**Tema 1: Integración Procesamiento Análisis**

**Visión general Big Data**

Las 7-8 V’s de Big Data:

* **Volumen**:
  + Contexto: Escalabilidad, Comunicaciones, Computación
  + Reto: Almacenamiento paralelo, elevado coste infraestructura
* **Velocidad:**
  + Contexto: Latencia, Agilidad uso, Tiempo respuesta, Seguridad
  + Reto: Elevado ando de banda, Necesidad escalabilidad, Mayor acceso a datos, Elevada concurrencia, Impacto mecanismos de seguridad
* **Variedad:**
  + Contexto: Tipografía del dato, Integración
  + Reto: Introducción nuevas plataformas y tecnologías, Acceso a datos, Consistencia del dato
* **Veracidad:**
  + Contexto: Calidad, Conformidad
  + Reto: Complejidad estrategias control calidad del dato, Mayor esfuerzo políticas gobierno del dato
* **Variabilidad:**
  + Contexto: Escalabilidad, Comunicaciones, Computación
  + Reto: Almacenamiento paralelo, elevado coste infraestructura
* **Valor:**
  + Contexto: Escalabilidad, Comunicaciones, Computación
  + Reto: Almacenamiento paralelo, elevado coste infraestructura
* **Visualización:**
  + Contexto: Escalabilidad, Comunicaciones, Computación
  + Reto: Almacenamiento paralelo, elevado coste infraestructura

**Importancia Big Data**

* Toma de decisiones informada
* Identificación de patrones y tendencias
* Mejora de la eficiencia operativa
* Personalización y marketing dirigido
* Avances en la investigación científica
* Transformación digital
* Predicción y prevención
* Seguridad cibernética

**Papel de la integración de la información en Big Data**

La **integración** se refiere al proceso de **reunir datos** de múltiples fuentes/formatos/fuentes en un entorno

* Accesibilidad de los datos
* Consolidación de datos
* Mejora de la calidad de los datos
* Análisis más completo
* Enriquecimiento de datos
* Ahorro de tiempo y recursos
* Soporte para análisis avanzados
* Cumplimiento normativo

**Retos y oportunidades en los proyectos de Big Data**

* Volumen de datos
* Variedad de datos
* Velocidad de datos
* Veracidad de datos
* Privacidad y seguridad
* Talento especializado
* Costos
* Adopción cultural

**Oportunidades en los proyectos de Big Data**

* Toma de decisiones basada en datos
* Innovación
* Personalización
* Eficiencia operativa
* Competitividad
* Descubrimiento de tendencias
* Mejora en la atención al cliente
* Generación de ingresos

**Casos de éxito como**: Google, Netflix, Walmart

**Técnicas y procesos para extraer información de los datos**

* **Análisis estadístico**
  + Resumen estadístico
  + Pruebas de hipótesis
* **Aprendizaje automático**
  + Clasificación: Clasificación en clases
  + Regresión: Predecir valores numéricos basado en datos previos
  + Agrupación: Identificar patrones y agruparlos
  + Procesamiento de lenguaje natural (NLP): Extraer información
  + Redes neuronales: Usa redes profundas para tareas complejas
* **Minería de datos**
  + Reglas de asociación: Identificar patrones entre conjunto de datos
  + Segmentación: Dividir en grupos homogéneos
  + Detección de anomalías: Identificar valores atípicos
* **Minería de texto**
  + Clasificación de documentos: Categorizar doc. textos
  + Análisis de temas: Identifica los principales temas
  + Búsqueda de información: Encontrar información relevante en colecciones de datos
* **Visualización de datos**
  + Gráficos: Representar valores para identificar patrones
  + Mapas de calor: Mostrar densidad de datos
* **Técnicas de extracción de características**
  + PCA (Análisis de Componentes Principales): Reducir dimensionalidad de los datos manteniendo lo importante
  + Selección de características: Identificar y seleccionar características mas relevantes para un problema
* **Procesamiento de Imágenes**
  + Detección de objetos
  + Reconocimiento de patrones
  + Seguimiento de objetos
* **Webscrapping**
  + Herramientas como **BeautifulSoap** o **Scrapy**

**Modelización de datos, razonamiento y resolución de problemas**

Enfoques de modelado de datos

* Bases de datos de nodos
* Bases de datos clave-valor
* Bases de datos de documentos

**Bases de datos de nodos**

Se utiliza para modelar y almacenar datos en forma de nodos y relaciones.

* **Casos de uso**: Para redes sociales, rutas de transporte.. etc.

**Grafo**: Representa unos datos usando un conjunto de nodos y aristas.

Componentes del Grafo:

* **Nodos**: Representa una entidad u objeto
* **Aristas**: Conexiones entre nodos
* **Adyacencia**: Colección de nodos adyacentes

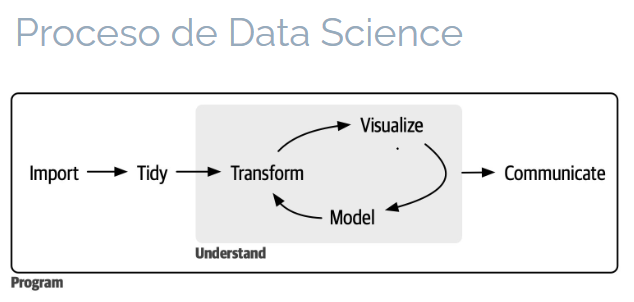
Tipos de Grafos:

* **Dirigido**: Cada arista tiene 1 dirección. Ej: hipervínculos
* **No dirigido**: Las aristas no tienen direcciones. Ej: grafo amistad
* **Ponderado**: Cada arista tiene un peso representado en valor numérico
* **Acíclico**: Grafo sin ciclos. Ej: Los arboles

Los **datos** y relaciones se **almacenarán** en un **SGBD** como **Neo4J**

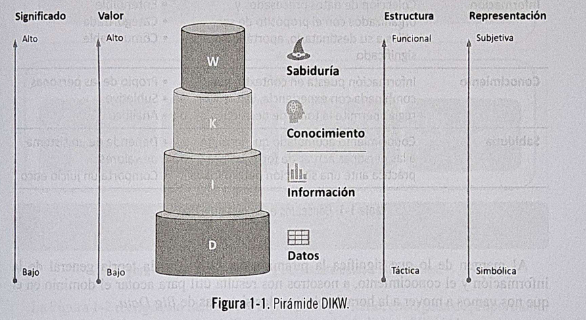
**Tema 2: R y Big Data**

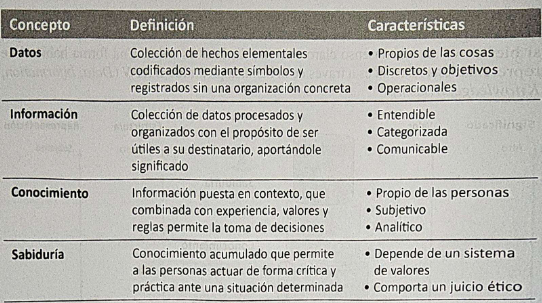
**R**: Es un lenguaje de programación utilizado para análisis de datos y visualización



**Tema 3: del Dato a la Información**

**Pirámide DIKW**: Data, Información, Conocimiento, Sabiduría





Caracterización del dato

* Datos al tipo
  + Simple
  + Compuesto
* Datos al formato
  + Estructurado: SQL
  + No estructurado: NoSQL
    - Datos textuales: documentos, mensajes, logs
    - Datos no textuales: audio, video, imágenes
  + Semiestructurado: XML,JSON
* Datos al generador
  + Generados por personas
  + Generados por máquinas
* Datos al tamaño
  + Pequeños (GB)
  + Medianos (TB)
  + Grandes (PB)
  + Muy grandes (EB)

Datos calientes (hotstorage): Requiere acceso frecuente

Datos fríos (cold storage): Permanece inactivo la mayor parte del tiempo

* Datos a su rol
  + Datos maestros: Clientes, empleados, productos
  + Datos operacionales: Facturaciones/Ventas
  + Datos externos: Datos meteorológicos
  + Datos analíticos: Hechos, dimensiones, perspectiva temporal

**Gestión de datos corporativos**: Procesos involucrados en el manejo de estos 4 tipos de datos

* Datos a su latencia: Tiempo total transcurrido entre un dato generado y puesto a disposición de los usuarios para ser consumido
  + En tiempo real
  + En lotes
* Datos a su sensibilidad
  + Riesgo alto: Obligatorio informar de cualquier incidencia
  + Riesgo medio: Están regulados y sujeto a restricciones
  + Riesgo bajo: Datos disponibles al publico

**BI**: Business Intelligence

**Computación en la nube**

Proporciona nuevas funcionalidades y capacidades.

Modelos de consumo (Ordenados por niveles):

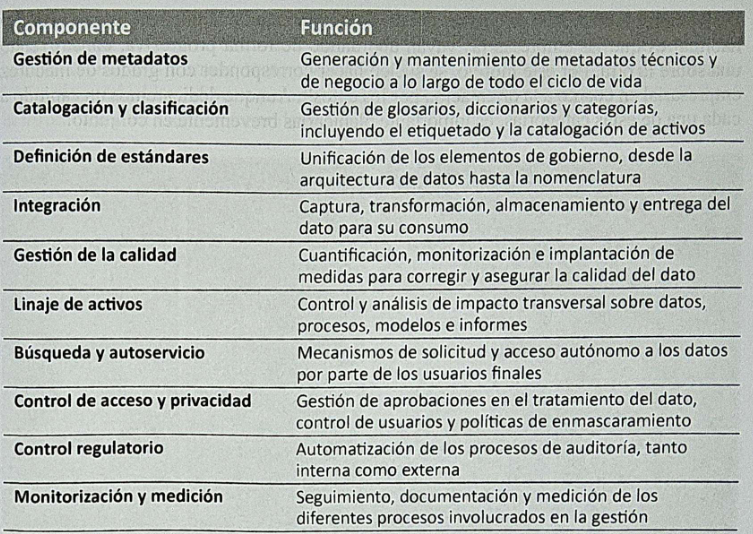
* **IaaS**: Infraestructura como Servicio. Permite desplegar rápidamente aplicaciones sin necesidad de modificarlas.
* **PaaS**: Servicio de Plataforma. Dispondremos de piezas que podremos ensamblar mediante código para conformar nuestra aplicación.
* **SaaS**: Software como Servicio. Donde consumimos una aplicación.

**Microservicios**: Separar y gestionar de manera individual distintos elementos que integran una solución.

**Nube híbrida**: Componentes se ejecutan en nubes de distintos proveedores (centro de datos local), comunicándose de manera segura

**API**: Interfaz de programación. Permite automatiza la mayoría de tareas de gestión.

**Gestión y gobierno del dato**



**Etapas de Análisis en la Explotación de la Información**

Sacado del libro:



Sacado de internet:

