nube (Cloud)?

Cloud Computing es un modelo de servicios de IT que pone a disposición de los usuarios un servicio bajo demanda de acceso a una red de ordenadores y servidores dentro de un pool compartido de recursos (red, servidores, almacenamiento, aplicaciones y





Características del cloud computing:



1. **Servicio bajo demanda:** el usuario puede dar de alta a servicios de computación sin interactuar con una persona contacto en el proveedor. Además, paga sólo por lo que usa.

1. **Acceso por Internet o VPN (o ambas):** generalmente accesible desde cualquier parte del mundo.

1. **Pooling de recursos:** los recursos de computación del proveedor están compartidos para atender a varios (muchos) clientes a la vez.

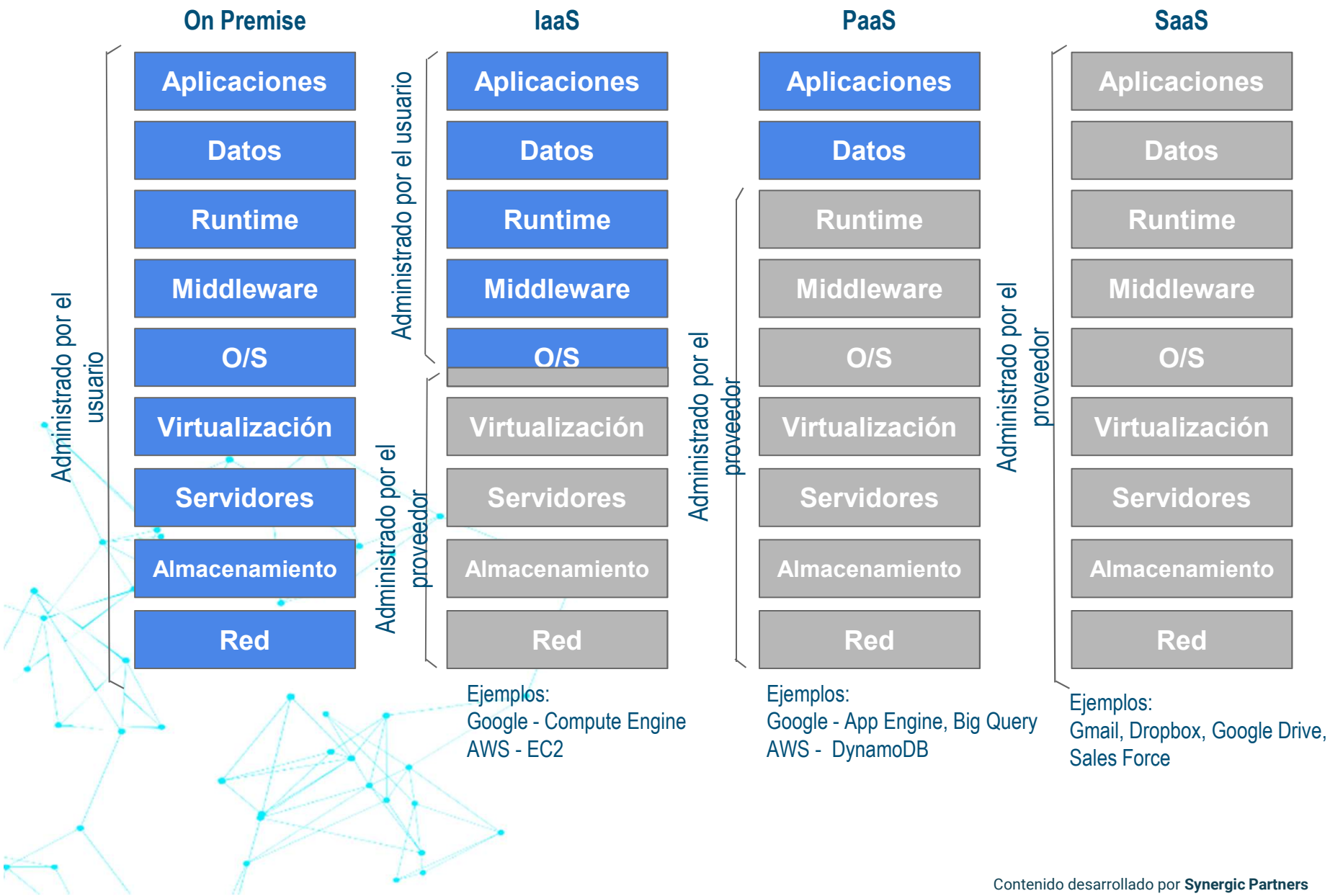
1. **Servicios elásticos:** el usuario puede escalar los servicios que usa con rapidez, en muchos casos automáticamente.

1. **Servicios medibles:** el uso de los recursos de la red utilizados por el usuario son 100% medibles y reportables con la finalidad de controlar costes, usos y seguridad de red.

Modelos de Servicios Cloud

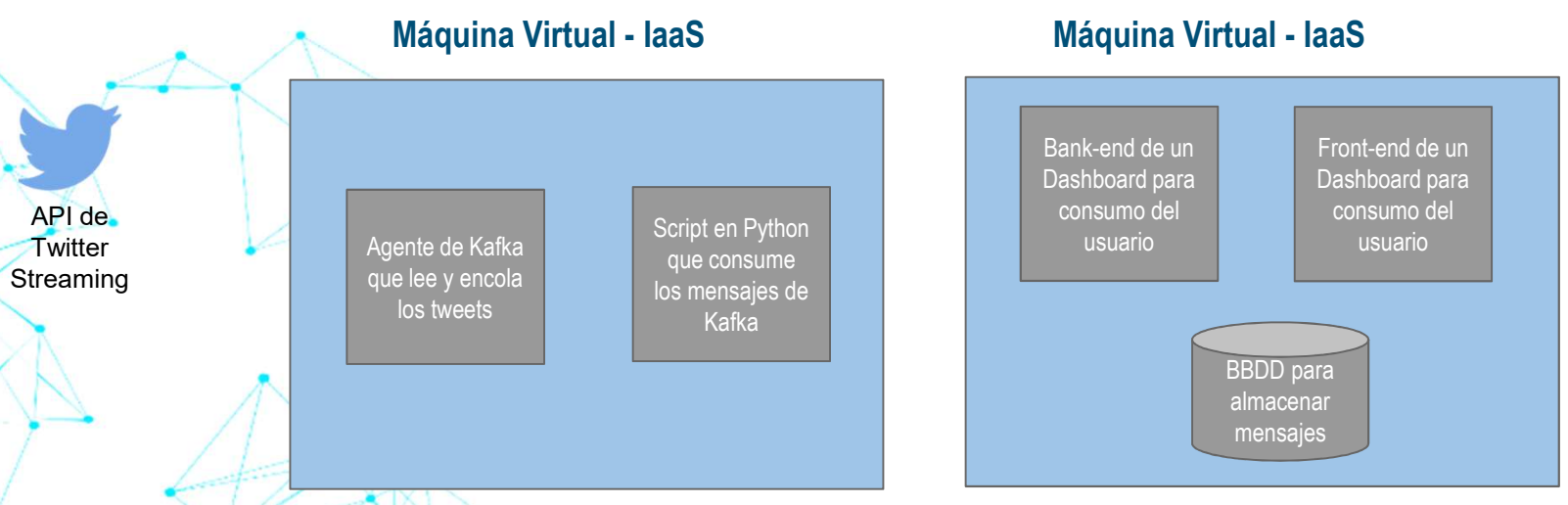
Contenido desarrollado por **Synergic Partners**

