
SKILLING
CENTER
TECMILENIO

SKILLING CENTER TECMILENIO

Fundamentos del Big Data

Datos en la nube y el internet de las cosas

Introducción

Víctor es el primer científico de datos en una empresa de servicios financieros. El departamento de tecnologías de la información pretende implementar la ciencia de datos en su operación. Víctor espera ser bastante ágil para mejorar procesos a partir de los datos, por lo que ha decidido plantear el uso del *cloud computing* para sus proyectos. Él reconoce que las transacciones de depósitos, pagos por compras en comercio electrónico, amortizaciones de créditos, dispersiones, inversiones, tasas, plazos y demás elementos de la empresa son de gran volumen para procesarlos desde la computadora del trabajo.

Conoce en este tema las razones por las cuáles Víctor considera que el cloud computing le puede resolver sus propósitos, así como la seguridad y confidencialidad de la información.

Explicación

La computación en la nube son servicios de un tercero de almacenamiento, procesamiento, conectividad, seguridad, software, entre otros.

Google Cloud (s.f.)

- **Lo define como "la disponibilidad bajo demanda de recursos de computación como servicios a través de Internet".**

Salesforce (s.f.)

- **Lo describe como "una tecnología que permite acceso remoto a software, almacenamiento de archivos y procesamiento de datos por medio de Internet".**

Explicación

Este servicio evita utilizar recursos tanto físicos como humanos en la inversión en infraestructura, configurar sistemas de seguridad y conectividad, licenciamiento para uso de software y disponerlo desde cualquier dispositivo conectado por internet, además de que se paga por lo que consume o recursos utilizados a un proveedor de la nube.

Explicación

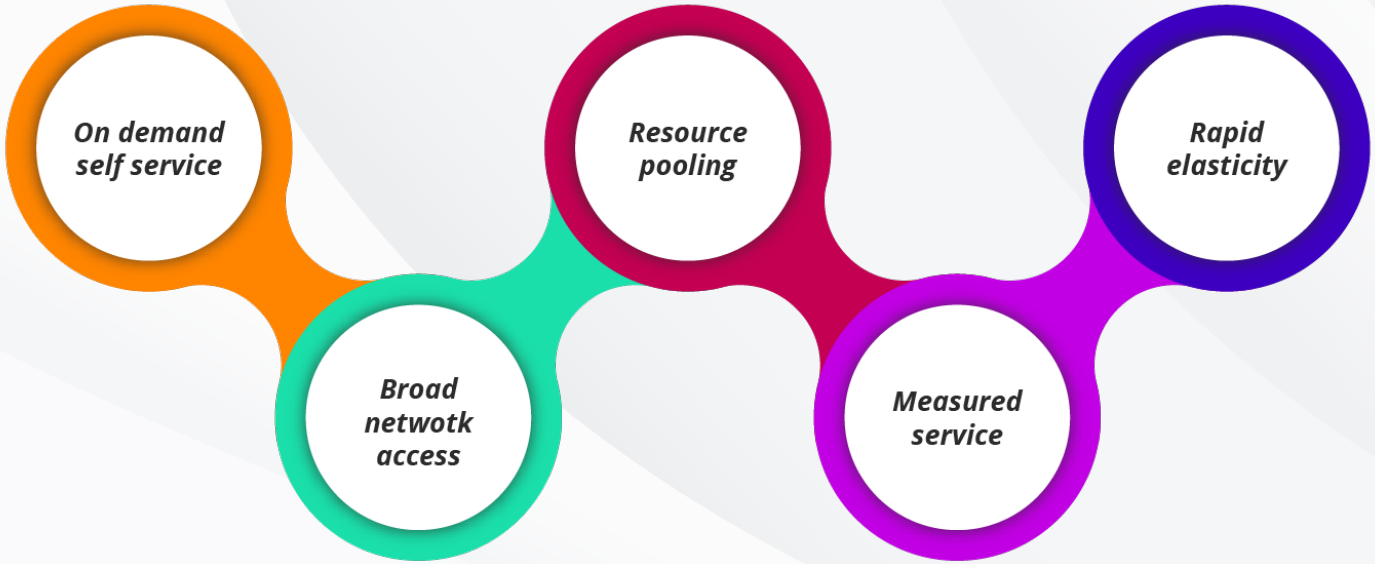
Los servicios del cloud computing se remontan al año 1950, cuando los *mainframes* o computadoras centralizadas de gran tamaño y altos costos, pero con capacidad de realizar operaciones en forma masiva, con limitaciones físicas por su tamaño, solo podían ser utilizados por una persona y su acceso era solo presencial.

