TEC | Tecnológico de Costa Rica

Minería de datos

Tecnológico de Costa Rica

Programa de Ciencia de Datos

Frans van Dunné

Agenda

TEC | Tecnológico de Costa Rica

- 8:00 – 9:30
  - Data warehouse vs Data Lake
- 9:20 – 9:35
  - Pausa
- 9:35 – 12:00
  - Data Lake
    - Características
    - Funciones

¿Qué es Big Data?

TEC | Tecnológico de Costa Rica

- Evolución del BI tradicional
- Nueva forma de trabajar con los datos
- Apoyado por nuevas herramientas
- ¿Cuando nace? Desde siempre…
- Cuando se hace conocido? Desde que salen herramientas fuera de lo tradicional…

¿Qué es Big Data?

TEC | Tecnológico de Costa Rica

Data Lake

TEC | Tecnológico de Costa Rica

- Un **data lake** es un repositorio de almacenamiento que contienen una gran cantidad de datos en bruto y que se mantienen allí hasta que sea necesario. A diferencia de un data warehouse jerárquico que almacena datos en ficheros o carpetas, un data lake utiliza una arquitectura plana para almacenar los datos.

¿Cuáles son los beneficios de un data lake?

TEC | Tecnológico de Costa Rica

- El principal beneficio de un data lake es la centralización de fuentes de contenido dispares. Una vez reunidas (de sus "silos de información"), estas fuentes pueden ser combinadas y procesadas utilizando big data, búsquedas y análisis que de otro modo hubieran sido imposibles. Las fuentes de contenido dispares a menudo contienen información confidencial que requerirá la implementación de las medidas de seguridad apropiadas en el data lake.

- Una vez que el contenido está en el data lake, puede normalizarse y enriquecerse. Esto puede incluir extracción de metadatos, conversión de formatos, aumento, extracción de entidades, reticulación, agregación, des-normalización o indexación.

Principales diferencias entre Data Lakes y Data Warehouses

TEC | Tecnológico de Costa Rica

- **Una Data Lake conserva todos los datos**
  - Durante el desarrollo de un data warehouse, se gasta una cantidad considerable de tiempo analizando las fuentes de datos, entendiendo los procesos de negocio y perfilando los datos. El resultado es un modelo de datos altamente estructurado diseñado para la generación de informes. Una gran parte de este proceso incluye tomar decisiones sobre qué datos incluir y no incluir en el almacén.

  - Generalmente, si los datos no se utilizan para responder a preguntas específicas o en un informe definido, pueden excluirse del almacén. Esto se hace generalmente para simplificar el modelo de datos y también para conservar el costoso espacio en el almacenamiento de disco que se utiliza para hacer el data warehouse.

Principales diferencias entre Data Lakes y Data Warehouses

TEC | Tecnológico de Costa Rica

- **Una Data Lake conserva todos los datos**
  - El data lake conserva todos los datos. No sólo los datos que se utilizan actualmente, sino los datos que se pueden utilizar e incluso los datos que nunca se van a ser utilizados sólo porque quizás podrían ser utilizados algún día. Los datos también se mantienen todo el tiempo para que podamos volver en el tiempo a cualquier punto para hacer el análisis.

  - Este enfoque se hace posible porque el hardware para un data lake suele ser muy diferente del utilizado para un data warehouse. La ampliación de un data lake a terabytes y petabytes puede hacerse de manera bastante económica.

Principales diferencias entre Data Lakes y Data Warehouses

TEC | Tecnológico de Costa Rica

- **Un Data Lake soporta todos los tipos de datos**
  - Los data warehouses generalmente se componen de datos extraídos de sistemas transaccionales junto con métricas cuantitativas y los atributos que las describen. Las fuentes de datos no tradicionales, como los registros del servidor web, los datos de sensores, la actividad de las redes sociales, el texto y las imágenes, se ignoran en gran medida. Se siguen encontrando nuevos usos para estos tipos de datos, pero consumirlos y almacenarlos puede ser costoso y difícil.

  - El enfoque del data lake abarca estos tipos de datos no tradicionales. En el data lake, guardamos todos los datos independientemente de la fuente y la estructura. Los mantenemos en su forma bruta y sólo los transformamos cuando estamos listos para usarlos. Este enfoque se conoce como "Schema on Read" en comparación con el "Schema on Write" que es el enfoque utilizado en el data warehouse.