Modelos de Servicios Cloud

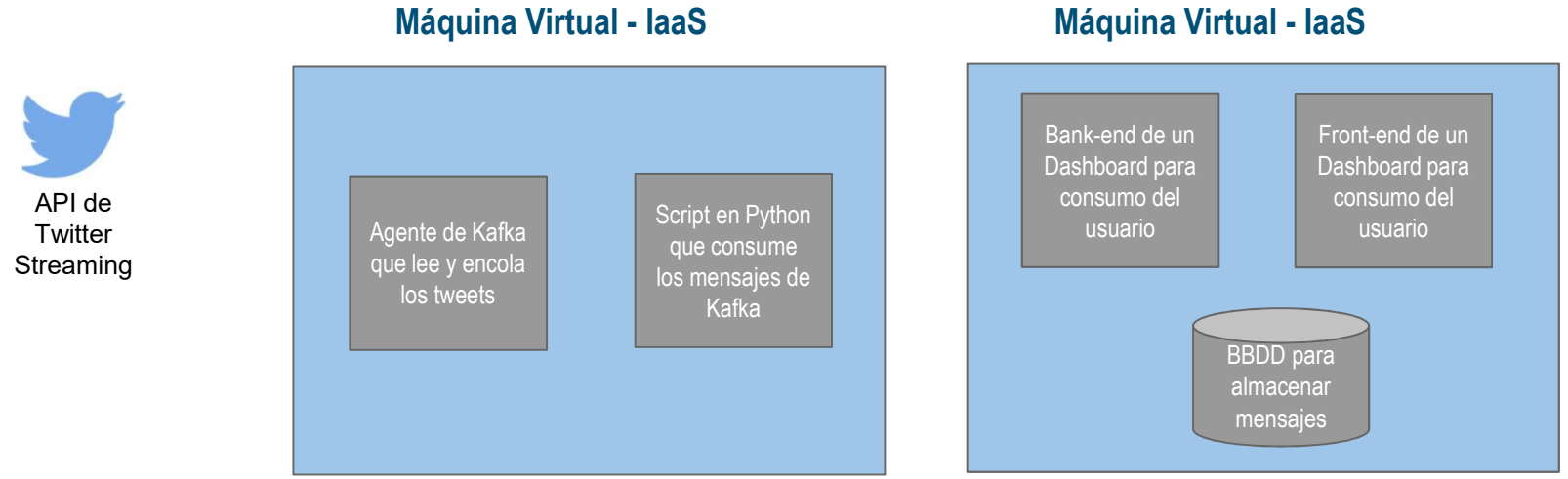


Ejemplo On Premise - Aplicación para Analizar Twitter en Real Time

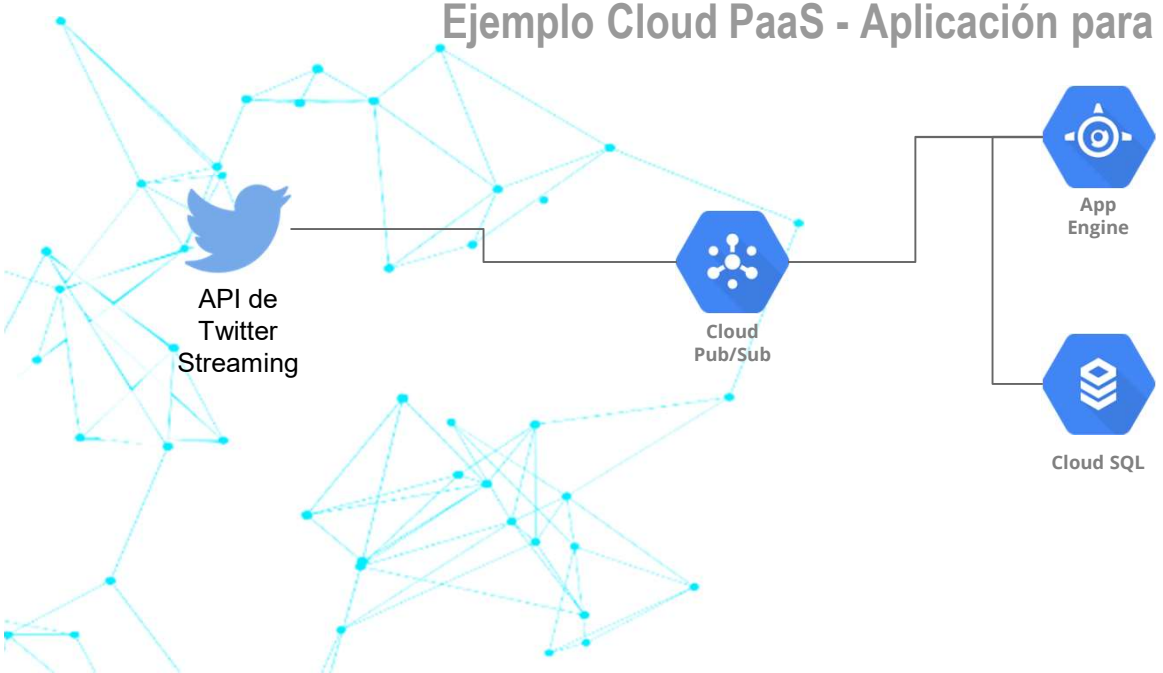
Ejemplo Cloud IaaS - Aplicación para Analizar Twitter en Real Time



Ejemplo Cloud IaaS - Aplicación para Analizar Twitter en Real Time



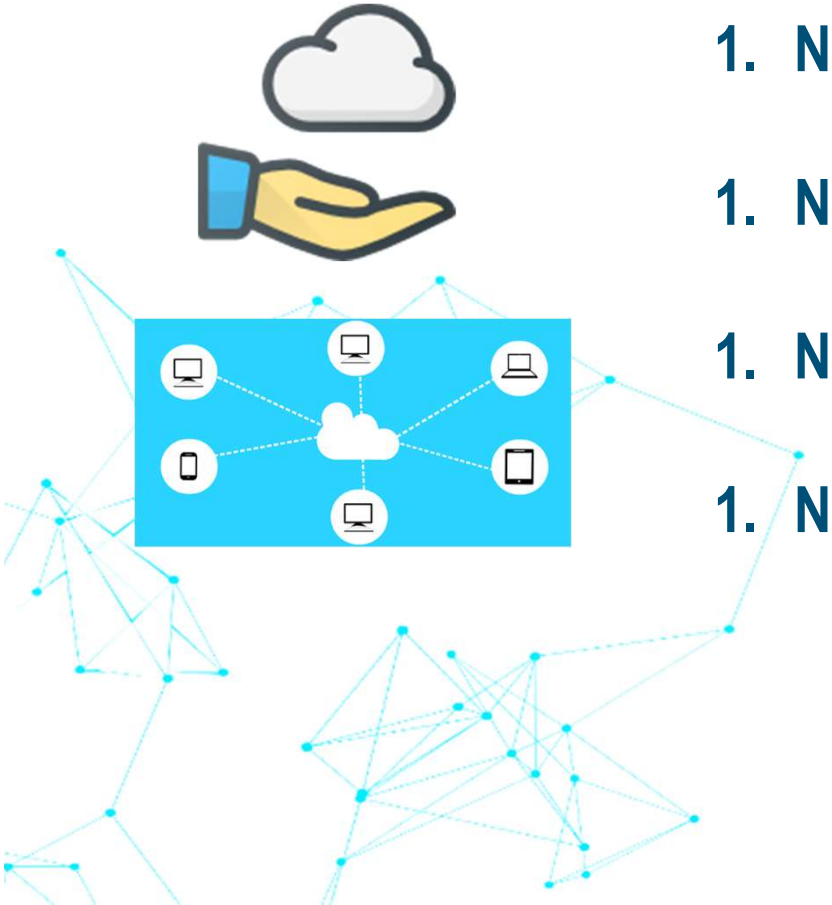
Ejemplo Cloud PaaS - Aplicación para Analizar Twitter en Real Time



- Autoescalable
- Tolerante a fallos
- Distribuido geográficamente
- Seguro

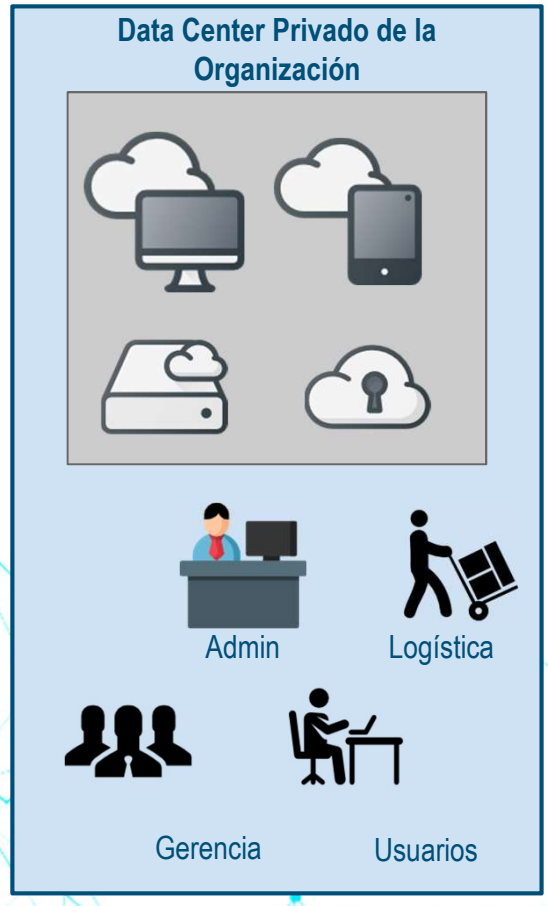
Tipos de Implementación

Tipos de Implementación:



1. Nube Privada
1. Nube Pública
1. Nube Comunitaria
1. Nube Híbrida

Tipos de Implementación - Nube Privada



Características:

- Todos el HW y SW de la infraestructura está dentro del data center privado de la empresa.
- Sin embargo ofrece los servicios de un proveedor en la nube:
  - Servicio bajo demanda, con mínima intervención humana.
  - Acceso a través de la red, en este caso la red corporativa.
  - Pooling de recursos.
  - Rápida elasticidad.
  - Servicio medible