Explicación

El desarrollo del modelo de negocio del cloud computing como se conoce hoy en día es el producto del avance tecnológico de la red de redes, Internet. Los servicios iniciaron con almacenamiento a gran escala, conectividad de red privada y disposición de información como Dropbox, y le siguieron los precursores de servicios con mayor alcance, aplicaciones empresariales y renta de computadoras virtuales como Salesforce, Amazon Web Services, Microsoft Azure, Apple iCloud, IBM, HP, Oracle, entre otros.

Explicación

Características del cloud computing



Fuente: Salesforce. (s.f.). *Cloud computing*. Recuperado de https://www.salesforce.com/mx/cloud-computing/

Explicación

Modelos de implementación

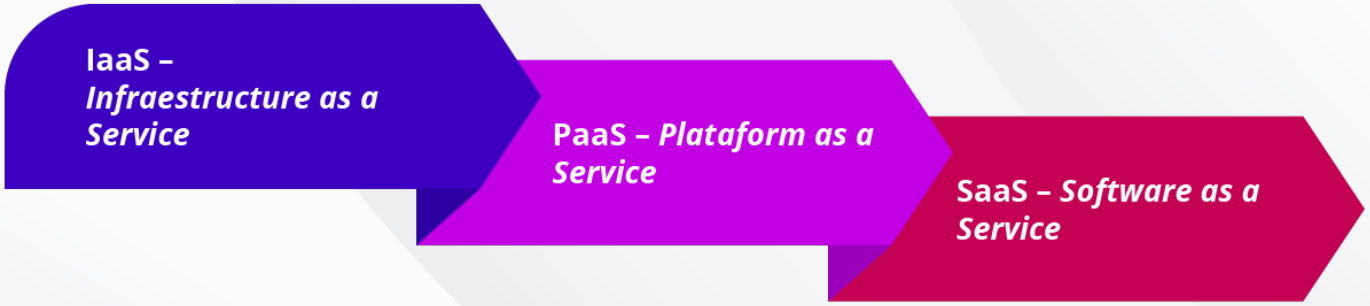
Son las formas de desplegar los servicios contratados en la nube.

Fuente: Salesforce. (s.f.). *Cloud computing*. Recuperado de https://www.salesforce.com/mx/cloud-computing/

Explicación

Modelos de servicio

Son los tipos de servicio que se pueden contratar en la nube.



Fuente: Salesforce. (s.f.). *Cloud computing*. Recuperado de https://www.salesforce.com/mx/cloud-computing/

La nube adecuada para cada empresa es aquella que se ajusta a sus necesidades y le permite cumplir con las regulaciones que le apliquen.

Explicación

Internet de las cosas

El cloud computing ha generado una variedad de valiosas tecnologías como el Big Data y el Internet de las cosas. Específicamente el Internet de las cosas (*Internet of Things*, IoT) se refiere a dispositivos físicos conectados a Internet. Oracle (s.f.) contempla en su definición que el propósito de la conexión a Internet de dichos dispositivos es el intercambio de datos con otros dispositivos y sistemas de Internet.

Explicación

Estos dispositivos no tienen capacidad de almacenamiento de datos, sino de transmitirlos a un servicio de cloud computing. Con la ayuda del IoT se entienden patrones de uso, desempeño y problemas más comunes, lo cual facilita la innovación y mejora de procesos. Alarmas, videocámaras, termostatos, electrodomésticos, monitores, smartwatches y muchos más dispositivos ahora pueden conectarse a internet para ser puestos a trabajar desde una aplicación.