Principales diferencias entre Data Lakes y Data Warehouses

TEC | Tecnológico de Costa Rica

- **Un Data Lakes soporta a todos los usuarios**
  - En la mayoría de las organizaciones, el 80% o más de los usuarios son "operacionales". Quieren obtener sus informes, ver sus KPIs o seleccionar el mismo conjunto de datos en una hoja de cálculo todos los días. El data warehouse suele ser ideal para estos usuarios porque está bien estructurado, fácil de usar y comprender y está diseñado para responder a sus preguntas.

  - El siguiente 10% más o menos, hace más análisis en esos datos. Utilizan el data warehouse como una fuente, pero a menudo vuelven a los sistemas de origen para obtener datos que no están incluidos en el almacén y a veces traen datos de fuera de la organización.

Principales diferencias entre Data Lakes y Data Warehouses

TEC | Tecnológico de Costa Rica

- **Los Data Lakes se adaptan fácilmente a los cambios**
  - En el data lake, por otro lado, como todos los datos se almacenan en bruto y siempre con accesibles a alguien que necesite utilizarlos, los usuarios tienen el poder de ir más allá de la estructura del almacén para explorar datos de nuevas maneras y responder a sus preguntas a su ritmo.

Principales diferencias entre Data Lakes y Data Warehouses

TEC | Tecnológico de Costa Rica

- **Los Data Lakes proporcionan una visión más rápida**
  - Debido a que los data lakes contienen todos los datos y tipos de datos, y a que permite a los usuarios acceder a los datos antes de que se hayan transformado, limpiado y estructurado, permite a los usuarios llegar a sus resultados más rápido que el método tradicional de data warehouse.

Principales diferencias entre Data Lakes y Data Warehouses

TEC | Tecnológico de Costa Rica

Principales diferencias entre Data Lakes y Data Warehouses

TEC | Tecnológico de Costa Rica

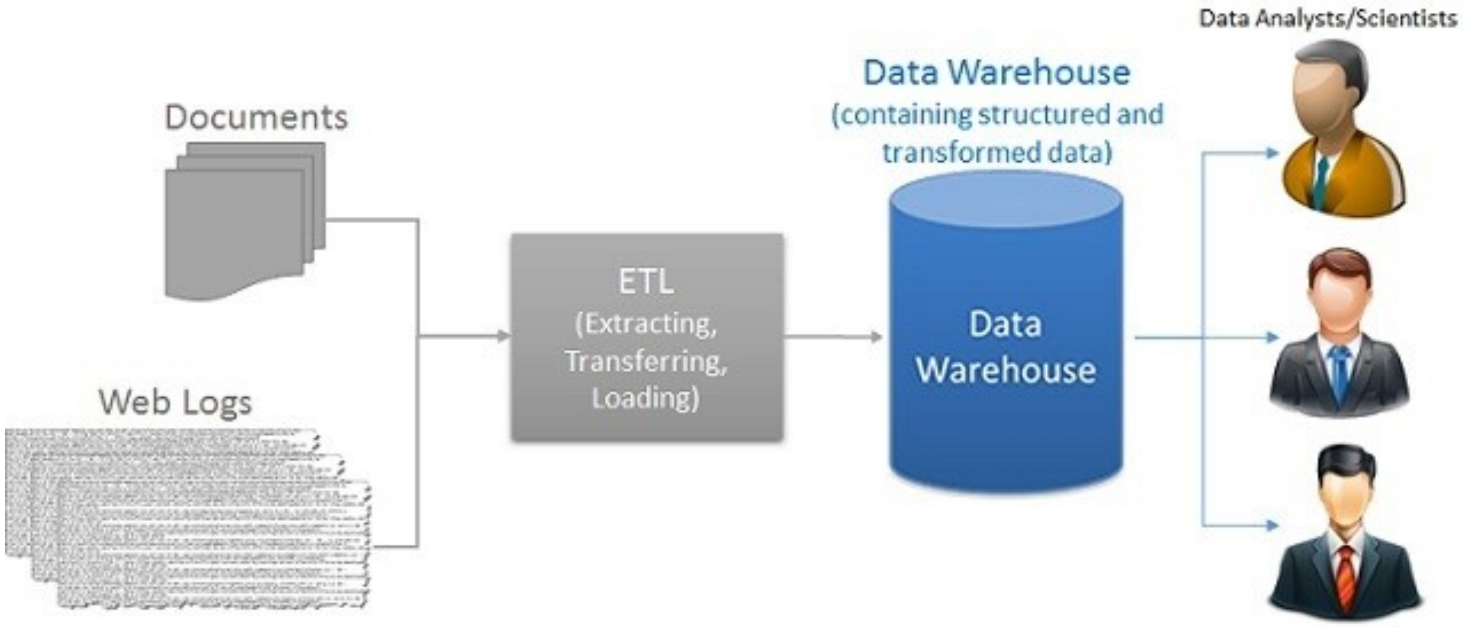
	DATA LAKE	DATA WAREHOUSE
Estructura de los datos	*Brutos (estructurados, semiestructurados y no estructurados)*	*Estructurados, procesados*
Finalidad de los datos	*Por definir, definida* ***Nota: Es posible que haya datos cuyo propósito no se haya definido (para uso futuro)***	*Definida*
Esquema	*On Read*	*On Write*
Usuarios	*Data Scientists*	*Usuarios empresariales*
Accessibilidad	*Gran accesibilidad y fácil actualización*	*Acceso y actualizaciones más complicadas y costosas*
Almacenamiento	*Almacenamiento distribuido y costes limitados (potencialmente ampliable a la nube)*	*Costes y revisión de costosos procesos de ingesta*

Principales diferencias entre Data Lakes y Data Warehouses

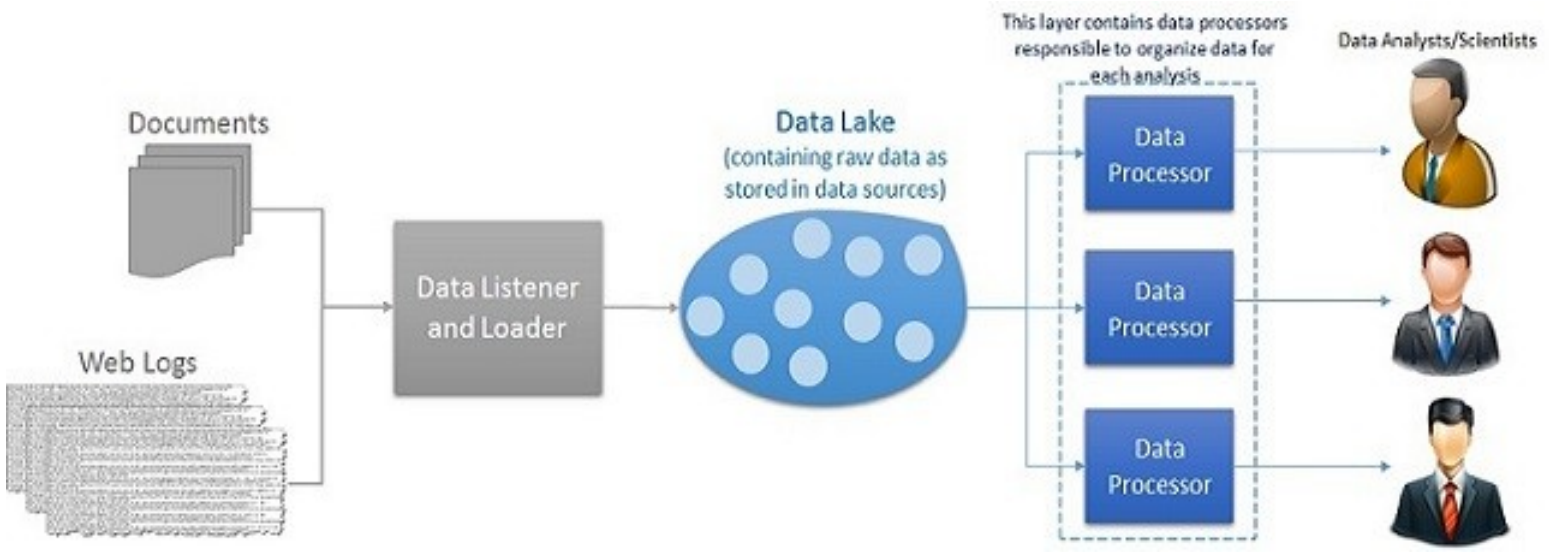
TEC | Tecnológico de Costa Rica

Características	Almacén de datos	Data Lake
Datos	Relacional de sistemas transaccionales, bases de datos operacionales y aplicaciones de línea de negocios	No relacional y relacional de dispositivos IoT, sitios web, aplicaciones móviles, redes sociales y aplicaciones corporativas
Esquema	Diseñado antes de la implementación de DW (esquema en escritura)	Escrito en el momento del análisis (schema-on-read)
Precio / rendimiento	Resultados de consulta más rápidos con un mayor costo de almacenamiento	Los resultados de las consultas se vuelven más rápidos usando almacenamiento de bajo costo
Calidad de datos	Datos altamente curados que sirven como la versión central de los reales	Cualquier información que pueda o no ser curada (es decir, datos sin procesar)
Usuarios	Analistas comerciales	Científicos de datos, desarrolladores de datos y analistas de negocios (usando datos curados)
Analítica	Informe por lotes, BI y visualizaciones	Aprendizaje automático, análisis predictivo, descubrimiento de datos y creación de perfiles

Data Lakes y Data Warehouses



Data Lakes y Data Warehouses



Ejercicio1

Ventajas y beneficios de un Data Lake

- Capacidad de obtener valor a partir de tipos ilimitados de datos
- Posibilidad de almacenar todo tipo de datos estructurados y no estructurados en un data lake, desde datos de CRM hasta publicaciones en redes sociales
- Mayor flexibilidad: no tiene que tener todas las respuestas por adelantado
- Posibilidad de almacenar datos en bruto: puede refinarlo a medida que su comprensión mejore
- Formas ilimitadas de consultar los datos
- Aplicación de una variedad de herramientas para obtener una idea de lo que significan los datos
- Eliminación de silos de datos
- Acceso democratizado a los datos a través de una única vista unificada de datos en toda la organización cuando se utiliza una plataforma de gestión de datos efectiva

Data Lake

TEC | Tecnológico de Costa Rica

- HDFS como servicio.
- Almacenamiento redundante.
- Escenarios:
  - Alta capacidad
  - Alta frecuencia
  - Alto rendimiento

- Almacenamiento de datos en su formato nativo
  - Estructurado, semi-estructurado y no estructurado.
- Almacenamiento ilimitado

Data Lake

TEC | Tecnológico de Costa Rica

- Confiable
  - Datos replicado 3 veces en una misma región.
  - Alta disponibilidad
- Optimizado para analítica
  - Creado para ejecutar grandes sistemas de análisis que requieren un rendimiento masivo.
  - Optimizado para el procesamiento en paralelo.

Ejercicio2

Características de un data lake

- Un único repositorio compartido de datos, normalmente almacenado en el Sistema de archivos distribuido (DFS). Los lagos de datos de Hadoop conservan los datos en su forma original y capturan los cambios a los datos y la semántica contextual a lo largo del ciclo de vida de los datos.Este enfoque es especialmente útil para las actividades de cumplimiento y auditoría interna.

Características de un data lake

Tecnológico de Costa Rica

- Incluye funcionalidades de orquestación y programación de trabajos (por ejemplo, a través de YARN). La ejecución de la carga de trabajo es un requisito previo para Hadoop empresarial.

Características de un data lake

- Contiene un conjunto de aplicaciones o flujos de trabajo para consumir, procesar o actuar sobre los datos.
- El fácil acceso de los usuarios es una de las características de un data lake, debido a que las organizaciones conservan los datos en su forma original. Ya sea estructurado, no estructurado o semiestructurado, los datos se cargan y almacenan tal cual. Los propietarios de datos pueden entonces consolidar datos de clientes, proveedores y operaciones, eliminando barreras técnicas e incluso políticas para compartir datos.