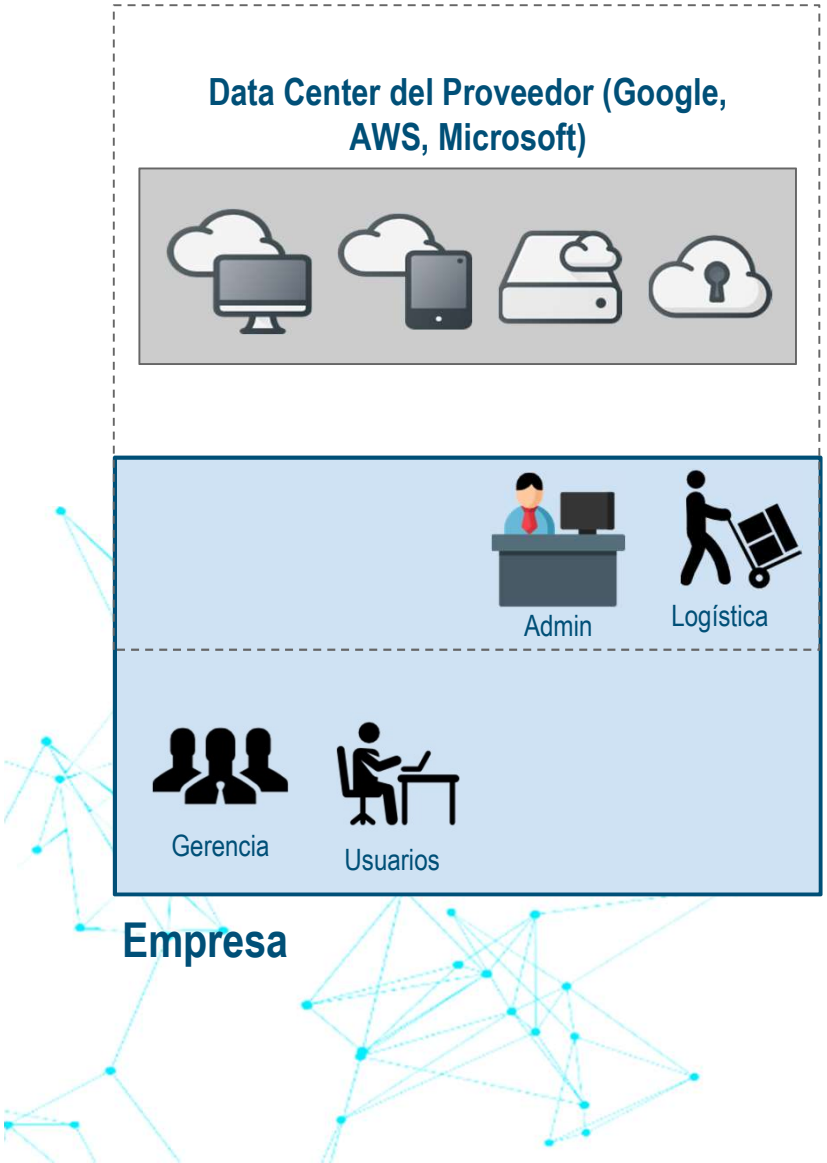
Ventajas:

- Máximo control por parte de la organización.
- Adecuada para modelos de negocio con restricción legal de ubicación, almacenamiento y manejo de los datos (por ejemplo: bancos)
- Control total sobre la seguridad.

Desventajas:

- Mayores costes fijos (electricidad, oficinas, personal, seguros, licencias, etc)
- La seguridad puede estar a riesgo si la organización no tiene experticia.
- Tener un data center no significa tener todos los servicios cloud funcionando adecuadamente.

Tipos de Implementación - Nube Pública



Características:

- Todos el HW y SW de la infraestructura está dentro del data center del proveedor.
- Cumple con estas características:
  - Servicio bajo demanda, con mínima intervención humana.
  - Acceso a través de la red, en este caso la red corporativa.
  - Pooling de recursos.
  - Rápida elasticidad.
  - Servicio medible

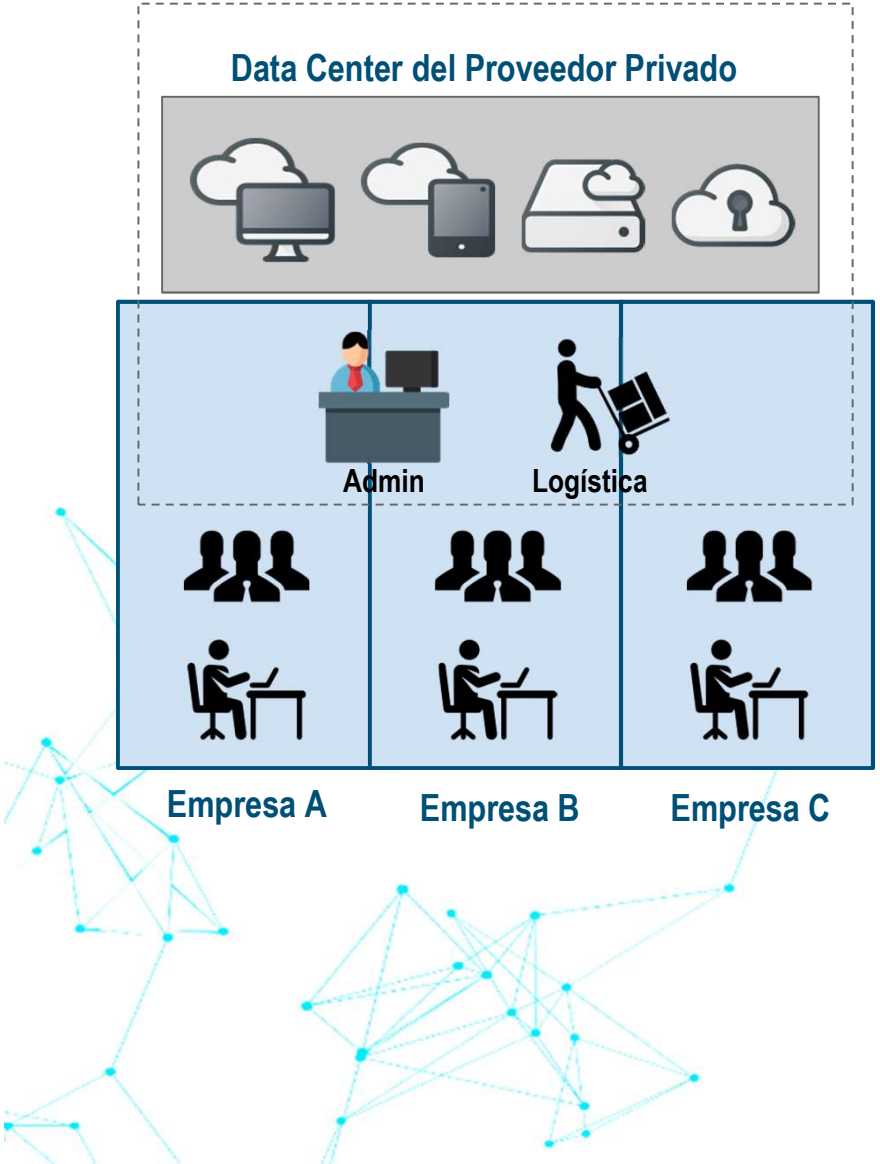
Ventajas:

- Menores costes y cambio de CAPEX por OPEX
- La seguridad en general es mejor que en data centers on premise.
- La empresa se enfoca en su negocio y no en IT.

Desventajas:

- En algunos modelos de negocio existen regulaciones que restringen el uso y almacenamiento de datos fuera de un país o región.

Tipos de Implementación - Nube Comunitaria



Características:

- Un proveedor privado suministra un data center con servicios cloud para compartir entre un grupo de empresas.

Ventajas:

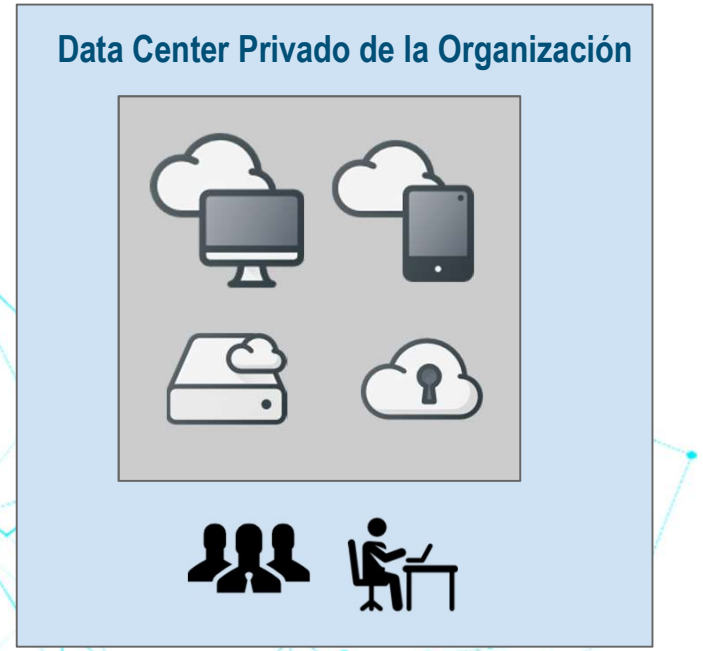
- Menores costes y cambio de CAPEX por OPEX
- La seguridad en general es mejor que en data centers on premise.
- La empresa se enfoca en su negocio y no en IT.
- Se controlan mejor temas regulatorios como ubicación y almacenamiento de los datos.

Desventajas:

- Agrega el trabajo de coordinación, contratos, temas legales entre las empresas participantes.

Tipos de Implementación - Nube Híbrida

Algunos servicios migrados a un proveedor público como Google, AWS, Microsoft



Características:

- Parte de los servicios están implementados en una nube privada de la empresa, en su propio data center.
- Parte de los servicios están implementados en una o varias nubes públicas (Google, AWS, Microsoft, etc)

Ventajas:

- Optimización de costes.
- Se mantiene total control sobre parte de los servicios.
- Puede ayudar a cumplir con temas legales y regulatorios.

Desventajas:

- Coordinar los servicios en privado con los de la nube pública puede ser complejo, costoso y llevar a problemas de seguridad.
- No elimina del todo la necesidad de tener en la empresa personal, procesos y recursos dedicados a IT y no al negocio principal de la empresa.

Oferta de Servicios
AMAZON AWS

Contenido desarrollado por **Synergic Partners**

IaaS - Computación

Amazon EC2

Amazon Lightsail

Amazon ECS

- Máquinas virtuales
- CPU + Almacenamiento
- SO disponibles:
  - Linux en distintas distribuciones.
  - Windows Server en distintas versiones
- Máquinas desde 1 vCPU compartido hasta 128 vCPUs dedicados y desde 0.6 GB de RAM hasta casi 1 TB de RAM
- Usos:
  - Desarrollo de aplicaciones
  - Aplicaciones en producción
  - Clusters de Máquinas, ej: Clouster Hadoop para Big Data
- Diferencias entre EC2 y Lightsail:
  - El usuario tiene total control sobre la configuración de una EC2. Lightsail son máquinas EC2 pre-configuradas.
  - Al ser pre-configuradas, Lightsail ahorra costes de gestión del servidor.
- Amazon ECS - EC2 Container Service:
  - Permite administrar contenedores, por ejemplo Dockers, para desplegar aplicaciones y servicios dentro de los mismos.

IaaS - Almacenamiento

- S3:
  - Almacenamiento en la nube de cualquier tipo de datos (ficheros de texto, CSV, vídeo, imágenes, script de software, etc)
  - Totalmente escalable.
  - Se paga por uso, por GB al mes.
  - Alta disponibilidad.
  - Disponible en varias regiones en el mundo de manera de buscar localizar los datos lo más cercano posible a las aplicaciones.

- Glacier:
  - Similar a S3, pero más económico.
  - Está diseñado para almacenar datos de manera segura pero a los que no necesitamos acceder con frecuencia.
  - Se paga por uso y por descargas. Si son muy pocas descargas al año, es mucho más barato que usar S3.
  - La disponibilidad es mejor que en S3 por la forma en la que está diseñado.

- EFS (Elastic File System):
  - Son los volúmenes de almacenamiento para las máquinas EC2

IaaS - Redes

Amazon VPC

- Amazon VPC (Red Privada Virtual):
  - Permite crear en AWS una red VNP con las mismas características y servicios que una VPN virtual (filtrado de rango de IPs, aislamiento lógico del resto de la red, seguridad, etc)

Elastic Load Balancing

- Elastic Load Balancing:
  - Distribuye automáticamente el tráfico entrante de las aplicaciones entre varias instancias de Amazon EC2.
  - Permite conseguir tolerancia a errores en las aplicaciones, ofreciendo la capacidad de balanceo de carga.

Amazon Route 53

- Amazon Route 53:
  - Es un servicio web DNS (sistema de nombres de dominio) escalable y de alta disponibilidad.

AWS Direct Connect

- AWS Direct Connect:
  - Permite establecer una conexión de red dedicada desde las instalaciones de una empresa a AWS.
  - Se utiliza principalmente para conectar el data center de la empresa directamente con su infraestructura y aplicaciones en AWS.

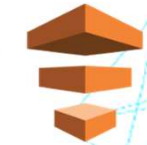
PaaS - Computación

- AWS Lambda:
  - Permite ejecutar código sin aprovisionar ni administrar servidores.
  - Se paga por el tiempo de cómputo que consuma
  - Servicio autogestionado, escalable y de alta disponibilidad.

- AWS Batch:
  - Permite ejecutar trabajos en modo batch si aprovisionar servidores.
  - El servicio es autoescalable y autogestionado.
  - Se paga por el tiempo de cómputo y recursos utilizados.

- Amazon EMR (Elastic Map Reduce):
  - Es un cluster Hadoop autogestionado.
  - Permite ejecutar trabajos de Hadoop, HBase, Spark, Hive, Flink, y otros.
  - Es escalable dinámicamente.
  - Se paga por tiempo de uso.

- Amazon Kinesis:
  - Permite el procesamiento de mensajería a tiempo real, similar a Apache Kafka.



- AWS Data Pipeline:
  - Permite ejecutar de manera automática labores de ETL.
  - Es autoescalable.

PaaS - Bases de Datos

Amazon DynamoDB

- Amazon DinamoDB:
  - Base de datos NoSQL
  - Es autogestionada y autoescalable.
  - Compatible con modelos de bases de dato clave valor y de documentos.

Amazon Aurora

- Amazon Aurora:
  - Base de datos relacional compatible con MySQL y PostgreSQL.
  - Es autoescalable, autogestionado y distribuido

Amazon Redshift

- Amazon Redshift:
  - Almacén de datos que permite realizar consultas SQL y conexión con muchas herramientas de BI ya en el mercado.
  - Diseñado para casos de uso de grandes volúmenes de datos (desde Tera hasta Petabytes)

Amazon ElastiCache

- Amazon ElastiCache
  - Permite implementar un almacén de datos en memoria en la nube.
  - Se utiliza para mejorar el desempeño de aplicaciones web.
  - Es una alternativa a bases de datos basadas en disco cuando la velocidad de acceso a los datos es crítica para el modelo de negocio.

PaaS - Inteligencia Artificial

Amazon Lex

Amazon Machine Learning

Amazon Polly

Amazon Rekognition

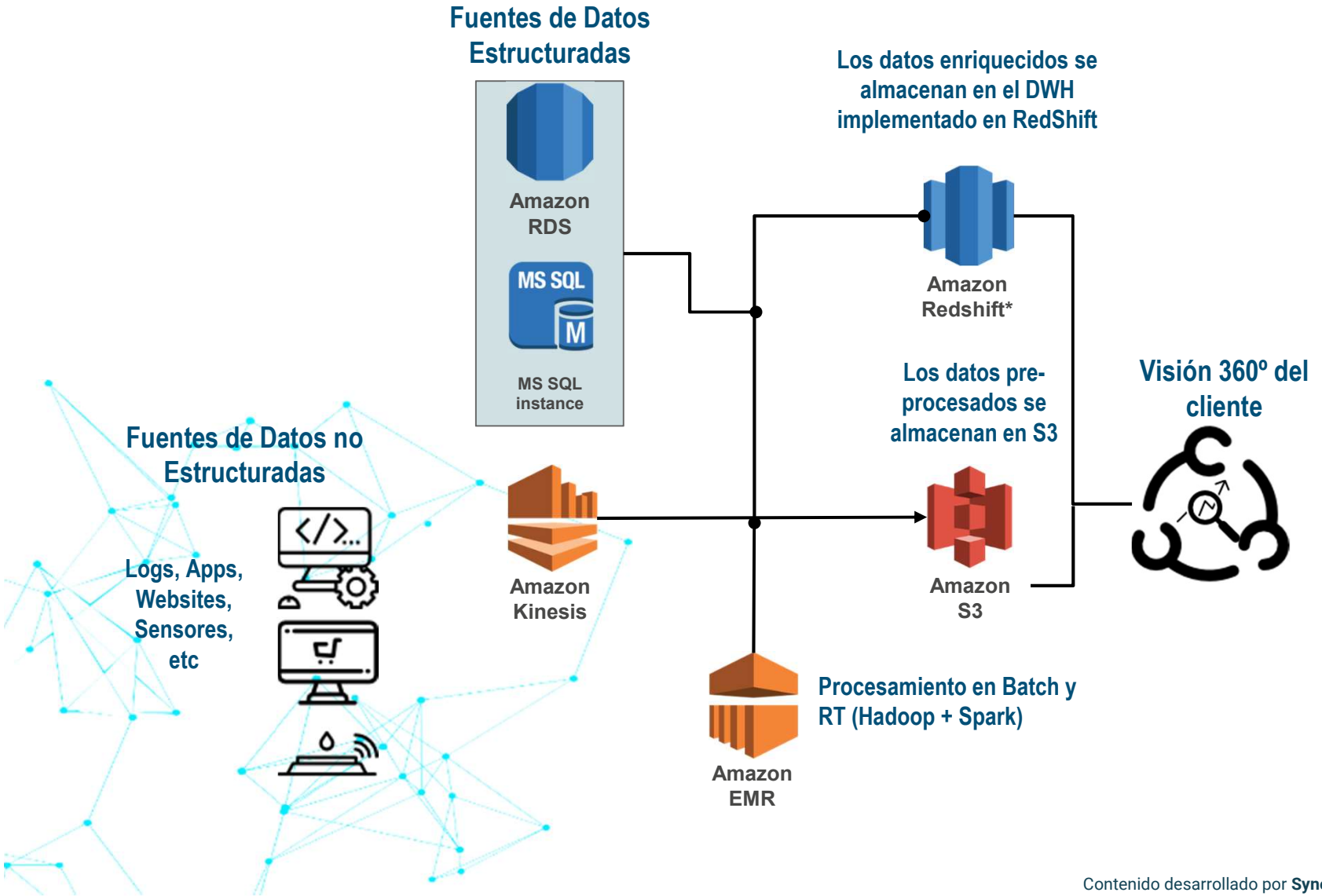
- Amazon Lex:
  - API que permite crear chatbots de voz y de texto.
  - Tiene modelos de Deep Learning entrenados para reconocimiento de voz y procesamiento de texto.
- Amazon Machine Learning:
  - Permite crear modelos de Machine Learning sin la necesidad de administrar la infraestructura donde se entrenan los modelos.
  - Tiene un API que permite exponer al mundo exterior los modelos entrenados.
- Amazon Polly:
  - Permite convertir texto a habla.
  - Tiene modelos de Deep Learning implementados para procesamiento del lenguaje natural (NLP)
- Amazon Rekognition:
  - Servicio de reconocimiento de imágenes.
  - Tiene modelos Convolucionales de Deep Learning entrenados para visión artificial.

Caso de Uso - Visión 360º del cliente

Situación:

- Hoy en día las empresas tienen una relación omni-canal con los clientes.
- Esto implica que obtienen información del cliente por varios canales: website, móvil, call center, redes sociales, oficinas físicas, entre otros.
- Para obtener una visión 360º del cliente, tienen que integrar todas estas fuentes de datos en un sólo sitio disponible a toda la organización para el posterior análisis de la información.
- A este sitio donde se integran todas estas fuentes de datos se le llama hoy en día **"Data Lake"** (Lago de datos).
- El siguiente diagrama muestra una posible implementación de Data Lake utilizando los servicios de Amazon Web Services.

Caso de Uso - Visión 360º del cliente



Oferta de Servicios
GOOGLE CLOUD PLATFORM (GCP)

Contenido desarrollado por **Synergic Partners**

IaaS - Computación

- Compute Engine:
  - Máquinas virtuales en la nube.
  - Disponibles desde 1 vCPU compartido, hasta 64 vCPU dedicados y desde 0.6 GB de RAM hasta 416 GB
  - SO disponibles linux y Windows Server en distintas versiones.

- GPU:
  - Máquinas virtuales dotadas con tarjetas aceleradoras gráficas
  - Disponibles desde 1 GPU hasta 8 GPUs por máquinas
  - Tarjetas NVIDIA
  - Muy eficientes para su uso en Deep Learning y en Rendering (efectos especiales, diseño gráfico, etc)

- Container Engine:
  - Permite desplegar contenedores tipo Docker en la nube.
  - Se utilizan para desarrollar aplicaciones y desplegarlas en ambientes autocontenidos.
  - Desarrollo de micro-servicios.

IaaS - Almacenamiento

- Cloud Storage:
  - Almacenamiento permanente en la nube para cualquier clase de archivos (vídeo, texto, CSV, scripts, sonido, etc)
  - Se paga por GB de uso.
  - Tiene modalidades de alta disponibilidad (más caro), y baja disponibilidad almacenamiento en frío (más barato)
  - Tiene modalidades de almacenamiento regional y global.

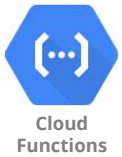
- Persistent Disk:
  - Volúmenes de almacenamiento para las máquinas virtuales Compute Engine y GPUs

IaaS - Redes

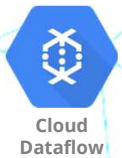
- Cloud VPN:
  - Servicio que permite crear una VPN (Red Privada Virtual) en la nube.
  - Permite aislar lógicamente una parte de la red del resto de servicios de la empresa.
  - Provee los mismos servicios que una VPN on premise.
- Cloud Load Balancing:
  - Servicio para balancear carga de tráfico entre aplicaciones.
  - Se usa para dar alta disponibilidad a las aplicaciones.
- Cloud DNS:
  - Servicio que permite implementar un servidor de nombres de dominio en la nube.
- Cloud Interconnect:
  - Servicio empresarial que permite conectar el data center de la empresa directamente a los servicios de Google Cloud.

PaaS - Computación

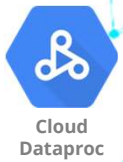
- App Engine:
  - Permite desarrollar y desplegar aplicaciones web.
  - Es autogestionado, el usuario no necesita configurar servidores
  - Es escalable automáticamente.
  - Soporta apps en Python, Java, PHP o Go.



- Cloud Functions:
  - Permite ejecutar aplicaciones sin servidores que se ejecutan en base a la lógica de negocio (eventos) o según un calendario.
  - Se paga por el tiempo de cómputo de la aplicación.



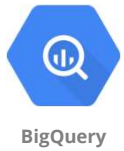
- Cloud Dataflow:
  - Permite crear un flujo ETL de extracción y transformación de datos.
  - Es autoescalable y robusto para aplicaciones de grandes volúmenes de datos.



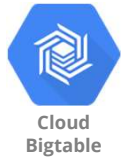
- Cloud Dataproc:
  - Permite desplegar un cluster Hadoop en la nube para ejecutar trabajos en Hadoop, Spark, Hive, Pig, entre otros

- Cloud Pub/Sub:
  - Servicio de encolamiento de mensajes similar a Apache Kafka.
  - Muy útil para aplicaciones de procesamiento de datos en Real Time.

PaaS - Bases de Datos

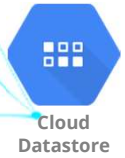


- BigQuery:
  - Implementar un Data Warehouse en el cual se pueden almacenar petabytes de datos y hacer consultas SQL en segundos (o milisegundos).
  - Es autogestionado, no hace falta administrar la base de datos desde el punto de vista de performance y escalado.



- Cloud Bigtable:
  - Base de datos NoSQL en la nube.
  - Servicio autogestionado y autescalable.
  - Base de datos columnar.

- Cloud SQL:
  - Servicio que permite implementar un servidor de bases de datos MySQL o PostgreSQL en la nube.
  - Similar a BigQuery pero para volúmenes de datos menor.



- Cloud Datastore:
  - Base de datos NoSQL en la nube.
  - Orientada a documentos.
  - Autoescalable y autogestionada.

