Campus Romero

Big Data

FUNDAMENTOS

Nombre del profesor

Campus Romero

ANTECEDENTES

El término Big Data nace en 1989 cuando Erick Larson lo utiliza por primera vez en un artículo sobre marketing, en éste escribe sobre los datos de los clientes y cómo se usarían. Ciertamente, obtiene una trascendencia social a partir de los años 90`s ya que nace el primer navegador web (www) “World wide web". A partir de esto, con la llegada del Internet se creó una red de conexión en el mundo y cualquier persona podía subir datos; convirtiéndose así en la primera generación de datos masivos.

Campus Romero

DEFINICIÓN

Es un conjunto de procesos, tecnologías y modelos basados en el almacenamiento masivo de datos, procesamiento y transformación de los mismos en conocimiento, para analizar lo que sucederá en un mundo complejo con muchas interacciones.

Campus Romero

LAS 5V

OTROS CONCEPTOS

LA PARALELIZACIÓN

Se refiere a dividir los datos en archivos más pequeños. Estos archivos más pequeños son enviados cada uno a una máquina diferente. De esta forma, cada máquina procesa una pequeña parte del fichero inicial en lugar de analizar el fichero completo.

LA ESCALABILIDAD

Indica su habilidad para reaccionar y adaptarse sin perder calidad, o bien manejar el crecimiento continuo de trabajo de manera fluida, o bien para estar preparado para hacerse más grande sin perder calidad en los servicios ofrecidos. Existen dos tipos de escalabilidad: la vertical y la horizontal.

Campus Romero

OTROS CONCEPTOS

LA ALTA DISPONIBILIDAD

Consiste en una serie de medidas tendientes a garantizar la disponibilidad de un servicio, es decir, asegurar que el servicio funcione durante las veinticuatro horas.

LA ENCRIPTACIÓN

Es el proceso de ofuscar datos mediante el uso de una clave o contraseña que consigue que, quienes accedan a ellos sin el password adecuado, no puedan encontrar ninguna utilidad en los mismos, puesto que resulta imposible descifrar su contenido

Campus Romero

TECNOLOGíAS ASOCIADAS

Campus Romero

OTRAS TECNOLOGíAS

HADOOP HDFS, GOOGLE CLOUD STORAGE Y AMAZON S3

Si queremos guardar grandes cantidades de datos, sean estructurados, semiestructurados o no estructurados

APACHE HIVE, APACHE SPARK, GOOGLE DATAPROC

Si queremos procesar gran cantidad de datos de forma batch o por lotes con una sintaxis SQL

APACHE SPARK

Si queremos procesar los datos en tiempo real y realizar procesos de aprendizaje de máquina

APACHE HBASE, GOOGLE CLOUD BIG TABLE, CASSANDRA, MONGODB Y COUCHDB

Si queremos hacer consultas de baja latencia utilizando una base de datos no relacional

Campus Romero

OTRAS TECNOLOGíAS

SQOOP

Si queremos extraer datos de una base de datos relacional y colocarlos en Hadoop

KAFKA, CLOUD PUB/SUB,CLOUD DATAFLOW

Si queremos extraer datos de otras fuentes y/o en tiempo real

ARTIFICIAL INTELLIGENCE GOOGLE PLATFORM, SPARK MACHINE LEARNING, SCALA, PYTHON, FRAMEWORKS DE ANÁLISIS DE DATOS COMO: TENSORFLOW, KERAS, PYTORCH

