Curso

1

MACHINE LEARNING & BIG DATA

Conceptos básicos, Estadística, Exploración y Data Quality

Fundamentos

JOSÉ NELSON ZEPEDA DOÑO

Cluster de Estudio: Advanced Analytics

Este material es el resumen de muchos autores que por medio de sus libros y documentos nos ofrecen fuentes riquísimas de conocimiento sobre los temas de Big Data y Machine Learning.
Algunas citas, figuras y tablas pueden ser encontradas de forma textual tal como lo indica el autor en su material original.

Nelson Zepeda
MIP • V 1.0
San Salvador El Salvador
Phone 503 79074137 • @nelsonzepeda733

Tabla de Contenido

Capítulo 1

Conceptos de Big Data

Los humanos han estado generando datos por miles de años.
Actualmente hay un repunte asombroso de la cantidad de datos que se producen y la variedad de los mismos.

Big Data es la combinación de data transaccional y data interactiva, la data transaccional es la que comúnmente se obtiene de los sistemas típicos de una empresa como un ERP, un CRM, un sistema hecho en casa, etc. mientras que la data interactiva proviene de fuentes tales como Internet, redes sociales, sensores, etc.

Pero, ¿Qué son los datos? Esta pregunta parece ser muy simple, sin embargo, su respuesta puede ir desde que "los datos son algo almacenado puntualmente en forma de letras y números para representar un hecho o suceso" hasta "todo lo que existe bajo el sol" siendo ambas respuestas validas por lo que se puede aseverar que todo es data y que el verdadero problema había sido capturar y preservar dicha data por temas de capacidad, costos y conectividad.

Cuando los datos son tratados adecuadamente, se puede obtener un valor inmenso de cara al negocio, la sociedad, la ciencia, etc. tal como se muestra en la siguiente figura.

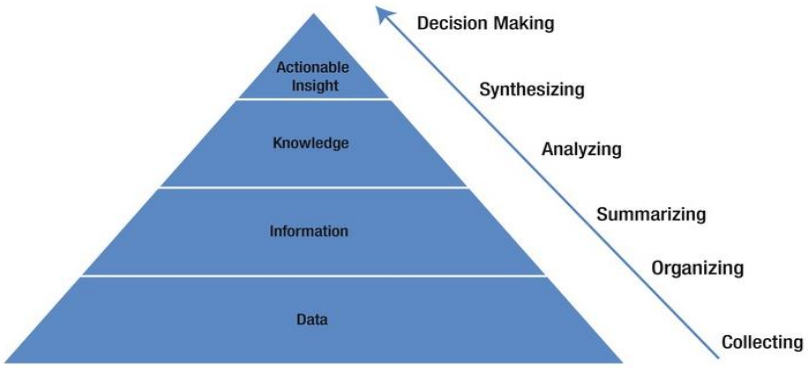


Figura 1-1 Transformando data en sabiduría.

Las 4Vs de Big Data

En la actualidad en donde ya existe más capacidad de procesamiento, almacenamiento y sobre todo capacidad de generar más y más data el concepto de big data cobra vida caracterizándose por dar una respuesta a aquellas necesidades de administrar un gran Volumen, alta Variedad y datos que se generan a gran Velocidad y sobre todo de forma Veraz.

Con las 4Vs a disposición de las industrias los datos han también evolucionado y una nueva característica los acompaña: la Complejidad.

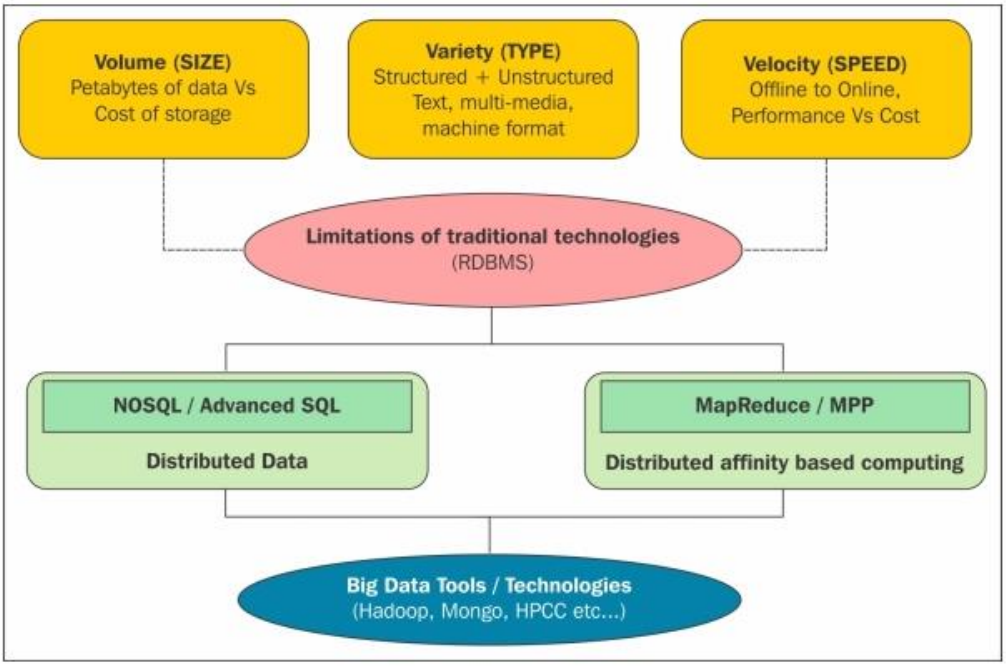


Figura 1-2 Big Data Vs

La velocidad, variedad, volumen, veracidad y complejidad hacen que big data tenga una naturaleza poli-estructurada, es decir tiene la naturaleza perfecta para hacer frente a datos estructurados, semi-estructurados y no-estructurados de manera rentable lo que nos permite concluir que el verdadero desafío es encontrar el valor de la data.

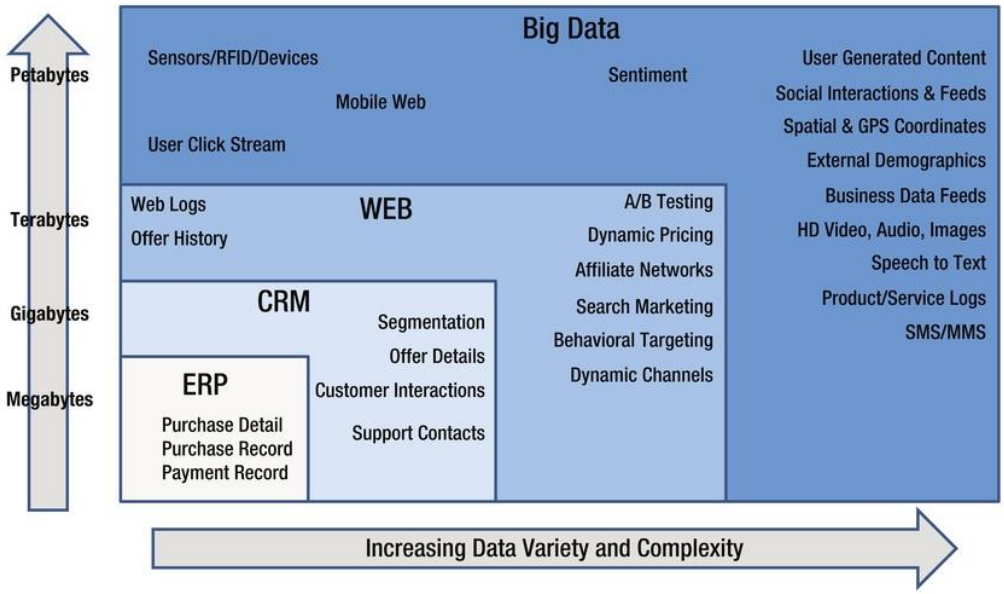


Figura 1-3 Generación de los datos

Para algunas industrias Big Data ha tomado una relevancia extremadamente importante, inicialmente se concebía como un concepto ligado totalmente a volúmenes de datos, luego fueron comprendiendo que no solo se trata de volumen sino también de complejidad en donde los tiempos de respuesta son importantes y el término "real time" puede ser toda una realidad.

1000 Gigabytes (GB) = 1 Terabyte (TB)
1000 Terabytes = 1 Petabyte (PB)
1000 Petabytes = 1 Exabyte (EB)
1000 Exabytes = 1 Zettabyte (ZB)
1000 Zettabytes = 1 Yottabyte (YB)

Tabla 1-1 Midiendo Big Data

En los últimos años el crecimiento ha sido exponencial ya que las diferentes industrias están generando y guardando cada vez más y más data y tratando incansablemente de generar valor.

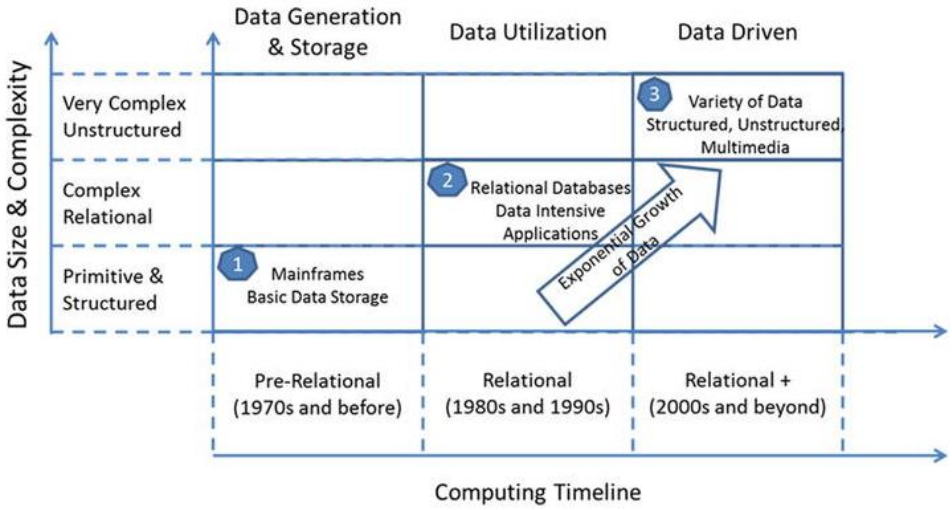


Figura 1-4

	Volume of Data	Velocity of Data	Variety of Data	Under -Utilized Data ('Dark Data')	Big Data Value Potential
Banking and Securities	High	High	Low	Medium	High
Communications & Media Services	High	High	High	Medium	High
Education	Very Low	Very Low	Very Low	High	Medium
Government	High	Medium	High	High	High
Healthcare Providers	Medium	High	Medium	Medium	High
Insurance	Medium	Medium	Medium	Medium	Medium
Manufacturing	High	High	High	High	High

Figura 1-5 Big data en las diferentes industrias

Big Data y el Negocio

Otro aspecto importante de Big Data es comprender en donde, cuando y como puede ser utilizada, la siguiente ilustración indica como Big Data puede estar totalmente alineada a los objetivos estratégicos del negocio.

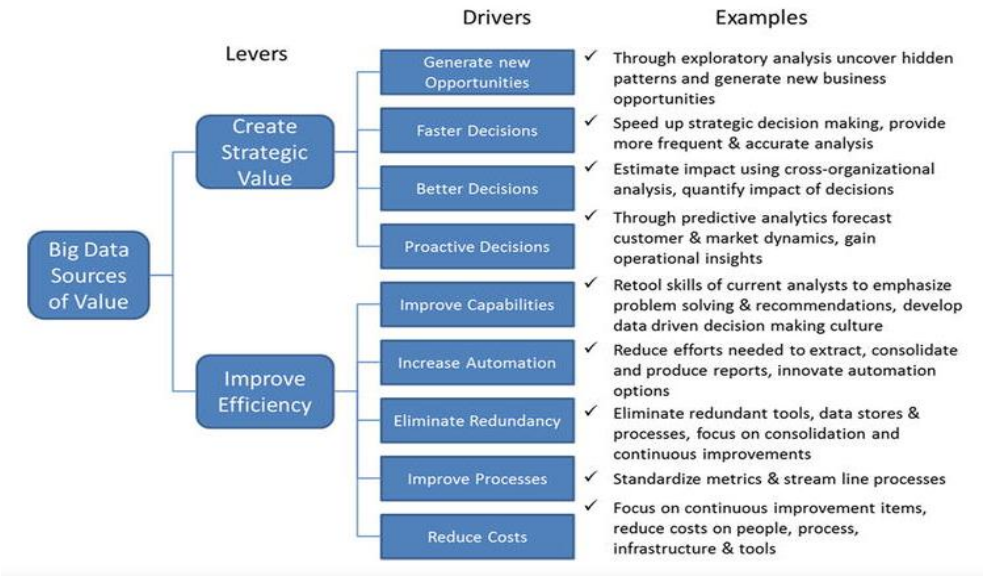


Figura 1-6 Drivers de Big Data

Las diferentes industrias están desarrollando diversos casos de uso para dar respuestas y soportar las estrategias de las áreas de negocio que van desde mercadeo hasta la cadena de abastecimiento.

Retail		Manufacturing	
✓ Customer Relationship Management ✓ Store Location & Layout	✓ Fraud Detection & Prevention ✓ Supply-Chain optimization ✓ Dynamic Pricing	✓ Product Research ✓ Engineering Analysis ✓ Predictive Maintenance	✓ Process & Quality Metrics ✓ Distribution Optimization
Financial Services		**Media & Telecommunications**	
✓ Algorithmic Trading ✓ Risk Analysis	✓ Fraud Detection ✓ Portfolio Analysis	✓ Network Optimization ✓ Customer Scoring	✓ Churn Prevention ✓ Fraud Prevention
Advertising & Public Relations		**Energy**	
✓ Demand Signaling ✓ Targeted Advertising	✓ Sentiment Analysis ✓ Customer Acquisition	✓ Smart Grid ✓ Exploration	✓ Operational Modeling ✓ Power-Line Sensors
Government		**Healthcare & Life Sciences**	
✓ Market Governance ✓ Weapon Systems & Counter Terrorism	✓ Econometrics ✓ Health Informatics	✓ Pharmacogenomics ✓ Bioinformatics	✓ Pharmaceutical Research ✓ Clinical Outcomes Research

Figura 1-7 Casos de uso de big data

Y tal como en muchos otros proyectos, identificar el caso de uso para soportar una necesidad de negocio es la primera fase para trabajar con big data de forma acertada.

La siguiente figura describe los macro pasos que deben seguirse para llevar a cabo un proyecto de esta naturaleza.

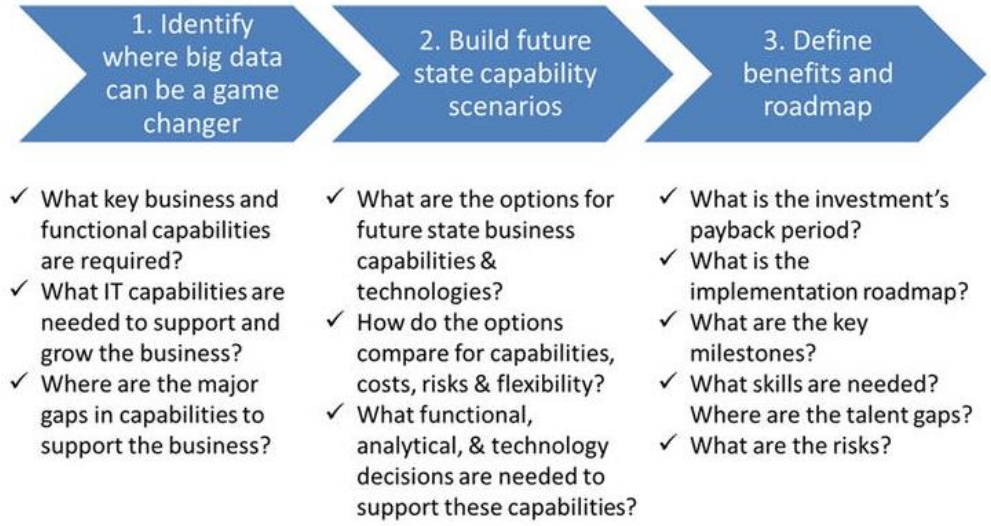


Figura 1-8 Big data roadmap

El científico de datos.

Adicional al recurso tecnológico, un aspecto fundamental es el recurso humano, desarrollar este tipo de conocimiento en las personas suele ser un trabajo muy duro y costoso, generalmente aquellas posiciones que suelen ser totalmente técnicas tienen mejores oportunidades de ser cubiertas pues en el mercado ya hay muchas personas especializándose en cada una de las herramientas del ecosistema big data, pero tal como se planteaba en el roadmap de big data, se requiere de muchos elementos no necesariamente técnicos para garantizar el éxito de un proyecto y es aquí donde nace la necesidad del Científico de Datos.

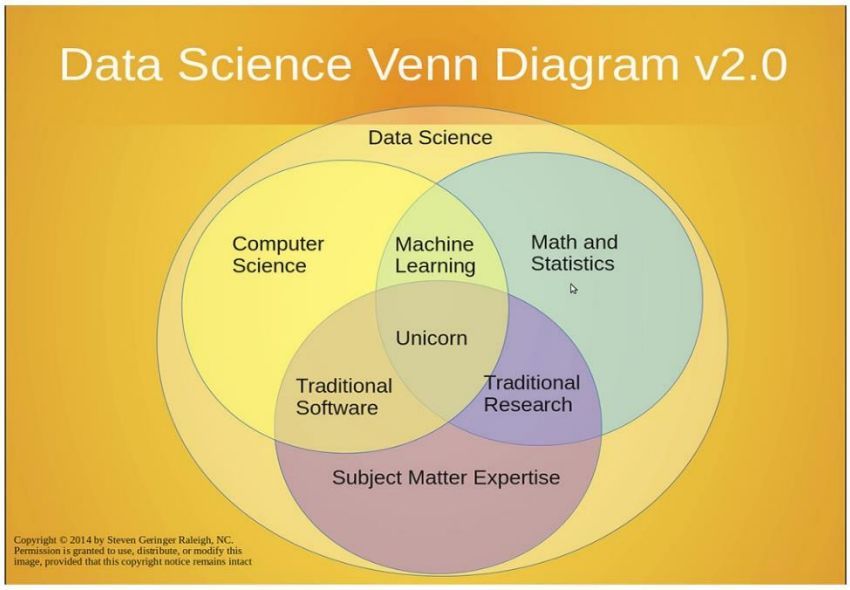


Figura 1-9 Habilidades del científico de datos

La figura anterior muestra con toda claridad todo lo que conlleva ser un científico de datos, y dada la complejidad de todas las cosas que debe conocer y aplicar, buscar este tipo de personas es equivalente a buscar "unicornios" para las organizaciones.