PaaS - Inteligencia Artificial

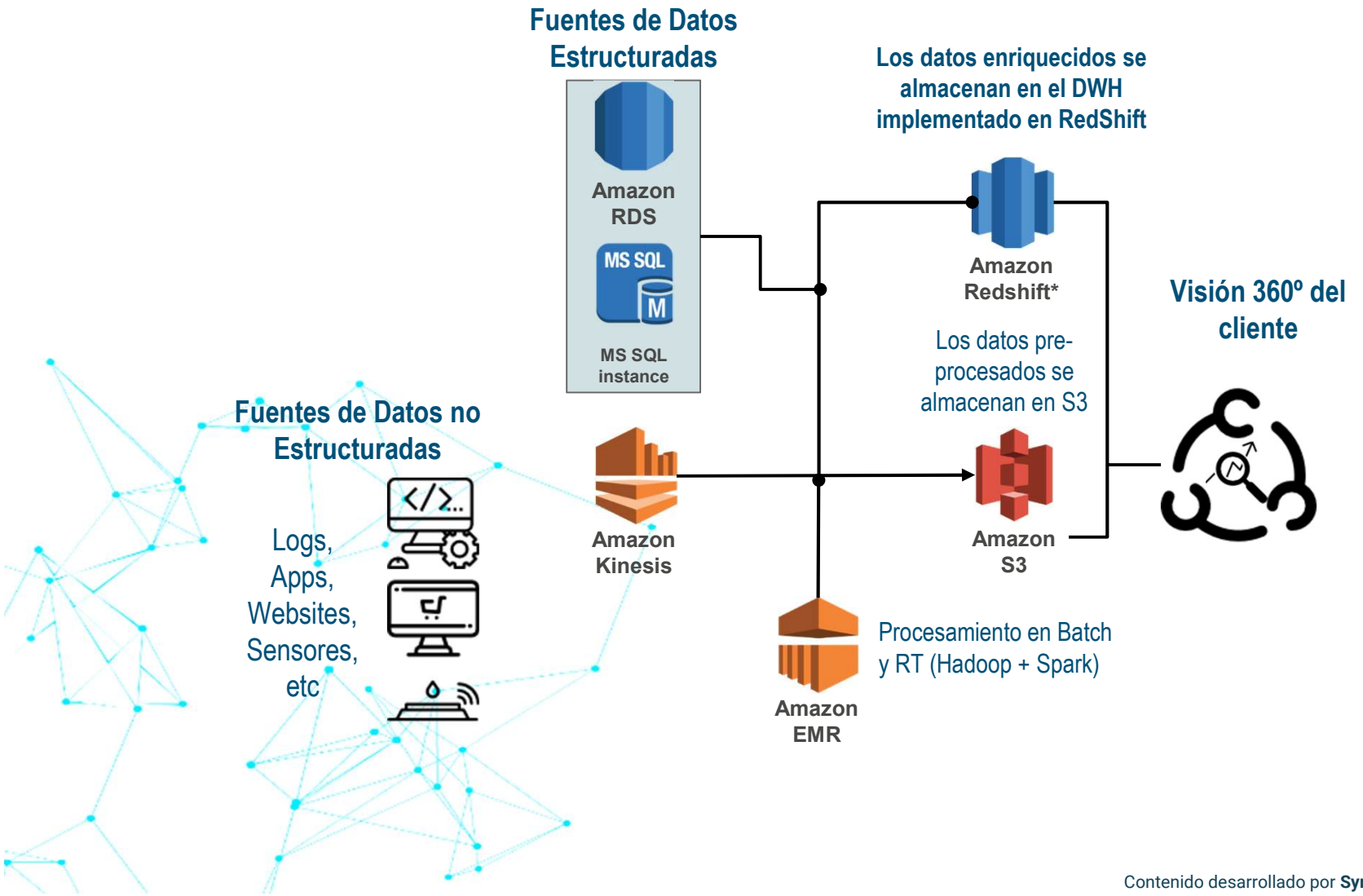
- Cloud Machine Learning:
  - Permite desarrollar modelos de machine learning con TensorFlow.
  - El desarrollador no tiene que gestionar la infraestructura. La misma se adapta automáticamente a las necesidades del modelo en cuanto a tamaño de los datos y cómputo necesario durante el entrenamiento.
  - Se paga sólo por el tiempo de cómputo, no de desarrollo.
- Distintas APIs con modelos de Deep Learning ya implementados:
  - **Cloud Vision API:** modelos pre-entrenados de visión artificial.
  - **Cloud Speech API:** modelos pre-entrenados de reconocimiento de voz.
  - **Cloud Video Intelligence API:** modelos de tratamiento de vídeo.
  - **Cloud Natural Language API:** modelos de procesamiento del lenguaje Natural.
  - **Cloud Translation API:** modelos para traducir texto entre distintos idiomas.

Caso de Uso - Visión 360º del cliente

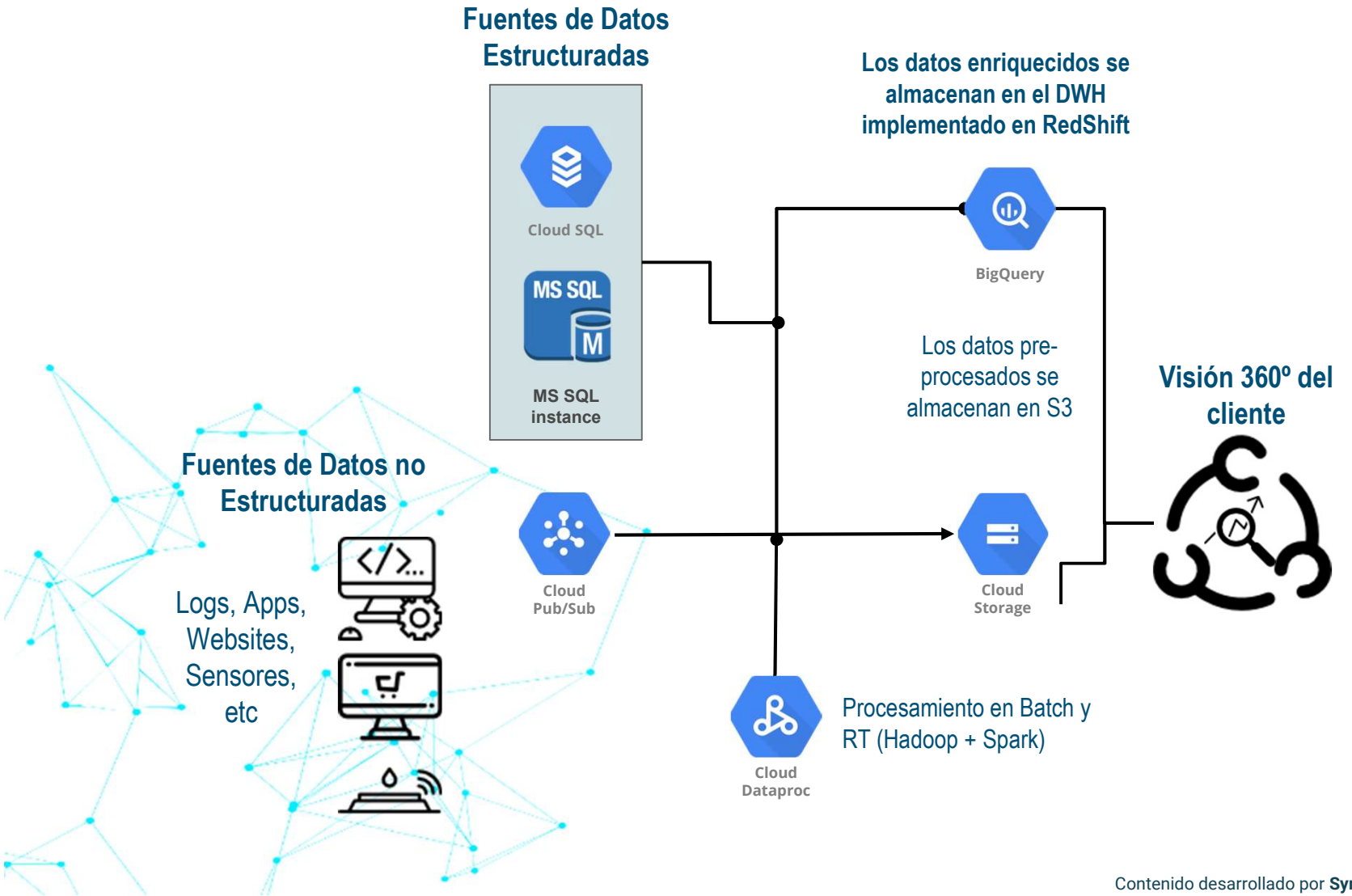
Situación (Mismo Caso de Uso explicado para AWS):

- Hoy en día las empresas tienen una relación omni-canal con los clientes.
- Esto implica que obtienen información del cliente por varios canales: website, móvil, call center, redes sociales, oficinas físicas, entre otros.
- Para obtener una visión 360º del cliente, tienen que integrar todas estas fuentes de datos en un sólo sitio disponible a toda la organización para el posterior análisis de la información.
- A este sitio donde se integran todas estas fuentes de datos se le llama hoy en día **"Data Lake"** (Lago de datos).
- El siguiente diagrama muestra una posible implementación de Data Lake utilizando los servicios de Amazon Web Services.

AWS
Caso de Uso - Visión 360º del cliente



GCP
Caso de Uso - Visión 360º del cliente



Oferta de Servicios
MICROSOFT AZURE

Contenido desarrollado por **Synergic Partners**

IaaS - Computación

Virtual Machines

- Virtual Machines:
  - Servidores virtuales en la nube.
  - SO disponibles: linux y Windows en distintas versiones.
  - Disponibles con CPUs y con GPUs

Container Instance

- Container Instance:
  - Permite desplegar contenedores tipo Docker en la nube.

IaaS - Almacenamiento

Storage

Disk Storage

- Storage
  - Almacenamiento permanente en la nube para cualquier clase de archivos (vídeo, texto, CSV, scripts, sonido, etc)
  - Se paga por GB de uso.
  - Tiene modalidades de alta disponibilidad (más caro), y baja disponibilidad almacenamiento en frío (más barato)
  - Tiene modalidades de almacenamiento regional y global.
- Disk Store:
  - Volúmenes de almacenamiento para las máquinas virtuales de CPUs y GPUs

IaaS - Redes

Virtual Network

Load Balancer

DNS de Azure

ExpressRoute

- Virtual Network:
  - Servicio que permite crear una VPN (Red Privada Virtual) en la nube.
  - Permite aislar lógicamente una parte de la red del resto de servicios de la empresa.
  - Provee los mismos servicios que una VPN on premise.
- Load Balancer:
  - Servicio para balancear carga de tráfico entre aplicaciones.
  - Se usa para dar alta disponibilidad a las aplicaciones.
- DNS de Azure:
  - Servicio que permite implementar un servidor de nombres de dominio en la nube.
- ExpressRoute:
  - Servicio empresarial que permite conectar el data center de la empresa directamente a los servicios de Azure.

