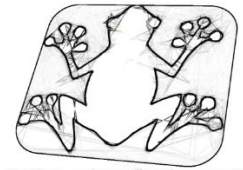


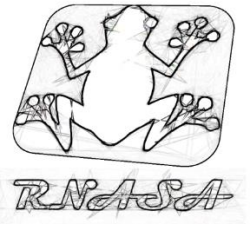
Fundamentos del Big Data

Conceptos Básicos

Grupo de Investigación RNASA-IMEDIR

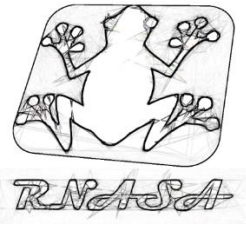
¿Qué es Big Data?

Hechos (en un minuto)



- Email users send more than 204 million messages
- Mobile Web receives 217 new users
- Google receives over 2 million search queries
- YouTube users upload 48 hours of new video
- Facebook users share 684,000 bits of content
- Twitter users send more than 100,000 tweets
- Apple receives around 47,000 application downloads
- Brands receive more than 34,000 Facebook 'likes'
- Instagram users share 3,600 new photos

¿Qué es big data?

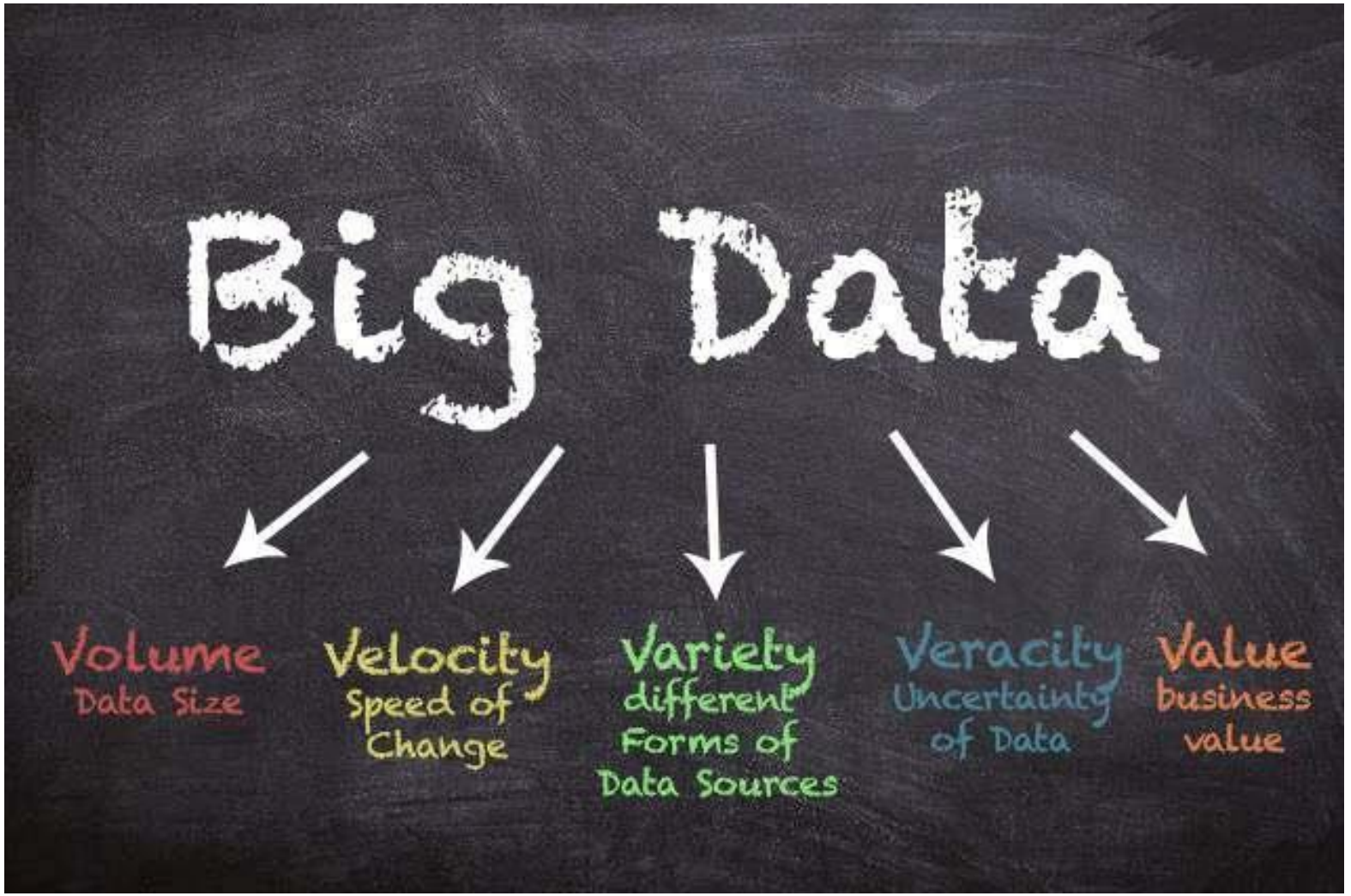


- Volúmenes de datos masivos que no pueden ser tratados mediante técnicas convencionales:

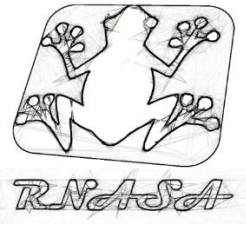
  - Captura
  - Transferencia
  - Almacenamiento
  - Gestión / Mantenimiento / Consulta
  - Análisis (Extracción de conocimiento)
  - Visualización

¿Qué es Big Data?

Características

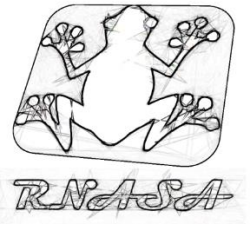


¿Qué es Big Data?
Características



- Las V’s del Big Data:
  - Volumen → Gran cantidad de datos
  - Variedad → Diferente naturaleza
  - Velocidad → Captura / Análisis / Crecimiento
  - Veracidad → Calidad de los datos
  - Valor → Análisis / Extracción de conocimiento

¿Qué es Big Data?
Características



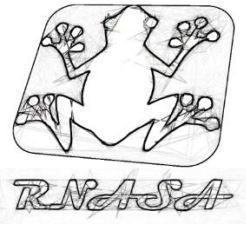
- Volumen

  - Se refiere al gran volumen de información que se maneja.

  - Los datos se acumulan con un crecimiento exponencial, requiriendo ampliar continuamente el almacenamiento de datos.

  - Cuando se habla de bases de datos masivas se refiere a magnitudes del orden de petabytes o exabytes

¿Qué es Big Data?
Características



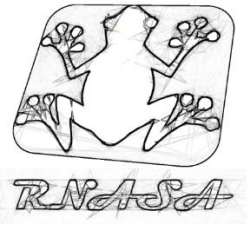
- Variedad

  - Necesidad de agregar información procedente de una amplia variedad de fuentes de información independientes: redes sociales, sensores, máquinas o personas individuales.

  - En general son datos desestructurados, así como gráficos, texto, sonido o imágenes.

  - Estos datos no pueden gestionarse fácilmente con bases de datos relacionales y las herramientas de inteligencia de negocio tradicionales.

¿Qué es Big Data?
Características

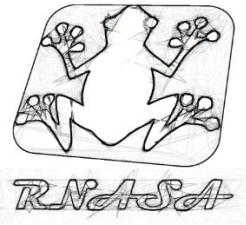


- Velocidad

  - Es la enorme velocidad en la generación, recogida y proceso de la información.

  - Hay que ser capaz de almacenar y procesar en tiempo real millones de datos generados por segundo por fuentes de información tales como sensores, cámaras de videos, redes sociales, blogs, páginas webs,...

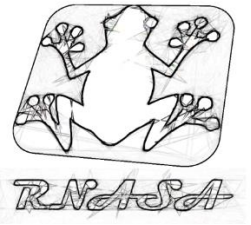
¿Qué es big data?
Características



- Veracidad

  - Se debe analizar inteligentemente un gran volumen de datos con la finalidad de obtener una información verídica y útil que nos permita mejorar nuestra toma de decisiones.

¿Qué es Big Data?
Características



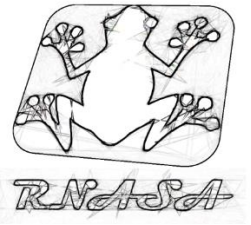
- Valor

  - Es la creación de una ventaja competitiva al identificar y procesar los datos claves, permitiendo así, por ejemplo:

    - Monetizar los datos.
    - Obtener nuevos clientes.
    - Generar fidelidad.
    - Reducir costes.
    - Mejorar la imagen de marca

DIFERENCIAR BIG DATA DE...

BUSSINESS INTELLIGENCE

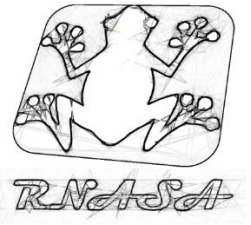


- Big Data se puede entender como una evolución del concepto Business Intelligence.

- En Business Intelligence:
  - Se captura información de una organización y tras un análisis, se obtienen resultados que tienen la finalidad de ayudar a la toma de decisiones estratégicas en la empresa.
  - La información proviene de fuentes de datos estructuradas.
  - Los datos se agrupan en un servidor central y se analizan de forma offline, estructurándose en una base de datos relacional.