La estrategia actual seguida por los departamentos de RRHH es construir equipos multidisciplinarios en donde cada integrante contenga uno o varios de los skills requeridos.

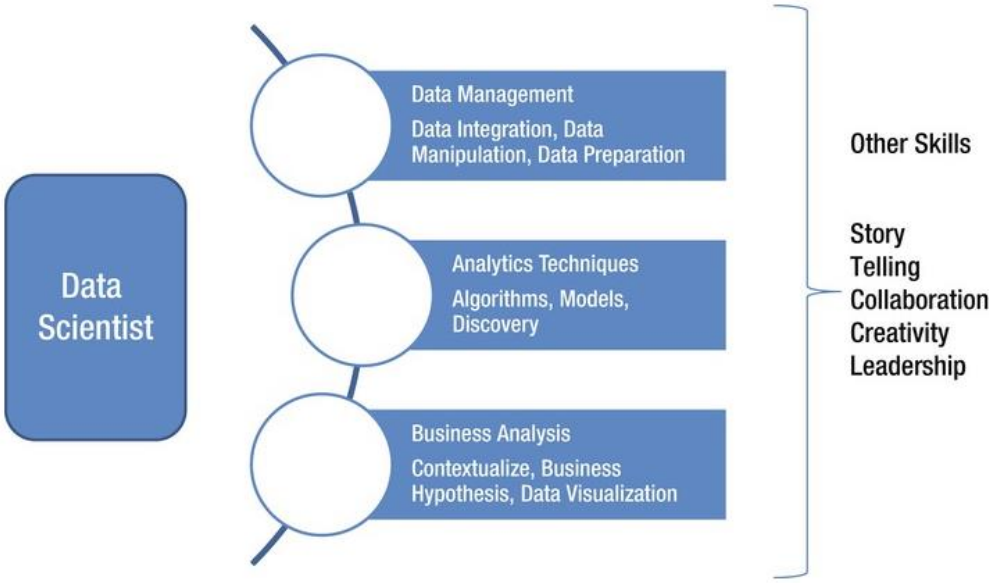


Figura 1-10 Skills de un científico de datos

El Ecosistema Big Data

El ecosistema de herramientas disponibles para trabajar big data está compuesto por muchos componentes, en este caso nos enfocaremos a los componentes de Hadoop como implementación de big data puesto que es el framework más popular entre las diferentes industrias.

Hadoop

La arquitectura de Hadoop está basada en 2 grandes componentes:

- HDFS: componente que maneja la computación distribuida y el almacenamiento.
- MapReduce: Componente que maneja el procesamiento en paralelo.

Hadoop es un proyecto que nació a finales del año 2002 y ha venido evolucionando hasta alcanzar la versión 2.X lanzada en el 2012 y mejorada en el 2014.

La demanda de Hadoop ha sido tal que actualmente en el mercado se encuentran diferentes distribuciones que ofrecen interfaces graficas / web, herramientas de monitoreo, IDEs para desarrollo, sub-paquetes y por supuesto no puede faltar el soporte.

Algunas de las distribuciones son:

1. Cloudera
2. HortonWorks
3. MapR
4. Amazon Elastic MapReduce

Independientemente de la distribución, el ecosistema de Hadoop tiene componentes claves que se describen en la siguiente figura.

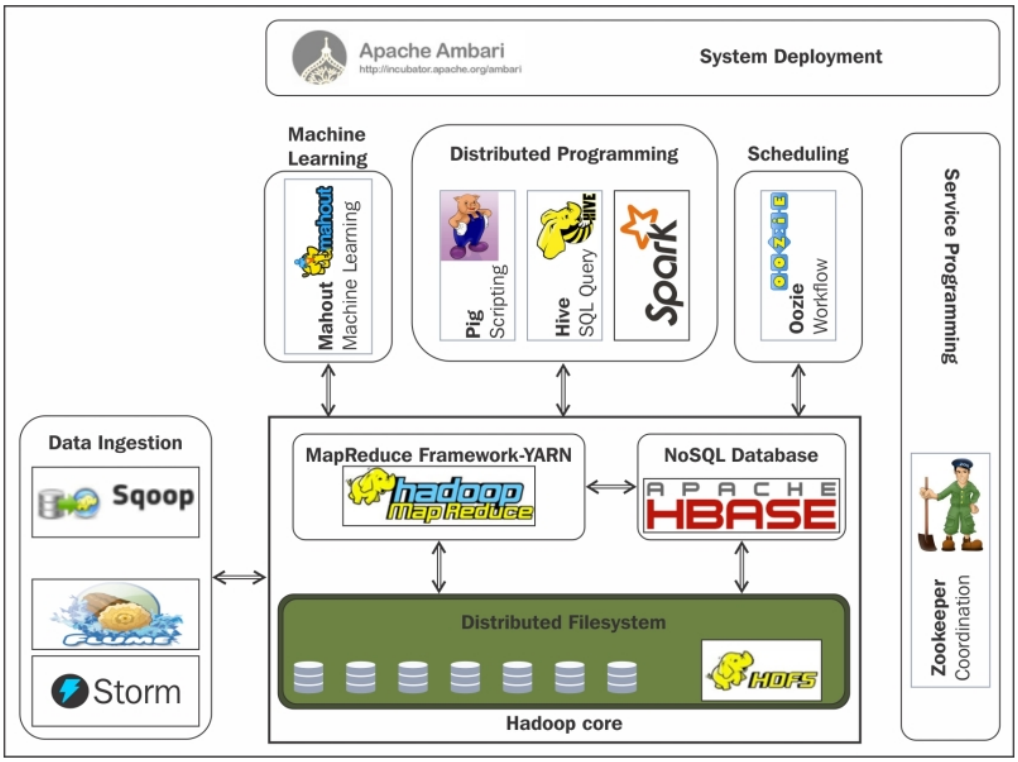


Figura 1-11 Ecosistema Hadoop

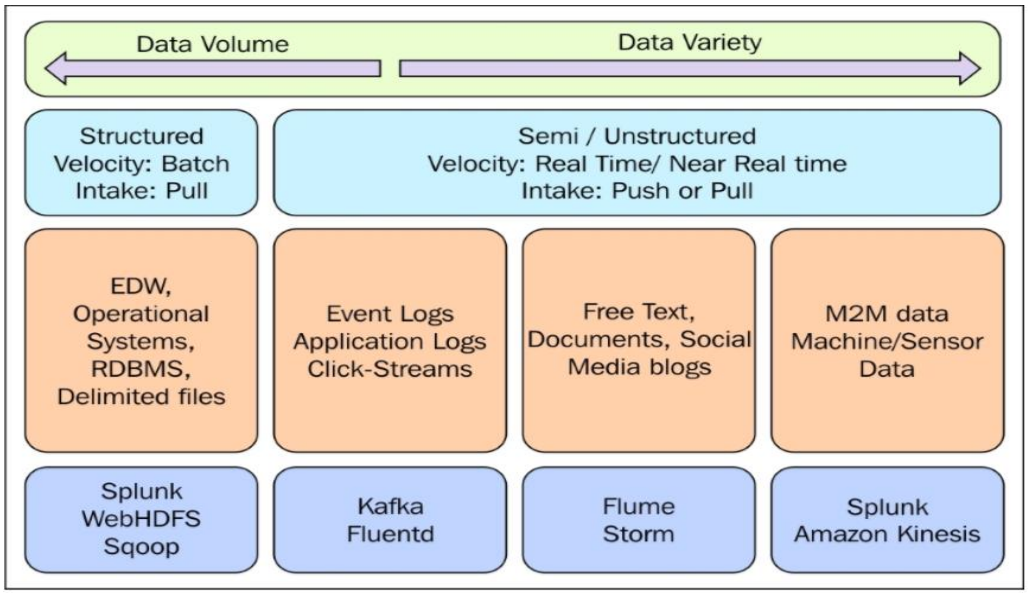


Figura 1-12 uso de las principales herramientas de Ingestión

Capítulo 2

Data Lake

Se destinan millones de dólares, pero se construye mal, gastamos años y sigue sin funcionar, ¿Alguien entiende que es un Data Lake?