La solución

TEC | Tecnológico de Costa Rica

- Los data lakes por sí solos son sólo medios para un fin. Para lograr el objetivo final de proporcionar conocimientos empresariales, se necesita inteligencia de máquina impulsada por servicios de metadatos universales. Los servicios de metadatos universales catalogan los metadatos adjuntos a los datos, tanto dentro como fuera de Hadoop, y también capturan los tags proporcionados por el usuario sobre el contexto empresarial de los datos.

Almacenar todas las cosas

TEC | Tecnológico de Costa Rica

- **Información de la planta de fabricación** sobre velocidades de producción, errores o estadísticas de seguridad.
- **Entrada de RFID** y código de barras de los almacenes, incluidos temas de almacenamiento, envío y logística.
- **Estadísticas de participación** del usuario del sitio web de la compañía.
- **Interacciones en las redes sociales** con los clientes.
- **Registros de correo electrónico**, chat y teléfono desde soporte.
- **Datos de las campañas**de marketing.
- Aportes de **ventas B2B y B2C del CRM.**
- Datos que se obtengan a través de los **aparatos que estén conectados a la red.**
- etc.

Ejercicio3

EMC²

UNDERSTANDING DATA LAKES

WHAT IS A DATA LAKE?

Data lake is one place to put all the data enterprises may want to use, including structured and unstructured data

HOW DO DATA LAKES WORK?

The concept can be compared to a water body, a lake, where water flows in, filling up a reservoir and flows out.

DID YOU KNOW?

The information in the Digital Universe will grow 10 times by 2020.

In the last 10 years, companies have started using data lakes to deal with the enormous amounts of data.

Data lakes help reveal complex business issues and build predictive models to address these. Companies ranging from restaurants to mining corporations use data lake solutions in their everyday analytics.

WHO IS USING DATA LAKES?

BUSINESS & DATA ANALYSTS

Analyze reports on specific data in the organization to provide business insight

DATA ARCHITECTS

Responsible for designing, creating, deploying and managing an organization's data architecture

DATA SCIENTISTS & APP DEVELOPERS

Perform statistical analysis on big data to identify trends, solve business problems and optimize performance

WHY ARE DATA LAKES IMPORTANT?

BUILD APPLICATIONS

Platform for businesses to get at the data and quickly build the views, and data-driven applications they really need

FLEXIBILITY & ACCESSIBILITY

Provide flexibility and accessibility in moving large amounts of data from data warehouse to perform analytics

RETAIN DATA AUTHENTICITY

Data Lakes allow you to store and analyze the information in different formats, retaining data authenticity