Si queremos realizar análisis de datos en entornos de Big Data

OTRAS TECNOLOGíAS

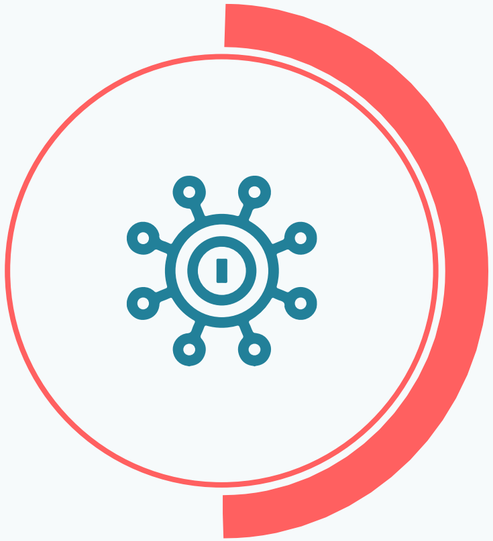
SCALA, PYSPARK

Si queremos implementar sistemas simples de enriquecimiento de datos

JAVA, SCALA

Si queremos implementar sistemas complejos sobre Big Data

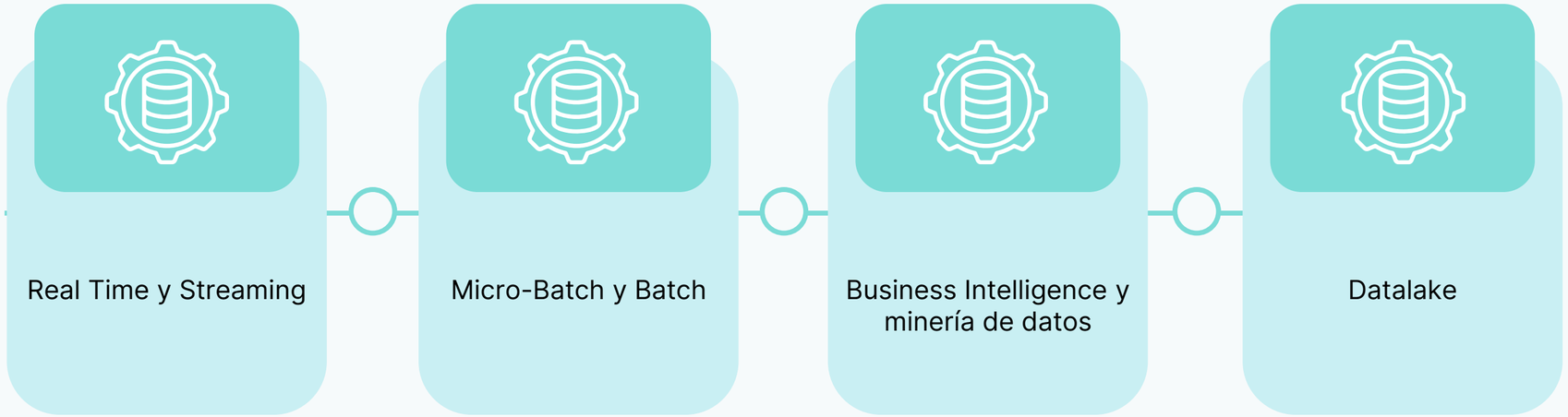
SOLUCIONES BASADAS EN BIG DATA



Desde 2011, año en el que comienza a oírse hablar del fenómeno, se ha vivido una constante evolución tecnológica enfocada a desarrollar la capacidad de las tradicionales herramientas de software de base de datos, incapaces de gestionar el creciente volumen de datos relativo a la actividad de los usuarios y los procesos de negocio de las diferentes empresas.

Campus Romero

SOLUCIONES BASADAS EN BIG DATA



Campus Romero

IMPORTANCIA

Lo que hace que Big Data sea tan útil para muchas empresas es el hecho de que proporciona respuestas a muchas preguntas que las empresas ni siquiera sabían que tenían. En otras palabras, proporciona un punto de referencia. Con una cantidad tan grande de información, los datos pueden ser moldeados o probados de cualquier manera que la empresa considere adecuada. Al hacerlo, las organizaciones son capaces de identificar los problemas de una forma más comprensible.

Campus Romero

BENEFICIOS

PRINCIPALES CAMBIOS EN LAS EMPRESAS

Ningún dato se perderá, esto se debe a que recopilan mucha información todos los días las empresas, pero se usa una pequeña proporción de esta. Los grandes volúmenes de datos pueden ser una mina de oro si se analizan de forma efectiva para extraer insights.

La migración a la nube es una decisión impostergable ya que la incorporación de soluciones asociadas a Cloud Computing ya no es una elección; por el contrario es algo esencial para las empresas.

El uso de Inteligencia Artificial y de Machine Learning se extenderá. Este dúo tecnológico que transforma grandes datos aparentemente difíciles de procesar en información simplificada y comprensible les permitirá a las compañías optimizar su rendimiento.

Campus Romero

PRINCIPALES CAMBIOS EN LAS EMPRESAS

Los Chief Data Officers (CDO) tomarán un rol central en las organizaciones. El Director de Datos se convirtió en una figura predominante. Se espera que cada vez más organizaciones promuevan a Directores de Datos en puestos jerárquicos, ya que sus decisiones pueden ser fundamentales para potenciar los resultados de negocio.

La computación cuántica se asoma, estas son máquinas súper potentes que se rigen por los principios de la mecánica cuántica, capaces de realizar análisis de proporciones de datos impensados, ganarán protagonismo.

Fundación Romero

www.fundacionromero.org.pe