Si se construye adecuadamente, existirá valor para el negocio en esta fascinante tecnología relacionada a los Data Lakes, y lo más importante es que si logramos generar ese valor, el Data Lake será un activo dentro del balance general de nuestra organización y no un simple gasto.

A medida que la explosión del Big Data va cubriendo más industrias y el dominio sobre este tipo de tecnologías se vuelve más natural, las organizaciones van almacenando una cantidad interminable de datos en estructuras llamadas Data Lake, lo cual nos deja entrever que el problema actual no es recolectar datos sino que el verdadero desafío sigue siendo obtener algo útil de ese mar de registros.

Antes de entrar a conceptos técnicos y debatir si es bueno o malo, veamos lo que tenemos actualmente:

Un Data Warehouse, las herramientas para Inteligencia de Negocios y hasta los reporteadores más simples del mercado fueron diseñados para contestar preguntas del pasado; cuestionamientos como *¿Cuáles fueron las ventas del mes pasado?*, eran contestadas por medio de alguna herramienta que hacia una lectura a nuestros datamarts o su equivalente, y éramos capaces de brindar respuestas que no solo proporcionaban un numero en frio, sino que podíamos acompañarlas con dimensiones de geografía, canales de venta, distinción de productos, etc.

Adicional al hecho de que todo estaba diseñado para contestar cosas del pasado, estaba pensado exclusivamente para los usuarios de negocio con el fin de ayudarles a tomar mejores decisiones.

Con la llegada del Big Data, Machine Learning, Data Lakes y todo el ecosistema actual de analytics, los sistemas ya no están diseñados solamente para contestar preguntas de negocio relacionados a eventos pasados a los gerentes de una empresa, hoy en día los sistemas están diseñados para que personas como tú y yo tomemos mejores decisiones y mejoremos nuestro diario vivir.

Para muchos de nosotros ahora es natural abrir una aplicación en nuestro dispositivo móvil y buscar cómo llegar a una dirección, ver el clima, hacer un plan de vacaciones, comprar un producto, revisar redes sociales, en fin una lista interminable.

Lo que algunas veces no dimensionamos es que todas estas aplicaciones utilizan tecnologías que son capaces de entender nuestro lenguaje natural, mostrarnos un mapa, capturar imágenes o entender si tenemos una sonrisa en nuestro rostro se basan en Big Data y Machine Learning, en donde cada registro es procesando a altas velocidades

con el objetivo de generar valor para el usuario y en forma general para la organización y algunos casos porque no decirlo generar valor para la sociedad.

Ahora bien, esta labor democratización cuyo objetivo es sacar a la luz el valor oculto en los datos y empoderar a las personas comunes para que tomen mejores decisiones en las actividades de su día a día, las organizaciones deben almacenar esas grandes cantidades de registros y luego procesarlas de manera correcta y eficiente, es decir no se podría hacer sin Big Data.

Con todo lo antes planteado, podemos remarcar las siguientes diferencias entre los sistemas tradicionales para análisis de datos vs la nueva generación de sistemas y arquitecturas:

- La escalabilidad del almacenamiento y el poder de procesamiento son diferentes.
- En el enfoque tradicional, la data proviene de sistemas relacionales y estructurados, en la nueva era del Big Data la data puede provenir de todo tipo de fuentes incluyendo las no estructuradas.
- La velocidad de procesamiento de los sistemas tradicionales es menor.
- La complejidad de los algoritmos que se pueden aplicar sobre la data.
- El enfoque tradicional ofrece reporteria y cubos con drill-downs, el nuevo enfoque es mucho más visual incluyendo mapas de calor, graficas de N dimensiones, etc. El Story teller es una realidad y una necesidad.

  La siguiente figura tomada del libro Data Lake Development with Big Data nos muestra de forma más grafica las diferencias entre ambos enfoques.

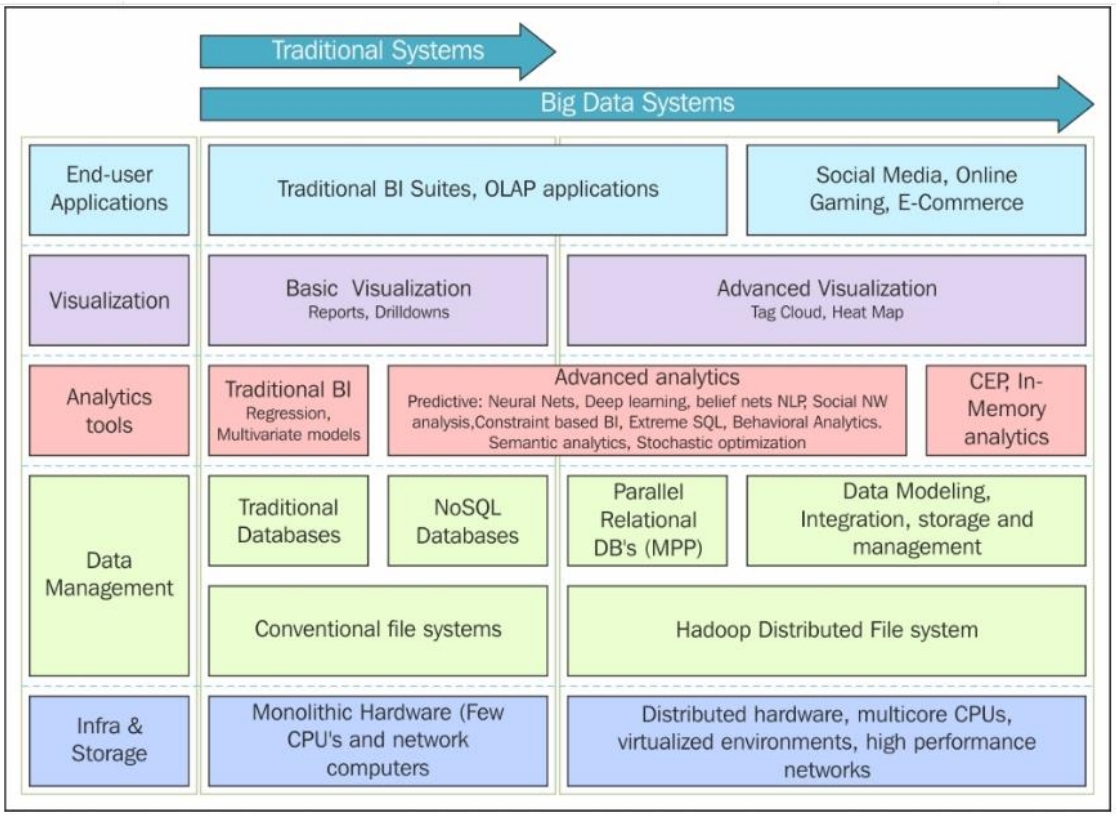


Figura 2-1 Diferencias entre el enfoque tradicional y el enfoque de Big Data

Pasemos ahora a dar la definición de Data Lake:

Un Data Lake, es un repositorio que almacena una gran cantidad de datos estructurados, semi-estructurados y no estructurados en su formato natural, es decir todo está almacenado de forma plana y los datos se van procesando/preparando según sea necesario. Debe ser reconocido como un punto de integración de la data para propósitos de análisis, no como un puente o colaboración entre los sistemas operacionales.[1]

Un Data Lake se espera que siempre esté disponible para proporcionar conclusiones, hallazgos y recomendaciones relevantes a partir de la data que almacena mediante el uso de diferentes análisis, herramientas y algoritmos de Machine Learning.