DIFERENCIAR BIG DATA DE...

BUSSINESS INTELLIGENCE

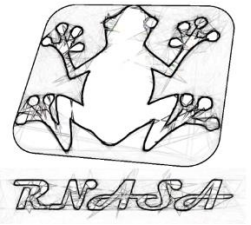


- En Big Data:

  - Se procesa información no estructurada: lenguaje natural, información de redes sociales, información sensores, dispositivos (wearables), etc
  - Los datos se almacenan de forma distribuida que permiten manejar cantidades más grandes de información de forma más ágil.
  - Se emplea procesamiento paralelo masivo de datos, mejorando la velocidad del análisis.

DIFERENCIAR BIG DATA DE...

DATA WAREHOUSE

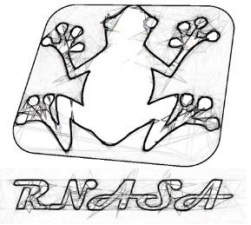


•Un **Data Warehouse** es un almacén de datos, creado con el fin de permitir la toma de decisiones en la organización.

• Una **solución Big Data** es al mismo tiempo almacén de datos y una tecnología de proceso, análisis y visualización de datos.

DIFERENCIAR BIG DATA DE...

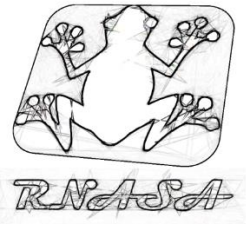
DATA MINING



• Big Data y minería de datos se relacionan por el uso de grandes conjuntos de datos para su procesamiento y análisis, sin embargo, divergen en su operativa.

• La minería de datos busca información concreta accediendo a partes pequeñas y específicas de los datos dentro de esos grandes conjuntos.

Análisis en el Big Data

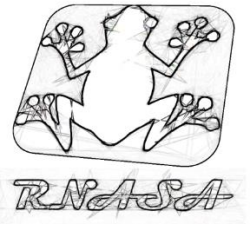


• La cuestión clave no es tener la capacidad para recolectar y almacenar una gran cantidad de datos.

• Con la acumulación de los datos no sé alcanzan ventajas competitivas: es necesario saber organizarlos, refinarlos, y convertirlos en información relevante que permita ganar posiciones en el mercado.

• Los datos tienen sólo valor potencial, es su análisis y sistematización el que permite incrementar la capacidad de innovar y obtener ventajas en las organizaciones.

Análisis en el big data



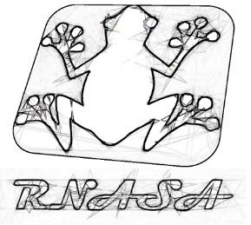
• Estas afirmaciones, nos lleva a hacer preguntas y encontrar respuestas, para la empresa y la sociedad.

• La gestión correcta de los datos genera una conciencia en las administraciones, empresas y ciudadanos, que los datos y su análisis son un activo de las sociedades modernas.

• Es evidente que se tienen que desarrollar nuevos perfiles para cuidar y sacar el máximo de esos activos.

PERFILES PROFESIONALES

LOS 7 PERFILES PROFESIONALES DEL BIG DATA

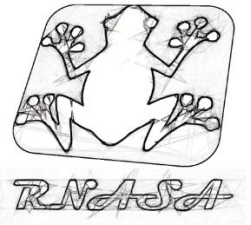


• **Chief Data Officer** (CDO): es el responsable de asegurar que la organización es data driven. Lidera la gestión de datos y analítica asociada por el negocio y, por tanto, es responsable de los diferentes equipos especialidades en datos.

• **Data Scientists** (científico de los datos): son los miembros clave del equipo de ciencia de datos. Permiten extraer conocimiento e información valiosa de los datos. Tienen visión general del proceso de extremo a extremo y pueden resolver problemas de ciencias datos, la construcción de modelos analíticos y algoritmos.

PERFILES PROFESIONALES

LOS 7 PERFILES PROFESIONALES DEL BIG DATA

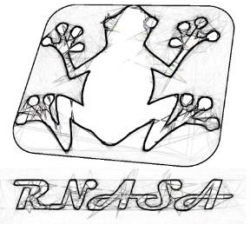


• **Citizen Data Scientist**: puede extraer valor, a través de su experiencia, explorando los datos, desde las unidades de negocio. Pueden ejecutar una serie simple de tareas analíticas utilizando herramientas de descubrimiento de datos.

• **Data Engineer** (ingeniero de datos): Se encarga de proporcionar los datos de una manera accesible y apropiada a los usuarios y Data scientists. Es un perfil especializado en infraestructura big data. Desarrolla y explota técnicas, procesos, herramientas y métodos que deben servir para el desarrollo de aplicaciones big data.