PaaS - Computación

App Service

Servicios en la nube

Data Factory

HDInsights

Stream Analytics

- App Service:
  - Permite desarrollar y desplegar aplicaciones web.
  - Es autogestionado, el usuario no necesita configurar servidores
  - Es escalable automáticamente.
  - Soporta apps en .NET, Java, Node.js, PHP y Python.
- Servicios en la nube:
  - Permite ejecutar aplicaciones sin servidores que se ejecutan en base a la lógica de negocio (eventos) o según un calendario.
  - Se paga por el tiempo de cómputo de la aplicación.
- Data Factory:
  - Permite crear un flujo ETL de extracción y transformación de datos.
  - Es autoescalable y robusto para aplicaciones de grandes volúmenes de datos.
- HDInsight:
  - Permite desplegar un cluster Hadoop en la nube para ejecutar trabajos en Hadoop, Spark, Hive, Pig, entre otros.
  - Una particularidad de este servicio en Azure es que permite desarrollar aplicaciones en RServer.
- Stream Analytics:
  - Servicio de encolamiento de mensajes similar a Apache Kafka.
  - Muy útil para aplicaciones de procesamiento de datos en Real Time.

PaaS - Bases de Datos

SQL Data Warehouse

- SQL Data Warehouse:
  - Implementar un Data Warehouse en el cual se pueden almacenar petabytes de datos y hacer consultas SQL en segundos (o milisegundos).
  - Es autogestionado, no hace falta administrar la base de datos desde el punto de vista de performance y escalado.

Redis Cache

- Redis Cache:
  - Base de datos en memoria para aplicaciones que requieren de muy baja latencia.

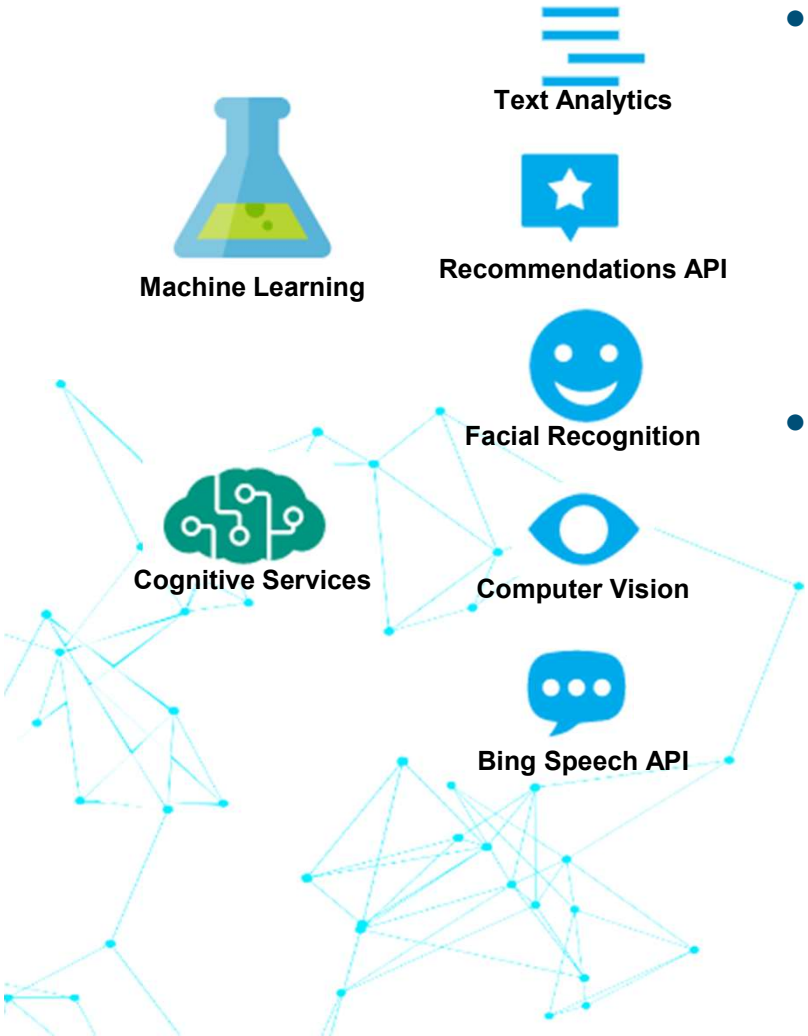
Azure Database for MySQL - PostgreSQL

- Azure Database for MySQL - PostgreSQL:
  - Servicio que permite implementar un servidor de bases de datos MySQL o PostgreSQL en la nube.
  - Similar a SQL Data Warehouse pero para volúmenes de datos menor.

Azure Cosmos DB

- Azure Cosmos DB:
  - Base de datos NoSQL en la nube.
  - Orientada a documentos.
  - Autoescalable y autogestionada.

PaaS - Inteligencia Artificial



- Machine Learning:
  - Permite desarrollar modelos de machine learning con CNTK y otras librerías.
  - El desarrollador no tiene que gestionar la infraestructura. La misma se adapta automáticamente a las necesidades del modelo en cuanto a tamaño de los datos y cómputo necesario durante el entrenamiento.
  - Se paga sólo por el tiempo de cómputo, no de desarrollo.

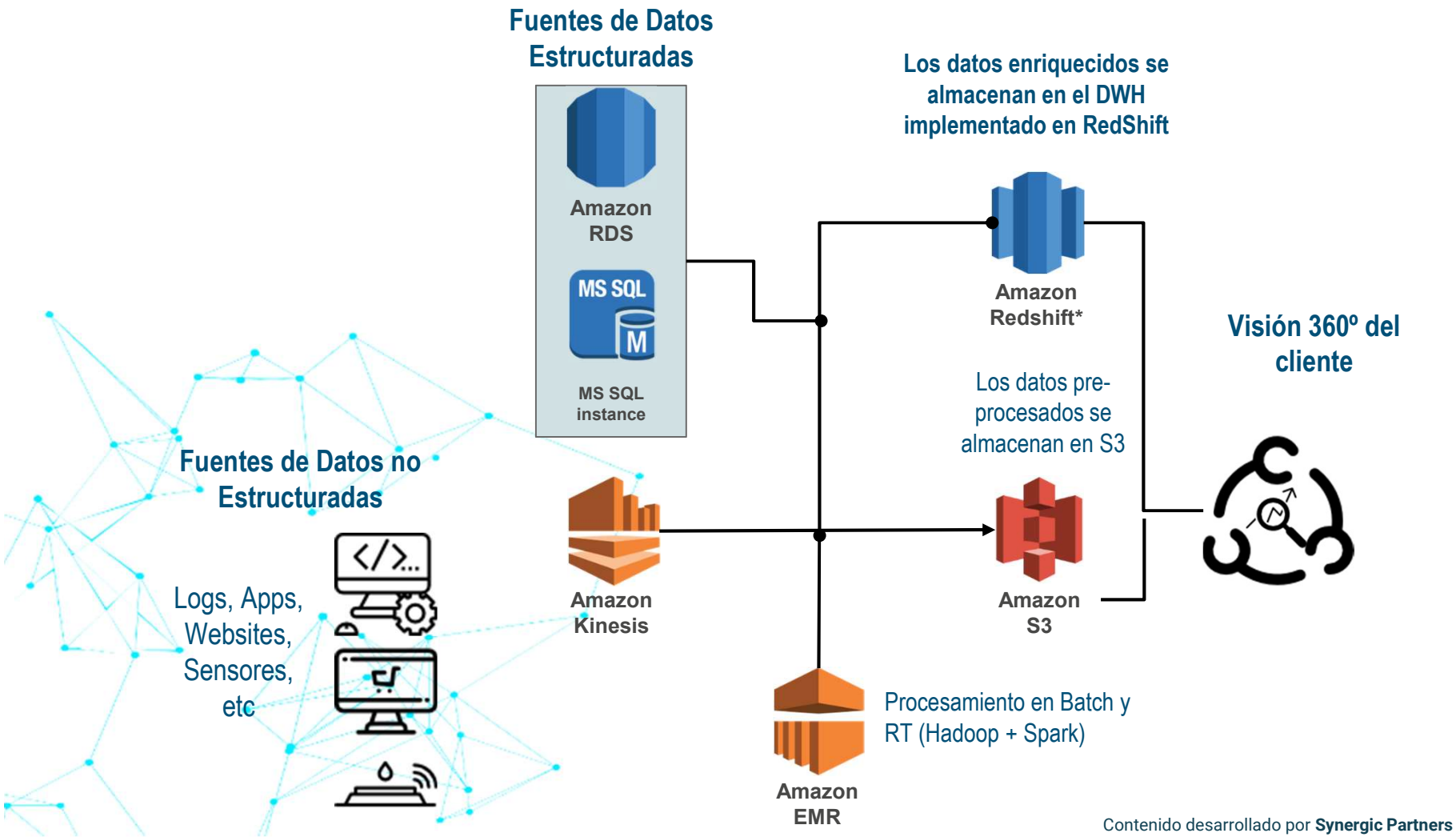
- Cognitive Services agrupa distintas APIs con modelos de Deep Learning ya implementados:
  - **Text Analytics:** modelos para análisis de texto.
  - **Recommendations API:** modelos para motores de recomendación de productos.
  - **Facial recognition:** modelos de reconocimiento facial.
  - **Computer vision:** modelos de procesamiento de imágenes.
  - **Bing Speech API:** modelos para convertir voz en texto y viceversa.

Caso de Uso - Visión 360º del cliente

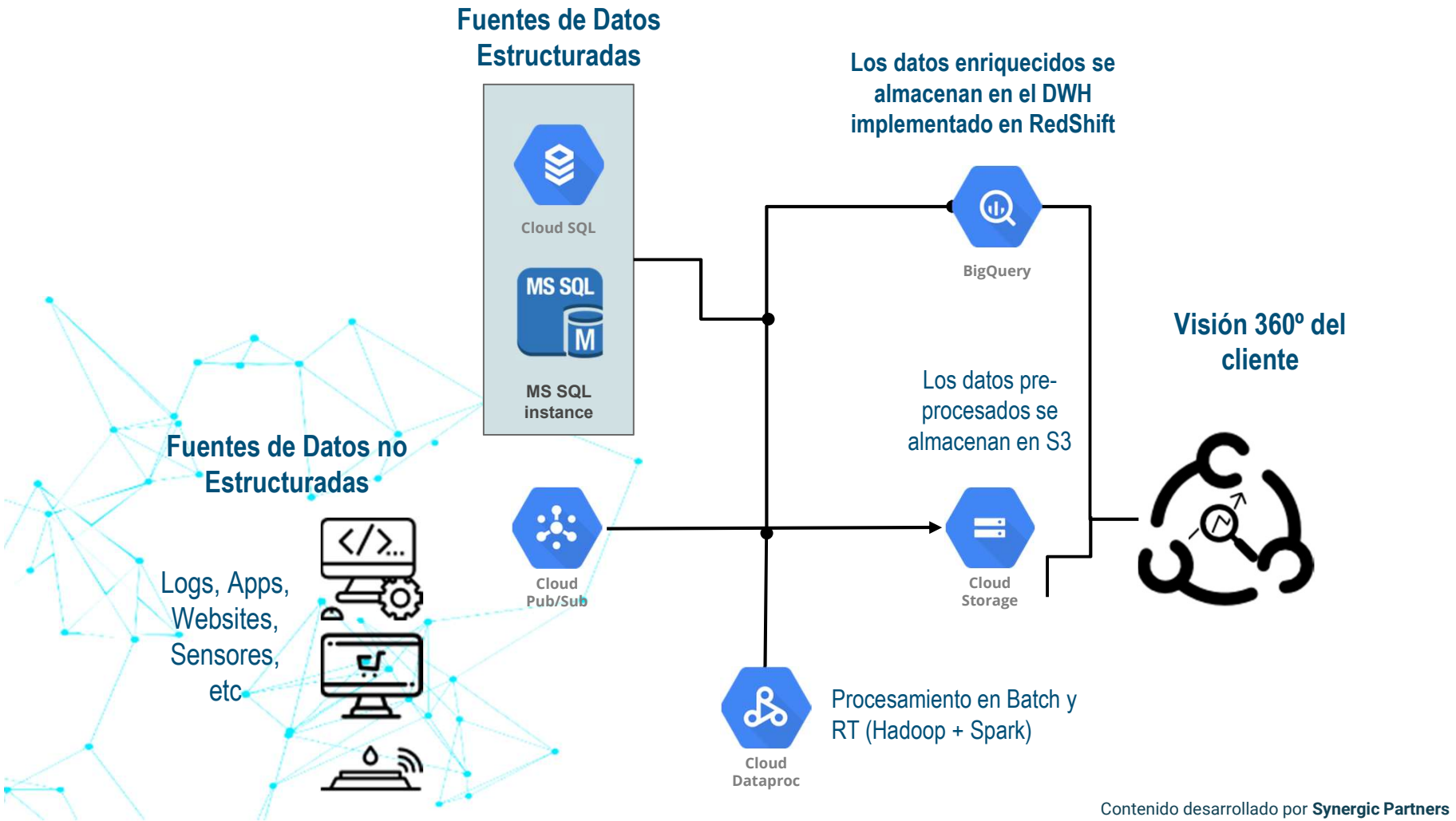
Situación
(mismo Caso de Uso explicado para AWS y Google):

- Hoy en día las empresas tienen una relación omni-canal con los clientes.
- Esto implica que obtienen información del cliente por varios canales: website, móvil, call center, redes sociales, oficinas físicas, entre otros.
- Para obtener una visión 360º del cliente, tienen que integrar todas estas fuentes de datos en un sólo sitio disponible a toda la organización para el posterior análisis de la información.
- A este sitio donde se integran todas estas fuentes de datos se le llama hoy en día **"Data Lake"** (Lago de datos).
- El siguiente diagrama muestra una posible implementación de Data Lake utilizando los servicios de Amazon Web Services.

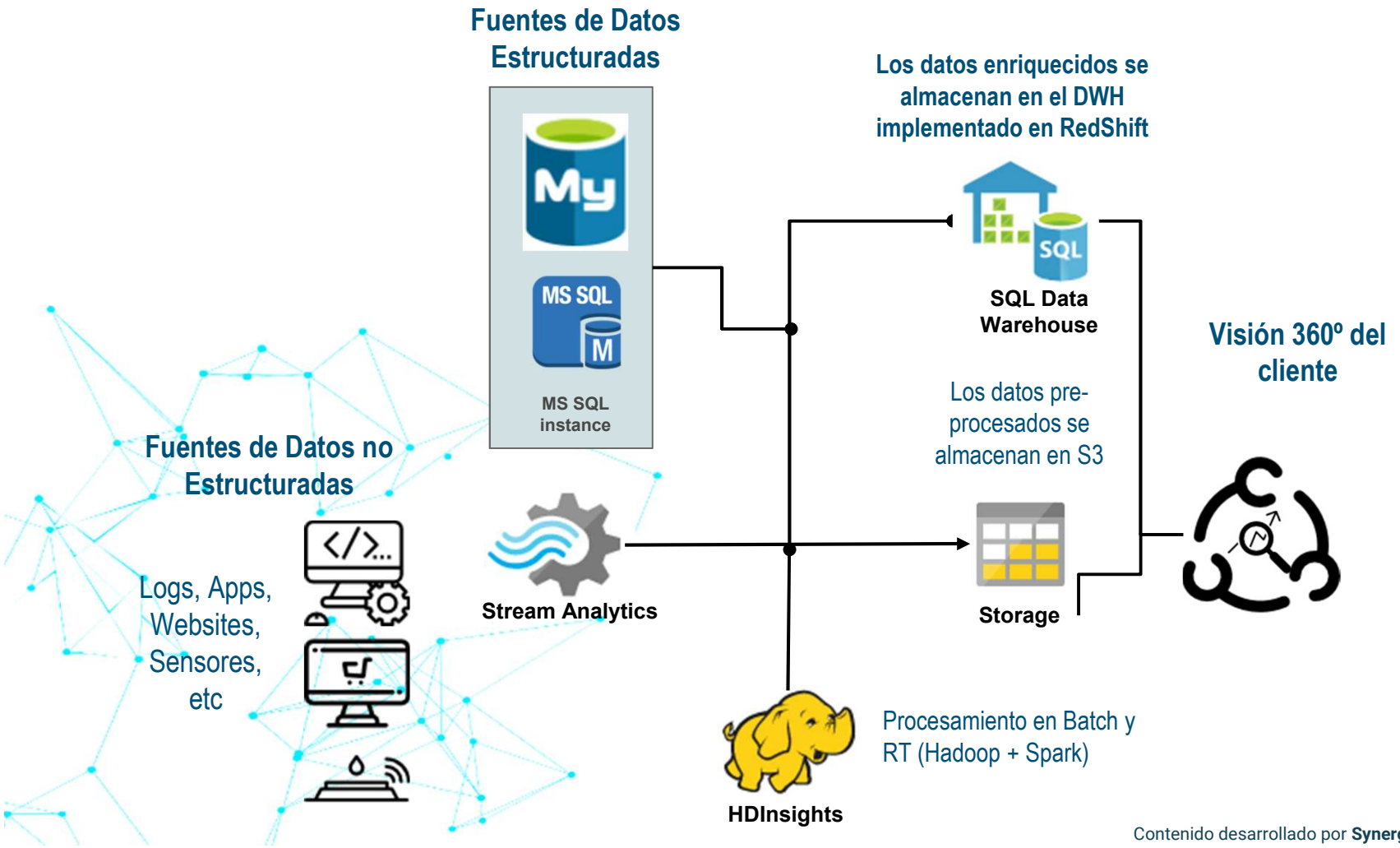
Caso de Uso - Visión 360º del cliente



Caso de Uso - Visión 360º del cliente



Caso de Uso - Visión 360º del cliente



TALLER

Contenido desarrollado por **Synergic Partners**