[1] Tomado de https://martinfowler.com/bliki/DataLake.html

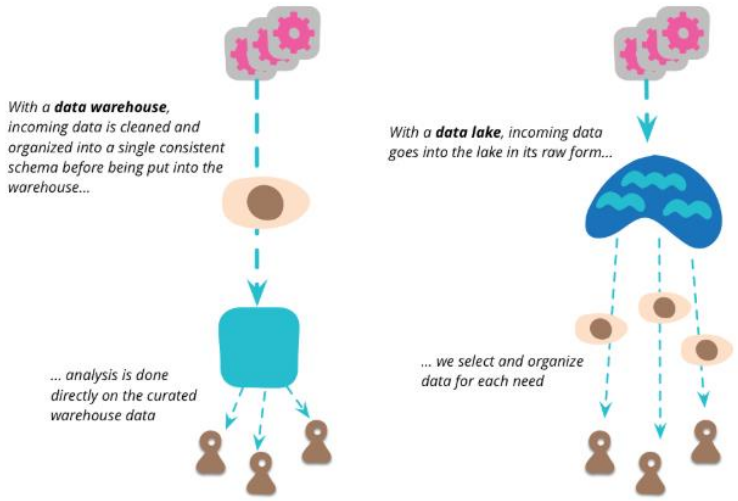


Figura 2-2 Los datos se van preparando según la necesidad

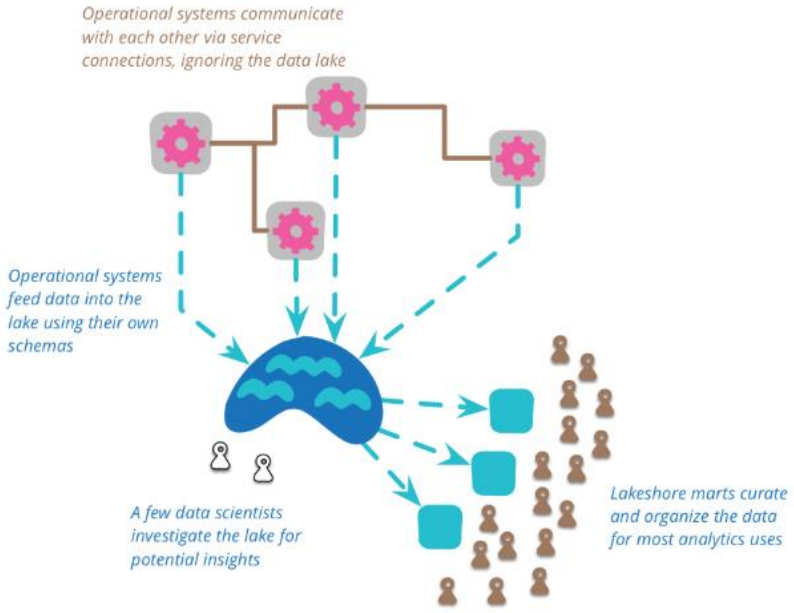


Figura 2-3 El Data Lake como un repositorio para todos los tipos de usuario

Como referencia la siguiente imagen nos da una idea más clara sobre los diferentes tipos de datos:

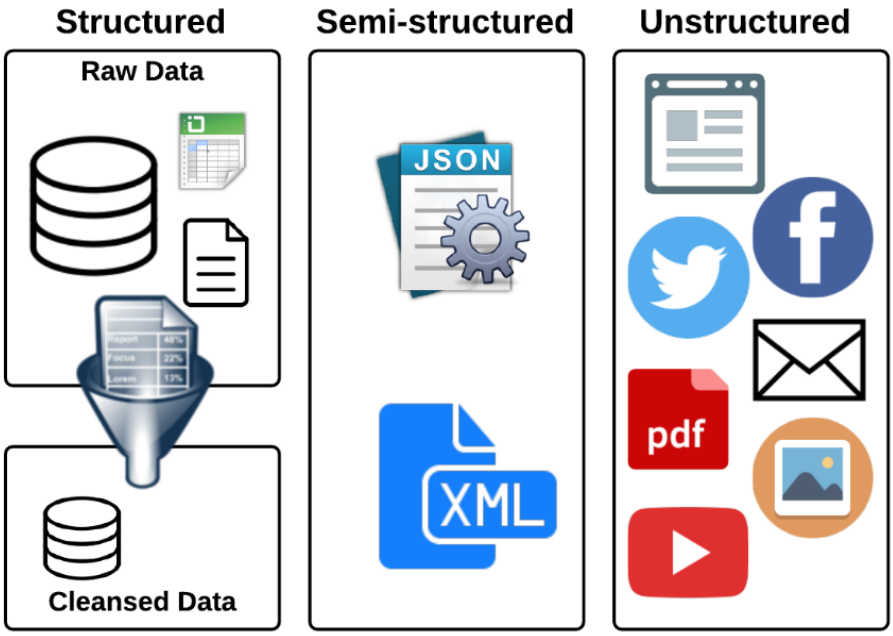


Figura 2-4: Tipos de datos

Beneficios de un Data Lake

Cuando existe un Data Lake que ha sido diseñado para almacenar toda la información producida por una organización y se tienen los medios adecuados para procesarla, presentarla y operativizarla nos encontramos ante un modelo centralizado cuyo conocimiento y administración consistente permitirá que se obtengan los siguientes beneficios:

- Data Governance
- Data Lineage
- Generación de Business Intelligence
- Trazabilidad de la información a través de los datos históricos
- Machine Learning para análisis predictivos y algoritmos no supervisados
- Capacidad para generar información en tiempo real

Ahora bien, para que estos beneficios sean una realidad, es necesario que 2 aspectos se cumplan:

- Se debe tener acceso a la tecnología necesaria para adquirir, almacenar y procesar grandes volúmenes de datos estructurados y no estructurados.

- Se debe tener las habilidades necesarias para el procesamiento de analíticos en tiempo real o casi tiempo real sobre estos grandes volúmenes de datos de manera iterativa.

Definición Técnica de un Data Lake

Funcionalmente hablando, un Data Lake es un repositorio de datos capaz de alojar los datos producidos o necesarios para una organización sin importar la naturaleza de los mismos o si provienen de fuentes no estructuradas.

Los siguientes conceptos definen un Data Lake:

- Data Lake es un gran repositorio de datos capaz de almacenar datos de cualquier naturaleza en su formato original (crudos) y también de ponerlos a disposición de la organización.

- Un Data Lake no es solamente "Hadoop". Un Data Lake utiliza diferentes herramientas. Hadoop provee un sub conjunto de todo el ecosistema actual de Big data.

- Un Data Lake no es la típica base de datos ya que se deben considerar los conceptos y herramientas para NoSQL y procesamiento en memoria.

- Un Data Lake no se puede implementar de forma aislada. Tiene que implementarse junto con un depósito de datos ya que complementa varias de sus funcionalidades.

- Un Data Lake, además de guardar datos estructurados, semi estructurados y no estructurados, debe ser capaz de alojar data proveniente de sensores y logs cuya característica es el movimiento rápido es decir stream.

- Un Data Lake esta optimizado para el procesamiento de datos, su enfoque no es el manejo de transacciones.

- Permite modelar los datos no solo bajo la perspectiva relacional tradicional:

    - Se puede modelar un grafo para encontrar las interacciones entre sus elementos: Neo4J
    - Como un almacen de documentos: MongoDB
    - Como un modelo columnar: Hbase
    - Como un modelo indexado optimizado para búsquedas: Riak

- Un Data Lake da la flexibilidad de guardar cualquier tipo de datos y que esta data cruda sea consumida dando paso a crear múltiples puntos de vista dentro de la organización por lo cual es mandatorio implementar controles para garantizar la consistencia de los datos:

    - Políticas y governance
    - Master Data Management

- Research Data Management
- Controles de seguridad y accesos.

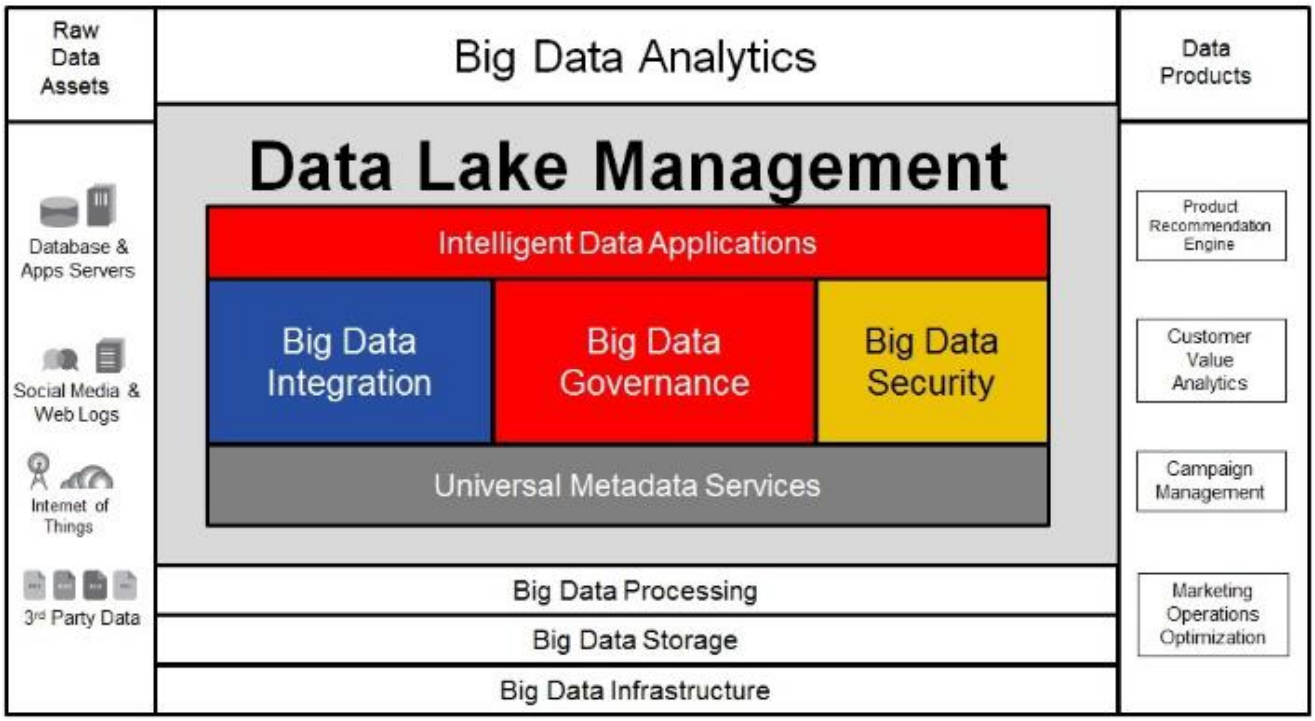


Figura 2-5: Datalake Management[2]

[2] https://blogs.informatica.com/2016/10/26/data-lake-management-poorly-architectured-data-lakes-make-fishing-insights-difficult/#fbid=QWDNUg1p3Ar

Capítulo 3

Machine Learning

En 1959 Arthur Samuel pionero de la Inteligencia Artificial, lanzo el primer concepto de Machine Learning.

Machine Learning son un conjunto de métodos/algoritmos diseñados para encontrar patrones y tendencias en los datos. Se encuentra en la intersección entre las matemáticas y estadística con la ingeniería de software y ciencias de la computación.

Los algoritmos se pueden clasificar en dos grandes categorías:

Familias de técnicas de ML

- Aprendizaje Supervisado: En este proceso de aprendizaje la variable de salida está bien definida (variable objetivo), es decir estas técnicas nos son útiles cuando nos interesa hacer predicciones sobre una variable objetivo.

- Aprendizaje No Supervisado: Este proceso de aprendizaje no implica tener una variable objetivo bien identificada, su objetivo no es hacer predicciones.

ML puede conceptualizarse como un ecosistema de 3 componentes principales:

- Modelo: El modelo es un objeto creado durante la fase de entrenamiento y es prácticamente el sistema que genera las predicciones lo más acertado posible mediante la identificación de las relaciones entre las diferentes variables y sus patrones.

- Parámetros: Estos son los atributos que utiliza el modelo para evaluar sus decisiones y establecer la conexión entre las diferentes variables.

- Learner: Este es el componente que hace los ajustes al modelo mediante la evaluación de las predicciones versus el mundo real.

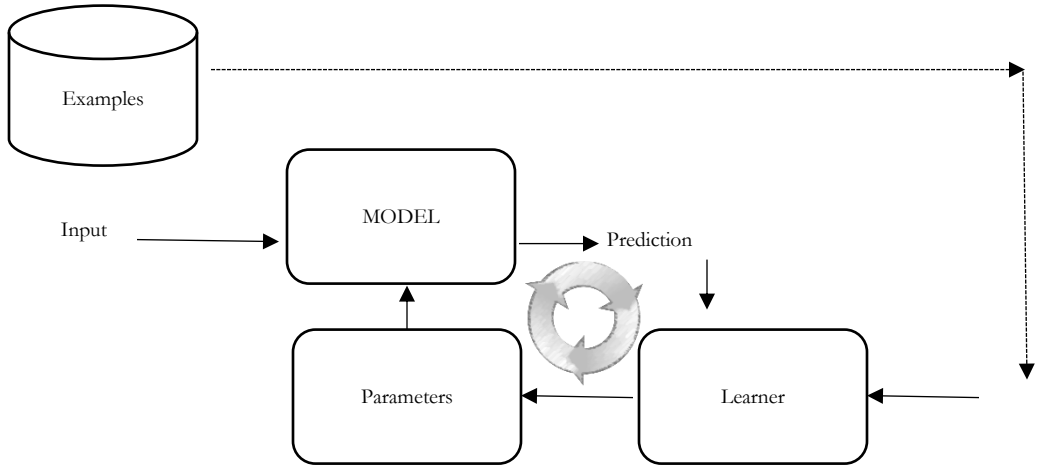


Figura 3-1 Vista de Alto Nivel de un sistema de ML

Mientras más datos tengamos, el modelo de ML que construyamos ofrece mejor calidad y su capacidad predictiva se ve mejorada sustancialmente.

Tal como se planteaba en los párrafos anteriores, existen dos grandes familias de algoritmos, iniciaremos esta sección hablando de los algoritmos supervisados, es decir nos enfocaremos en los algoritmos que son útiles en aquellos problemas en donde existe una variable objetivo a predecir/analizar y lo primero que debemos considerar es que este tipo de algoritmos se divide en 2 sub-categorías:

- Clasificación: Cuando la variable a predecir pertenece a un grupo bien definido de clases ejemplo, predecir si un cliente volverá a comprar o no.

- Regresión: Cuando la variable a predecir es de tipo numérica, específicamente un número Real, ejemplo, predecir el valor de una casa dada su ubicación, tamaño, etc.

Los diferentes algoritmos de tipo supervisado son:

- Métodos heurístico: Classification Trees y Nearest Neighbor

- Métodos de separación: Neural Networks, Support Vector Machines.

- Métodos de regresión: Logistic Regression

- Métodos Probabilísticos: Bayesian Methods

Por otro lado, abordando la otra familia de algoritmos nos encontramos con los algoritmos de tipo no supervisados, los cuales también pueden ser reconocidos como algoritmos de descubrimiento de patrones, popularmente denominados algoritmos de clustering.

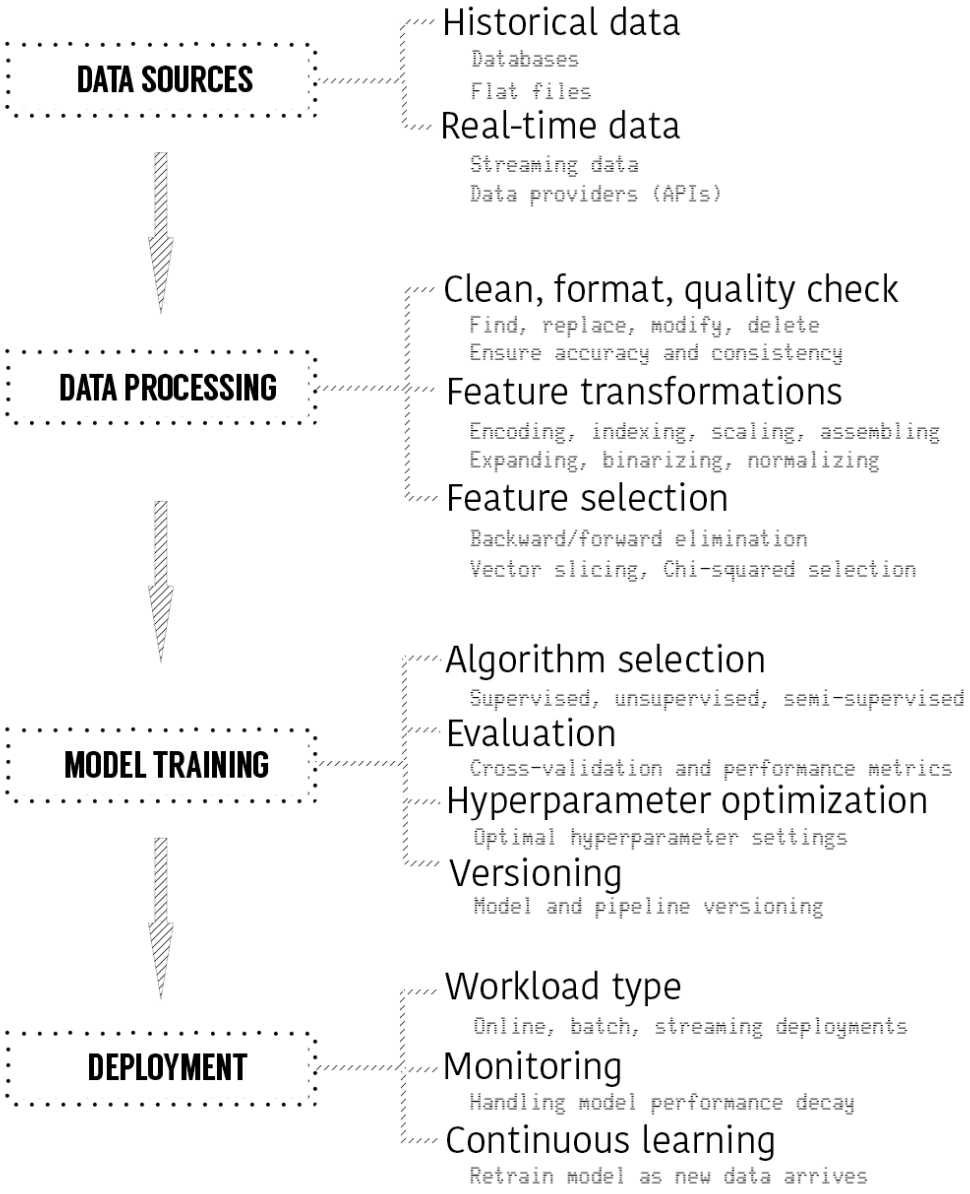
Learning Type	Technique	Model	Algorithm
Unsupervised	Clustering	Partition	K-Means
			K-medoids
		Hierarchical	Agglomerative
			Divisive
	Association	Association	Association
	Dimension Reduction	Principal Component Analysis	PCA

Figura 3-2 Algoritmos no supervisados.

Metodología de Trabajo en ML

ML tambien tiene una metodologia de trabajo bien definida, en donde uno de los principales pasos es el tratamiento adecuado de los datos, es decir limpieza, transformación y algunas tecnicas mas avanzadas como la estandarización, pasando por la revision de la calidad y capacidad predictiva mediante diferentes métodos en donde se evalúan aspectos como exactitud, recall y otros parámetros, hasta lograr un modelo aceptable el cual pueda ser puesto a disposición de la organización.

En la siguiente figura se puede apreciar mas detalle de las tareas a realizar en cada paso del ciclo.



Herramientas para ML

Las herramientas para soportar las actividades de ML son una gran cantidad, entre las más populares destacan:

- Lenguaje R
- Python
- Weka
- Knime
- RapidMiner
- Azure ML Studio
- TensorFlow
- BigML
- SkyTree
- IBM Watson
- MLIB Spark
- Julia
- Jupyter
- Etc.

Este curso utilizaremos principalmente lenguaje R y Weka ya que son herramientas al alcance de todos y tienen una gran comunidad que apalanca su desarrollo y mejora continua.

R es un entorno y lenguaje de programación con un enfoque al análisis estadístico.

Fue creado en 1993 por Robert Gentleman y Ross Ihaka, aunque sus raíces provienen del lenguaje S, desarrollado por John Chambers y Rick Becker en los laboratorios de AT&T. En abril de 2017 se liberó la versión 3.4 presentando mejoras en rendimiento, interfaz de usuario, gestión de memoria y un compilador de código de bytes de tipo JIT (Just in Time).

Por formar parte de un proyecto colaborativo y abierto existe una gran cantidad y variedad de paquetes que permiten que el trabajo en R sea mucho más fácil.

A finales de 2009, la cantidad de paquetes disponibles para trabajar superaba los 2,000 y los temas iban desde estadística bayesiana hasta econometría y series temporales.

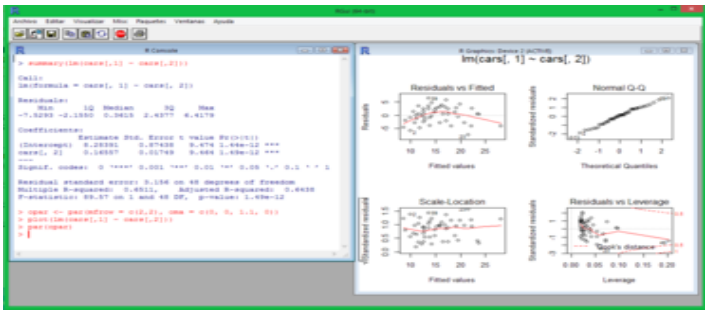


Figura 3-3 Entorno Lenguaje R

Por otro lado, Weka, es una herramienta que ofrece un entorno completo para el análisis del conocimiento y ML, está diseñado para codificar lo menos posible por medio de su extensa colección de técnicas para pre-procesamiento de datos y modelado.

La siguiente tabla muestra todas las técnicas disponibles en Weka

Algorithm	Weka Family	Description
ZeroR	Rules	Este es uno de los métodos más simples; se basa en la tabla de frecuencias y hacer una predicción basada en la moda.
OneR	Rules	Este método se basa en la generación de una regla por cada atributo existente en la base de datos que se está analizando.
Naive Bayes Classifier	Bayes	Este método se basa en el teorema bayesiano, los diferentes atributos se consideran condicionalmente independientes de la variable a pronosticar.
MultiClass Classifier	Meta	Este método se conoce como un meta-clasificador y se aplica a problemas en donde la variable objetivo debe ser clasificada de forma binaria o con múltiples clases.
Input Mapped Classifier	Meta	Este método se utiliza en problemas de clasificación o bien puede ser aplicado como entrada para otro algoritmo.
AdaBoost M1	Meta	Este es un algoritmo basado en arboles de decisión específicamente para problemas de clasificación.
MultiClass Classifier Updateable	Meta	Es un método para problemas de clasificación con multi-clases, aunque también se puede utilizar para problemas de tipo binario. Este algoritmo está basado en el concepto de Gradiente Estocástico SGD (por sus siglas en ingles).
Multilayer Perceptron	Functions	Este método está basado en redes neuronales bajo el concepto de capas.
Decision Stump	Trees	Esta es una técnica basada en arboles de decisión de un nivel. Usualmente se combina con otras técnicas.
Hoeffding Tree	Trees	Técnica basada en arboles de decisión, está diseñado para trabajar bajo escenarios en donde la data para el aprendizaje es limitada.
J48	Trees	Técnica basada en arboles de decisión específicamente del algoritmo diseñado por Ross Quinlan.
Random Tree	Trees	Esta técnica se basa en seleccionar de forma aleatoria diferentes atributos del dataset y luego construir un árbol de decisión.
REP Tree	Trees	Reduces Error Pruning, está basado en la aplicación de regresiones sobre arboles de decisión.
RandomForest	Trees	Este es un bosque de muchos arboles de decisión.
SMO	Functions	Algoritmo basado en el concepto establecido por John Platt para el entrenamiento de Support Vector Machines en problemas de clasificación.

El entorno de trabajo de Weka es sumamente intuitivo y se fundamenta en el concepto de simple – CLI "Simple Command Line Interface".

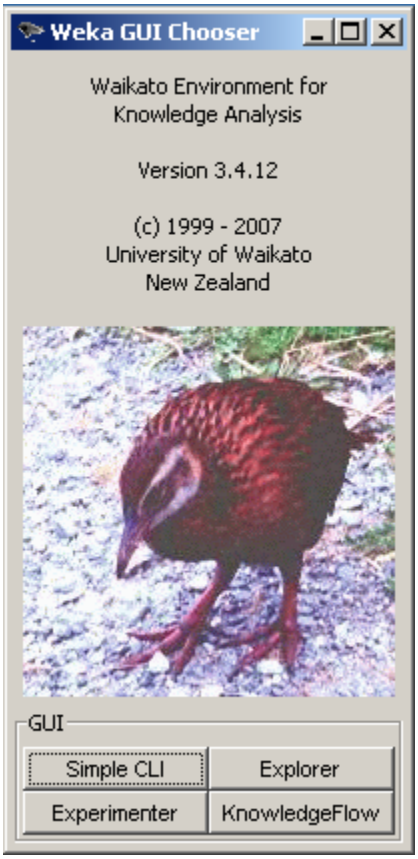


Figura 3-4 Pantalla Inicial Weka

Capítulo 4

Estadística Básica

Los encantos de esta ciencia sublime, las matemáticas, sólo se le revelan en toda su belleza a aquellos que tienen el coraje de profundizar en ella (Carl Friedrich Gauss).

Estadística

La estadística es la parte de las matemáticas que se encarga del estudio de una determinada característica de una población, recogiendo los datos, organizándolos en tablas, representándolos gráficamente y analizándolos para sacar conclusiones[3].

Existen dos tipos de estadística:

- Estadística Descriptiva: Realiza estudios sobre los datos, para resumir la información de la forma más sencilla y presentable posible obteniendo así los parámetros que distinguen las características de un conjunto de observaciones, es decir, trata del recuento, ordenación, clasificación y presentación de los datos.

- Estadística Inferencial: Realiza el estudio sobre un subconjunto de la población llamado muestra y, posteriormente, extiende/infiere los resultados obtenidos a toda la población. En otras palabras, la estadística inferencial utiliza los resultados de la estadística descriptiva y se apoya en el cálculo de probabilidades para la obtención de conclusiones sobre una población a partir de los resultados obtenidos de una muestra.

Como se puede comprender a partir de las definiciones dadas hasta el momento, el elemento diferenciador clave entre estos 2 grandes conceptos es que la estadística descriptiva tal como su nombre lo indica, utiliza cálculos matemáticos para descubrir las características de un grupo de datos, sin embargo no es capaz de decirnos porque estas características ocurren, es aquí donde entra en juego la estadística inferencial.

Antes de proseguir con el tema, debemos hacer notar que el concepto básico de población debe comprenderse como *"el grupo más grande acerca del cual se desea comprender algo"*, por ejemplo en una empresa, el grupo de todos los empleados podría ser una población. Una muestra es un grupo bajo estudio más pequeño perteneciente a una población, por ejemplo, todos los empleados que atienden directamente a los clientes.

[3] http://recursostic.educacion.es/descartes/web/materiales_didacticos/unidimensional_lbarrios/definicion_est.htm

Adicional hay que tener en mente que en la jerga estadística, cuando se analiza una muestra, los resultados y cálculos obtenidos se denominan **estadísticos**, pero cuando se obtienen conclusiones para toda una población, estos se denominan **parámetros.**

Esta diferencia entre términos y parámetros debe tenerse en cuenta ya que hasta las fórmulas matemáticas podrían tener ligeras variaciones a raíz de este tema.

