Análisis, Diseño y Procesamiento de Datos Aplicados a las Ciencias y a las Tecnologías (ADP)

Master Degree on Computational Intelligence and Internet of Things

University of Córdoba - Spain

16/10/2024

Análisis, Diseño y Procesamiento de Datos Aplicados a las Ciencias y a las Tecnologías (ADP)

Data-Driven Organizations

Juan Antonio Romero del Castillo – aromero@uco.es

Programa de contenidos – Data Driven Organizations

1) Data-Driven Organizations
   - Data science to make informed decisions.
   - Roles: data analyst, data scientist and data engineer in organizations.
   - Data pipeline and IT infrastructure for data science (use case: the public cloud).
   - Introducction to Amazon AWS public cloud.
2) Preparing the Data. The Elements of Data
   - The five Vs of data: volume, velocity, variety, veracity, and value
   - Data sources
   - Planning Your Pipeline
3) Design Principles and Patterns for Data Pipelines
   - Evolution of data architectures
   - Modern data architectures
   - Pipeline: Ingestion and storage
4) Labs and activities

Sessions planning (4 sesions of 2.5h)

Session 1 (16/10/2024)

- Presentación de la asignatura. Programa teórico y práctico. Evaluación. (5m).
- Enrole students to AWS ACAv3 y AWS ADE (15m).
- Explain AWS Certifications, Canvas y AWS ‘Claim your Badget’ (10m).
- Teory presentation: Data-driven org. - ADE M02_DataDrivenOrganizations_InstructorNotes.pptx (60m).
- Concepts needed to do the AWS Lab (30m)
- **Lab 1:** AWS ACAv3 Module 3 Guided Lab: Exploring AWS Identity and Access Management (IAM) (40m)

Session 2 (17/10/2024)

- Terminar presentation: data-driven organizations (15m)
- Presentation: Elements of Data. ADE M03_Elements_of_Data_InstructorNotes.pptx (30m)
- Concepts needed to do the AWS Lab (30m)
- **Lab 2:** ADE Module 6 Guided lab: Creating an Amazon RDS Database (45m)
- Ejercicio preguntas y respuestas sobre RDS (30m).

Sessions planning (4 sesiones de 2.5h)

Session 3 (23/10/2024)

- Presentation: Design Principles and Patterns - ADE M04_DesignPrinciplesAndPattern.pptx (45m)
- Concepts needed to do the AWS Lab (10m)
- **Lab 3:** ADE Module 4 Lab: Querying Data by Using Athena (90m)

Session 4 (11/11/2024)

- Resumen repaso (10m)
- Concepts needed to do the AWS Lab (20m)
- **Lab 4:** AWS ADE Module 7: Performing ETL on a Dataset by Using AWS Glue (90)

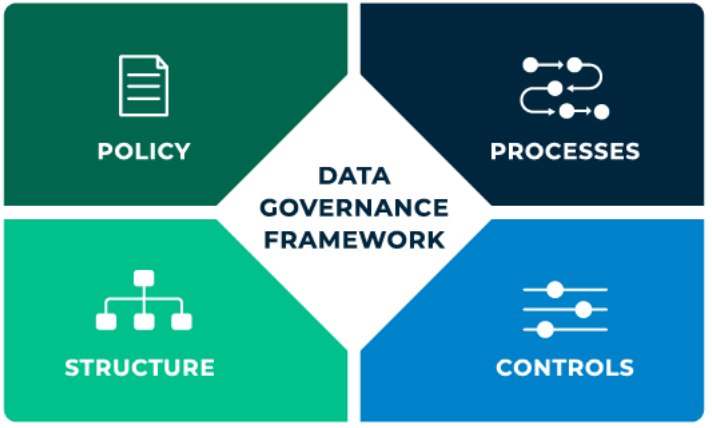
Conceptos Previos

POC. Data Democracy

- **POC**: A Proof of Concept is a prototype or a small-scale project designed to demonstrate the feasibility or practicality of a concept, idea, or technology. It is often used in business and technology to validate whether a particular approach or solution is viable before committing to a full-scale implementation
- **Democratizing access** to data refers to the idea of making information and datasets available and accessible in an equitable and open manner to a wide range of people and organizations, rather than restricting access to a select or limited group.

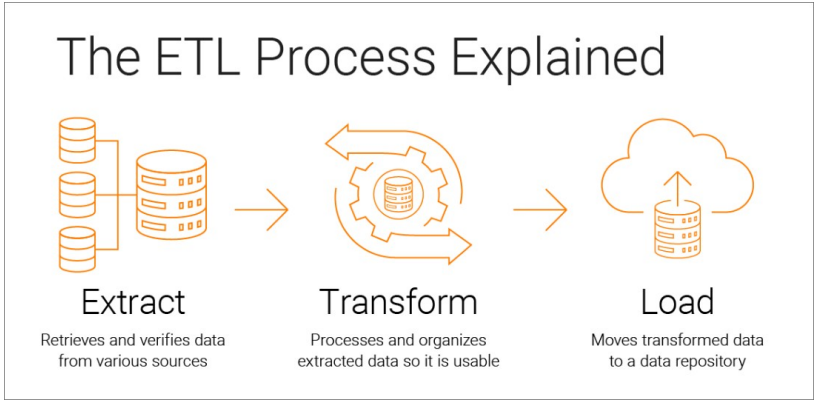
Data governance

- Refers to a set of practices and **policies** designed to effectively manage an organization's data. It involves defining responsibilities, processes and standards to ensure the quality, security, privacy and appropriate use of data. Data governance seeks to optimize data-driven decision making, reduce risks and promote transparency in information management.



Data Science Roles

- **Data scientist:** identify whether the data will be valuable (what to do?)
- **Data engineer**: mechanics and technical requirements (how to do?)
- **Data Science**: Data Science focuses on exploring data, discovering patterns, creating predictive models and deriving new information from raw data..
- **Raw data** refers to data that has not been processed, cleaned, or transformed in any way.
- **Business Intelligent** (BI): gathering, organizing and reporting to analyze an organization's past performance. There are many BI tools. Desde los 50-60 hasta la integración con la IA.
- **ETL process** (Extract, Transform, Load) it refers to a data integration process used to move and transform data from source systems to a destination system. Used in BI, data science, etc.



Business Intelligence vs Data Science

- **BI:**
  - Enfoque: BI se centra en la recopilación, análisis y presentación de datos históricos para proporcionar información que facilite la toma de decisiones empresariales.
  - Objetivo: Ayuda a las organizaciones a comprender su desempeño pasado y actual a través de informes y paneles interactivos.
  - Herramientas: Utiliza principalmente herramientas de generación de informes y visualización de datos para ofrecer información en tiempo real.
- **DS:**
  - Enfoque: Data Science se centra en la extracción de conocimientos y patrones a partir de datos, tanto históricos como en tiempo real, utilizando una variedad de técnicas y algoritmos.
  - Objetivo: Busca predecir tendencias futuras, descubrir información oculta y optimizar procesos mediante el análisis de datos.
  - Herramientas: Involucra programación, estadísticas avanzadas y técnicas de Inteligencia Artificial, aprendizaje automático y técnicas de minería de datos.

ML vs Non-ML

- **Non-ML** (Non-Machine Learning) applications rely on predefined algorithms and rules to perform tasks, are less data-dependent, and often lack adaptability. They require explicit programming, are typically used for tasks with well-defined rules, and are more interpretable.
- In contrast, **ML** (Machine Learning) applications **learn** from data, require substantial data for training, can **adapt** to changing conditions, automatically extract features from data, excel in **complex tasks** with complex or evolving patterns, may require more computational resources, and can be challenging to interpret. The choice between non-ML and ML depends on the specific problem and application requirements.

Datawarehouse, data lake, data silo, purpose-built storage

- **Datawarehouse:** a schema consisting of different databases made for one type of business reporting/analysis. It uses data from the data leak and can pour new data into the data lake.
- **Data silo:** is a large, sophisticated data store, but isolated from the rest and difficult to access and integrate with the rest of the warehouses. These difficulties are overcome with the data lake concept.
- **Data lake:** moder concept all databases from different sources, structured, semi-structured or unstructured, in a central point of truth, easily accessible, secure, etc.
- **Purpose-built storage:** a specific specialized storage for every app and goal: S3, DynamoDB, Aurora, SageMaker, RedShift, etc.

Big Data Framework

- Big data: Manejo de grandes y complejos volúmenes de datos (data science es un campo interdisciplinario de extracción de conocimiento y conclusiones que no se limita al big data).
- The framework is a set of tools and tech. Designed to efficiently process, store, manage and analyze big data.
- Frameworks:
  - Hadoop
  - Apache Spark
  - Apache Kafka: persistent record of events (kafka topics) and small services for processing topic in real time as soon as they happen and connections to internal and external systems.
  - Amazon MKS to run Apache Kafka.
  - Amazon Kinesis
  - Etc.

Generative IA

- Generative AI refers to deep-learning models that can generate high-quality text, images, and other content based on the data they were trained on.
- Artificial intelligence has gone through many cycles of hype, but even to skeptics, the release of ChatGPT seems to mark a turning point.
- Infografía AWS: https://d1.awsstatic.com/psc-digital/2023/gc-400/top-5-gen-ai-info/top-5-generative-ai-questions-es.pdf

OLTP vs OLAP

- **OLTP:**
  - Eficiente para transacciones en tiempo real: inserta, actualiza, etc.
  - On Line **Transactions** Processing o Procesamiento de Transacciones On Line (as in RDBMS)
  - Transactional processing. Normally, each transaction, even if it is divided into several transactions, is guaranteed to finish well or not finish at all. Ex: a bank transfer requires subtracting from the sender and adding to the receiver, if the former is done but not the latter, there would be serious inconsistencies.
  - Short/small transactions (INSERT, DELETE, UPDATE, etc.)
  - Row storage is adecuate in this case.
  - Normalized tables.
  - Critical systems.
- **OLAP:**
  - On Line Analytical Processing para consulta y análisis más complejos de grandes cantidades de datos.
  - Complex transactions that assemble specific data structures and algorithms that the OLTP method would take a long time to perform.
  - Extracts data for analysis and decision making
  - Few but lengthy transactions
  - Non-normalized tables

Thanks!