Sin embargo, se deben resolver problemas como la encriptación de los datos, estandarización de la comunicación entre dispositivos y la privacidad de información sensible.

Actividad

Para reforzar tu comprensión del tema, contesta las siguientes preguntas:

1. Investiga y describe la manera en que el cloud computing puede ayudar a mejorar la movilidad urbana de la ciudad en donde vives. Menciona en general los objetivos del proyecto, el flujo general de información que incluye la generación, recolección, almacenamiento, procesamiento de la información, salidas, acciones en sitio y el aprendizaje del modelo con el tiempo de uso.
2. Considera un proceso productivo con el que estés familiarizado con su operación y menciona la conveniencia de aplicar el modelo de servicio en cloud computing denominado Software as a Service (SaaS).
3. Comenta los beneficios de utilizar el Internet de las cosas (IoT) en una línea de producción, específicamente en el tema del aseguramiento de la calidad.

Cierre

Ahora sabes cuáles son los beneficios que consideró Víctor para utilizar el cloud computing para sus proyectos. Estos son que puede adaptarse al volumen de procesamiento de información de la empresa financiera, no requiere inversión en infraestructura, no necesita mantenimiento físico por parte de la empresa, puede conectarse desde distintos dispositivos y le va a permitir tener transparencia sobre los recursos utilizados y sus costos.

Bibliografía

- Google Cloud. (s.f.). *¿Qué es cloud computing?* Recuperado de https://cloud.google.com/learn/what-is-cloud-computing?hl=es
- Oracle. (s.f.). *¿Qué es el IoT?* Recuperado de https://www.oracle.com/ar/internet-of-things/what-is-iot/
- Salesforce. (s.f.). *Cloud computing.* Recuperado de https://www.salesforce.com/mx/cloud-computing/

SKILLING CENTER TECMILENIO

Fundamentos del Big Data

Arquitectura y gobierno de Big Data

Introducción

La consultora en la que trabaja Miguel está realizando una evaluación sobre la estrategia de administrar con datos en una empresa financiera enfocada en microcréditos. Lo anterior se debe a que, en la empresa financiera, a pesar de que ya cuenta con modelos de *machine learning*, tableros que se actualizan diariamente y bases de datos donde la información por áreas del negocio está disponible para analítica, frecuentemente presenta errores en la información, fallas en el acceso, los reportes tienen retraso en fechas de entrega y están gastando más de lo presupuestado.

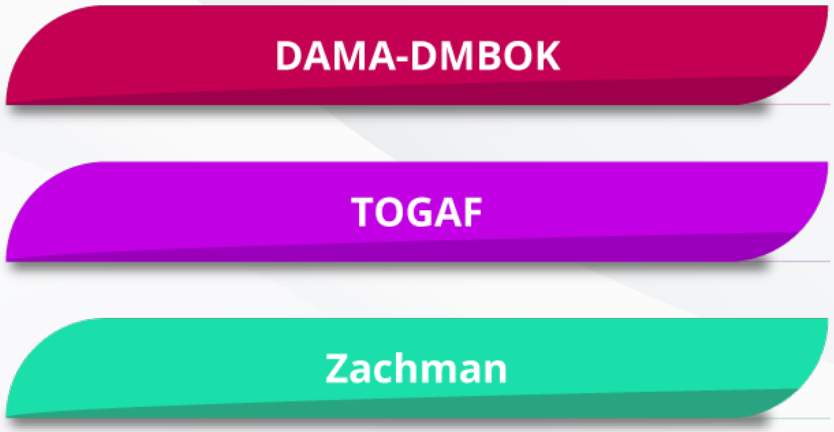
Miguel identificó dos aspectos importantes que están faltando a la empresa financiera: una clara y robusta arquitectura de datos y un eficiente gobierno de los datos. Únete a Miguel en la explicación que dará a la empresa sobre estos dos elementos que son pieza clave en la ciencia de datos.

Explicación

La arquitectura de Big Data

Consiste en la estructura que conecta la estrategia de datos con la estrategia del negocio. Para IBM (s.f.), la arquitectura de datos se encarga de cómo los datos son recolectados, transformados, distribuidos y consumidos por la organización. Esta estructura establece un plan para la gestión de datos y la manera en que estos pasan por los procesos de almacenamiento y análisis, denominado como flujo de datos (*data flow*).

Algunas metodologías más populares para la arquitectura de datos son:



Explicación

Por otro lado, la arquitectura de datos contempla tres tipos de modelos de datos:

Fuente: IBM. (s.f.). *What is data architecture?* Recuperado de https://www.ibm.com/topics/data-architecture

Explicación

Algunos proveedores de servicios que facilitan la configuración particular son Azure Data Factory, Amazon Web Services Step Functions y Google Cloud Functions.

Tanto las metodologías como los modelos se combinan según las necesidades de la organización, incluso el sistema puede tener más de una. Como componentes o capas del sistema se mencionan los siguientes (Yaseen y Obaid, 2020):

1. Fuentes de datos: incluye datos internos y externos.
2. Almacenamiento: datos guardados para futuro uso, considerando si son o no estructurados.
3. Almacenamiento analítico: datos ya procesados con herramientas analíticas.
4. Consumo: permite la interacción de los usuarios según sus necesidades.

Explicación

Los principales sistemas como fuentes de datos son (Dasgupta, 2018):

- Sistemas transaccionales
- Sitios web y aplicaciones
- Proveedores y terceros
- Sensores y otros componentes electrónicos

Explicación

Esquemas para el almacenamiento de información

1. *Repositorio con documentos, hojas de cálculo*: contiene solo datos estructurados generados por sistemas transaccionales.
2. *Data lake*: contiene datos en su formato original (*raw data*), como bases de datos relacionales, datos estructurados y no estructurados de bases NoSQL, datos semiestructurados como archivos en formato log y csv, archivos de imagen, video, audio, pdf.
3. *Data warehouse*: los datos están conectados entre sí con una estructura o clasificación determinada, que permite generar proyectos de analítica.
4. *Data mart*: similar que el data warehouse, pero los datos se clasifican para temas específicos del negocio, como finanzas, producción, mercadotecnia, etcétera.

Empresas como Microsoft, Amazon y Google con servicios de computación en la nube ofrecen la configuración propia para el negocio, incluyendo estimar los costos aproximados del servicio.

Explicación

Gobernanza del Big Data

Conforme avanza la adopción del Big Data en las organizaciones, se crean nuevas regulaciones que se tienen que respetar, por lo que se deben establecer tanto la autoridad como los procedimientos de control de la información que aseguren la calidad, seguridad, roles, facultades y medios de acceso y comunicación.

Explicación

Principales funciones de la gobernanza:

- Establecer estrategias de datos y políticas.
- Determinar los estándares de datos y la arquitectura adecuada al modelo de negocio.
- Asegurar el cumplimiento de normativas que aplican.
- Establecer mecanismos para la solución de problemas y mitigar riesgos de seguridad.
- Generar proyectos de gestión de datos.

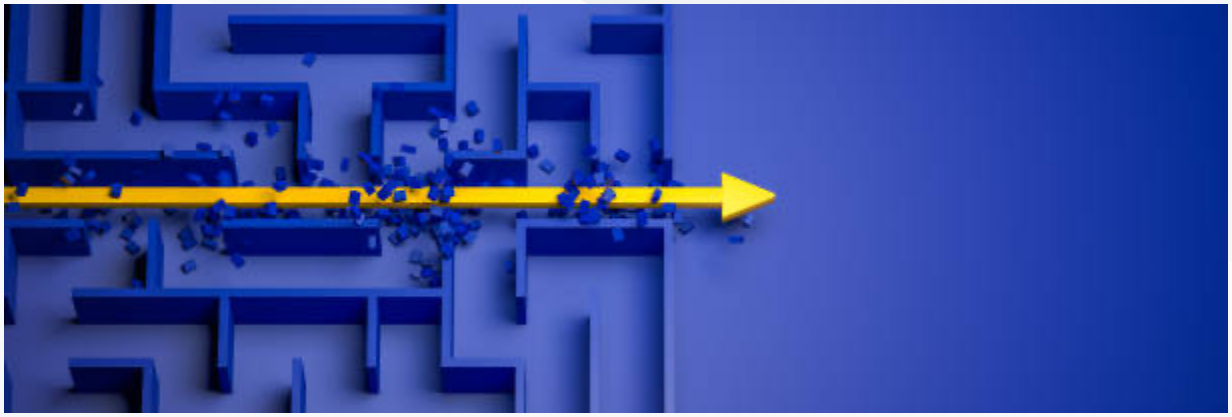
Google Cloud (s.f.) señala que el gobierno de datos “es imprescindible en todas las empresas de todos los sectores y tamaños, ya que los datos se han ido convirtiendo en su recurso más valioso al avanzar en su transformación digital”.

Explicación

Sin embargo, a pesar de reconocer las funciones y beneficios de la gobernanza de la información, en la realidad tiene importantes retos:

- La configuración de la gobernanza requiere de diversos servicios en plataformas de diferentes proveedores, además de crear la propia estructura adecuada para la empresa.
- Poco patrocinio por parte del negocio al no considerarlo como una prioridad en el presupuesto.
- Grado de complejidad en la implementación por falta de una cultura de ciencia de datos.

Como todo cambio o mejora, el viaje tiene su ciclo de vida y los resultados no son instantáneos. Se requiere de una fuerte y constante comunicación dentro de la empresa. Además, se puede recurrir a una estrategia de *data literacy* para facilitar la adopción de la ciencia de datos, sobre una base de gobernanza dentro de la organización.



Actividad

Para reforzar tu comprensión del tema, contesta las siguientes preguntas:

1. Describe las características generales que debe tener la arquitectura de datos en el área en donde laboras, considerando la información que utilizan. Elabora una tabla en donde clasifiques la información de entrada, su procesamiento y la salida, así como el medio o herramienta utilizada en cada etapa. Por último, menciona observaciones o recomendaciones para la mejora en el proceso.
2. Establece en general las características o funciones de la gobernanza de datos que aplican para el proceso analizado en el punto anterior.
3. Investiga y menciona en general la utilidad de la información pública que puedes disponer del censo de población y vivienda del Instituto Nacional de Geografía y Estadística (INEGI).

Cierre

Miguel ha finalizado su explicación y recomienda a la financiera trabajar en una arquitectura especialmente diseñada según sus necesidades, donde se incorpore un *data lake* y un *data warehouse*, debido a que la empresa puede crear valor al analizar la variedad de datos que recolectan de sus clientes y que los *data marts* necesitan centralizar la información.

También ha incluido en su recomendación desarrollar la gobernanza de datos para que estas nuevas estructuras se utilicen bajo el marco regulatorio y se puedan mejorar las prácticas que venían ocurriendo.

Bibliografía

- Dasgupta, N. (2018). *Practical Big Data analytics: hands-on techniques to implement enterprise analytics and machine learning using Hadoop, Spark, NoSQL and R*. Packt Publishing.
- Google Cloud. (s.f.). *¿Qué es el gobierno de datos?* Recuperado de https://cloud.google.com/learn/what-is-data-governance?hl=es
- IBM. (s.f.). *What is data architecture?* Recuperado de https://www.ibm.com/topics/data-architecture.
- Yaseen, H., y Obaid, A. (2020). *Big Data: definition, architecture & applications. International Journal on Informatics Visualization, 4*(1). Recuperado de http://dx.doi.org/10.30630/joiv.4.1.292

Tecmilenio no guarda relación alguna con las marcas mencionadas como ejemplo. Las marcas son propiedad de sus titulares conforme a la legislación aplicable, estas se utilizan con fines académicos y didácticos, por lo que no existen fines de lucro, relación publicitaria o de patrocinio.