PERFILES PROFESIONALES

LOS 7 PERFILES PROFESIONALES DEL BIG DATA



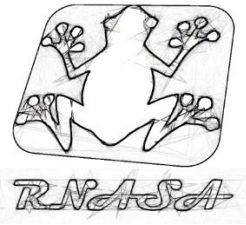
• **Data Steward** (administrador de datos): es responsable de mantener la calidad, disponibilidad y seguridad de los datos.

• **Business Data Analyst** (analista de datos): participa en las iniciativas y proyectos de análisis de datos. Es la persona que recoge las necesidades de los usuarios de negocio para los Data Scientist y presenta resultados obtenidos.

• **Data Artist:** son los responsables de crear los gráficos, infografías y otras herramientas visuales para ayudar a las diferentes personas a comprender datos complejos.

NECESIDADES

NUEVA INFRAESTRUCTURA TECNOLÓGICA



• La capacidad de aportar escalamiento de procesamiento masivo permite la identificación continua de información útil sepultada dentro de Big data.

• Integrar metodologías y tecnología para el descubrimiento y entendimiento de información basado en fuentes altamente escalables. Por ejemplo, Open Data, Linked Data, Social Data, Sentiment Analysis, Online Stream Analysis, Web Intelligence.

NECESIDADES

NUEVA INFRAESTRUCTURA TECNOLOGICA

Para ello es necesario que:

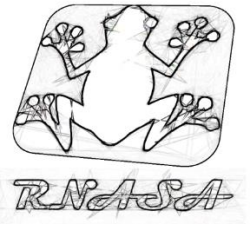
- Sea escalable de forma masiva a petabytes de datos (en la actualidad).

- Soporte y acceso a datos de baja latencia y toma de decisiones

- Tenga análisis integrado para acelerar el modelado de análisis avanzado y de los procesos.

NECESIDADES

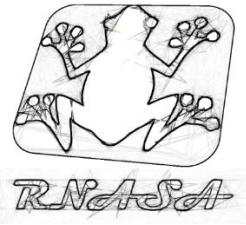
NUEVA INFRAESTRUCTURA TECNOLÓGICA



• Identificar las oportunidades de transformación y generación de valor basadas en el análisis de los datos, proveniente tanto de fuentes internas como externas a la organización.

• Desarrollar soluciones que permitan generar valor añadido y diferenciación a partir de los procesos de análisis de información sobre Big Data.

Riesgos en el Análisis de datos



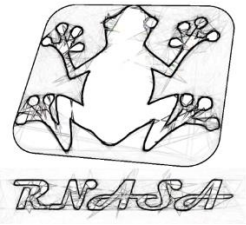
• Uno de los riesgos que presenta la búsqueda de información en el Big Data, es el descubrimiento de patrones no significativos.

• Estos patrones no relevantes se conocen en la estadística como principios de Bonferroni.

• Una gran cantidad datos como los que se analizan en los entornos Big Data, pueden “validar” cualquier patrón.

DATA MINING

ANÁLISIS DE LA INFORMACIÓN

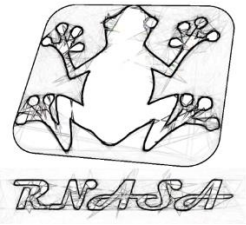


- Técnicas interdisciplinares (Inteligencia Artificial / Machine Learning) en entornos de procesamiento distribuido:

  - Redes de Neuronas Artificiales (ANN / DL)
  - Árboles de Decisión (DT)
  - Regresión (Simple / Multiple)
  - Redes Bayesianas
  - Support Vector Machines (SVM)

Tipos de Análisis

Modelos predictivos



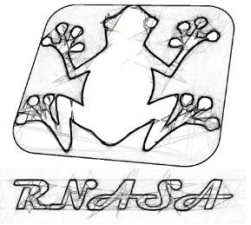
- **Evalúan qué probabilidad tiene un individuo de mostrar un comportamiento específico en el futuro.**

- **Buscan patrones discriminadores** en los datos para responder comportamiento.

- **Realizan cálculos en tiempo real** para evaluar un determinado riesgo u oportunidad, a fin de orientar una decisión adecuada.

TIPOS DE ANÁLISIS

MODELOS PREDICTIVOS



• Describen las relaciones entre los datos para poder clasificar a los individuos en grupos.

• Identifican diferentes relaciones entre individuos que pueden ser utilizadas para predecir también acciones futuras.

• Describen la relación entre todos los elementos de una decisión, la decisión a tomar y las variables y valores que determinan la propia decisión, con la finalidad de predecir los resultados mediante el análisis de muchas variables.