SPEED

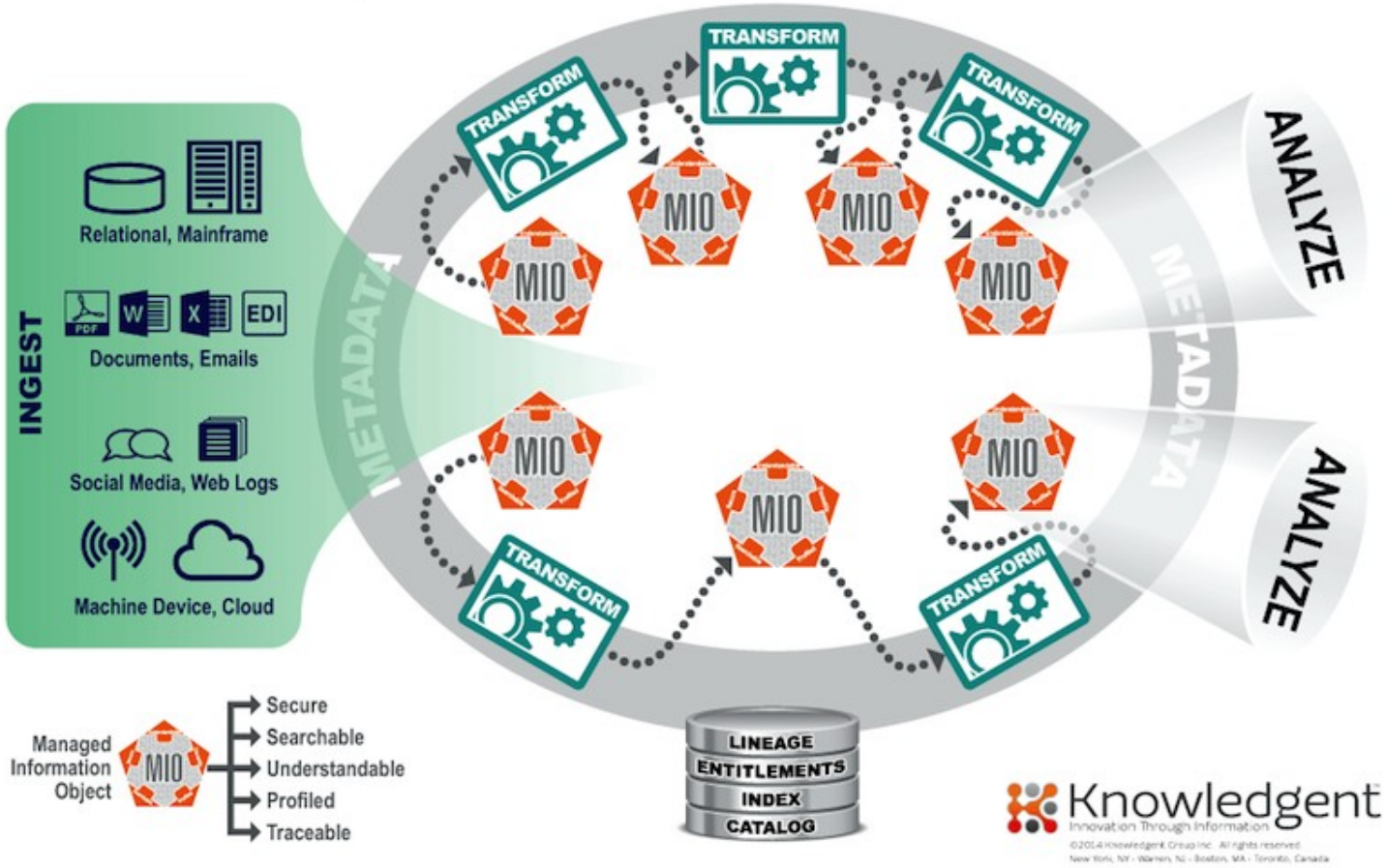
Ability to sift through immense quantities of data quickly

EXPLORE & ANALYZE

Ability to explore and analyze data to derive business value and benefit

EMC²

The Supervised Data Lake



Problemas con Data Lake

TEC | Tecnológico de Costa Rica

- **Los datos no estructurados requieren una programación especializada**. A pesar de que se produce después de que se almacenan los datos, necesitarás crear programas para acceder, ordenar, desinfectar y manipular los datos de una forma utilizable.
- **Tienes que planificar posibles casos de uso**. La planificación de posibles usos futuros te ayuda a aclarar los tipos de datos que tienes actualmente y si sus procesos que utilizas actuamente funcionarán en un futuro.
- **Mantenimiento**: Solo porque importes tus datos en formatos sin formato no significa que debas evitar limpiarlos. Asegúrate de que tu información se mantiene limpia para que tu lago no se convierta en un pantano.
- **El acceso al lago no es democrático**. En este punto, los analistas de datos deberían ser los únicos que tienen acceso al Dala Lake. Solo ellos entenderán cómo manipular los datos. Llegará el momento en el que los otros usuarios también puedan buscar para acceder a los datos necesarios, pero esa es todavía una mejora prevista para el futuro.
- **Acumulación de datos**: en algún momento, debes preguntarte qué harás con todos estos datos y por qué los estás manteniendo. Muchas compañías hablan de la importancia de almacenar todas las cosas para responder a preguntas futuras. Pero a veces esto se les va de las manos. A veces se aferran demasiado a esos datos incluso cuando no los necesitan todos.

Conclusión

TEC | Tecnológico de Costa Rica

- **En conclusión, el concepto de lago de datos / data lake, es recomendado para grandes volúmenes de datos de los que no se conocen a priori las estructuras analíticas. Por lo tanto, es un complemento del «data warehouse» que se mantiene como la estructura mejor adaptada al análisis repetitivo y comparativo de los datos estructurados de la empresa.**