A continuación se listaran los términos básicos más utilizados en el ámbito estadístico[4]:

- Población: Conjunto de elementos que se quiere estudiar.
    - Habitantes de un país
    - Alumnos en una universidad
    - Prendas de vestir fabricadas
- Muestra: Cualquier sub-conjunto de una población
- Variable estadística: Cada una de las características que se requiere estudiar de los elementos que componen la población en estudio. Estas variables pueden ser cuantitativas o cualitativas.
    - Salario Mensual: 1000, 300, 700, 500
    - Zona en que habita: Centro, Norte, Sur
    - Edad: 15, 18, 25, 13
- Individuo: Cada uno de los elementos que componen una población o muestra.
- Estadístico: Medida descriptiva de una muestra

Variables y sus clasificaciones

Una variable estadística es una característica que puede fluctuar y cuya variación es susceptible de adoptar diferentes valores, los cuales pueden medirse u observarse[5]. Las variables adquieren valor cuando se relacionan con otras variables y dependiendo de su naturaleza se pueden clasificar en diferentes tipos:

[4] https://es.slideshare.net/marketing2009/estadstica-descriptiva-e-inferencial

[5] https://es.wikipedia.org/wiki/Variable_estadistica

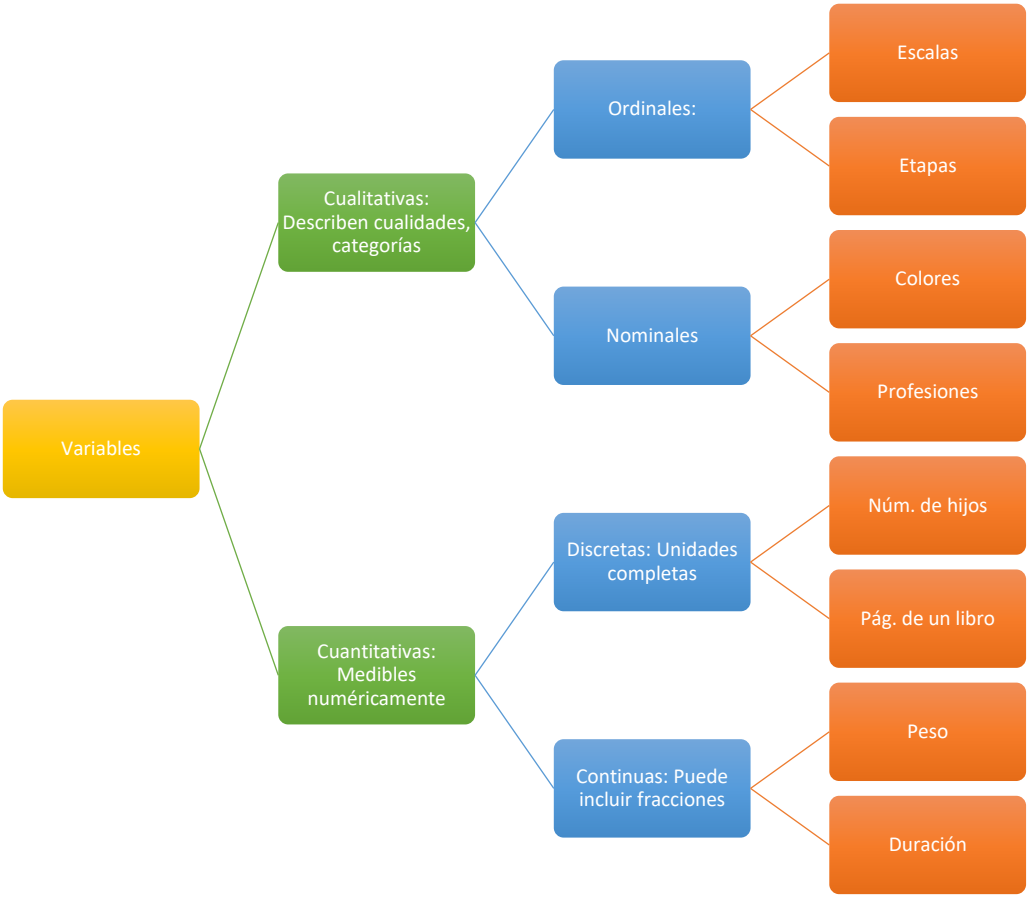


Figura 4-1 Tipos de variables.

Medidas de Tendencia Central

Las medidas de tendencia central, son medidas estadísticas que pretenden resumir en un solo valor a un conjunto de valores.

Nos sirven para describir características básicas de un conjunto de datos que contiene variables cuantitativas, y el objetivo de estas es sintetizar los datos.

Las medidas de tendencia central más comunes son:

- La media: Valor obtenido al sumar todos los datos y dividirlos por el número total de datos
- La mediana: Valor que ocupa el lugar central de todos los datos, ordenados de mayor a menor.
- La moda: Valor con la mayor frecuencia.

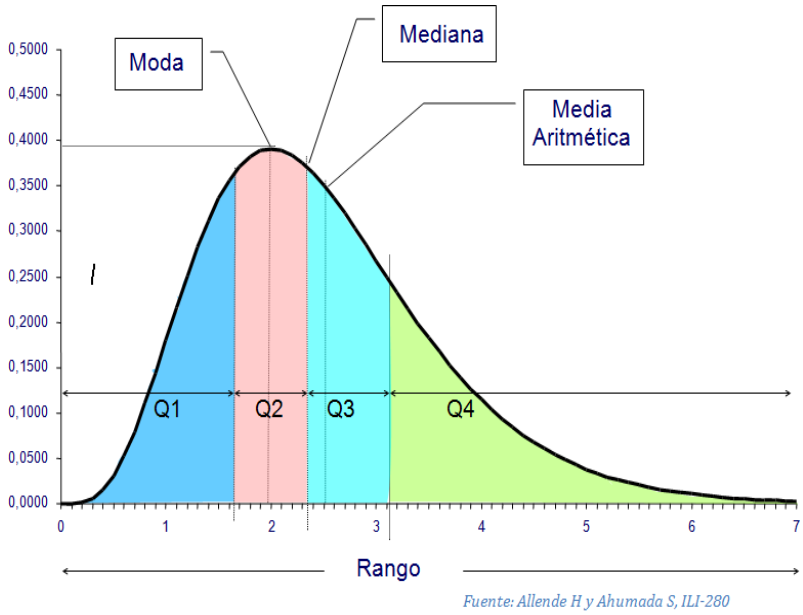


Figura 4-2 Medidas Tendencia Central.

Sin embargo, a pesar de que las medidas de tendencia central son las más populares, el análisis estadístico básico comprende:

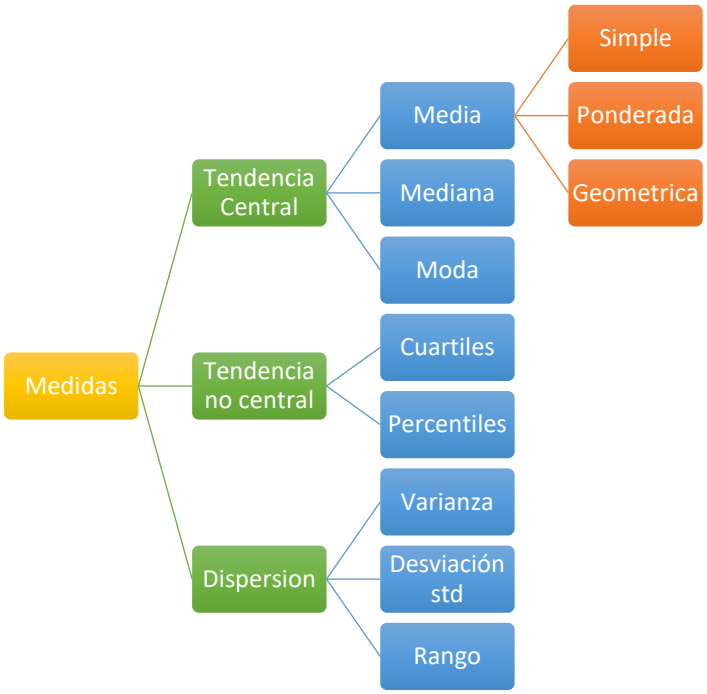


Figura 4-3 Familia de Medidas.

La media

En su expresión más simple, la media es comúnmente conocida como "el promedio", es el resultado de sumar todos los datos y dividir entre el número total de datos[6].

$\bar{X}$ es el símbolo de la **media aritmética.**

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{N}$$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{N}$$

Ejemplo: Los pesos en kilogramos de seis personas son: 84, 91, 72, 68, 87 y 78. ¿Cuál es el promedio?

$$\bar{X} = \frac{84 + 91 + 72 + 68 + 87 + 78}{6} = 80\ Kg$$

Las propiedades de la media aritmética son las siguientes:

- La suma de las desviaciones de las puntuaciones de una distribución respecto a la media es igual a cero

$$\sum(X_i - \bar{X}) = 0$$

- La suma de los cuadrados de las desviaciones de los valores de una variable con respecto a un número cualquiera se hace mínima cuando dicho número coincide con la media aritmética.

$$\sum(X_i - \bar{X})^2\ Mínimo$$

- Si a todos los valores de una variable se les suma un mismo número, la media aritmética queda aumentada en dicho número.
- Si todos los valores de una variable se multiplican por un mismo número, la media aritmética queda multiplicada por dicho número.
- La media se puede hallar solo para variables cuantitativas.

[6] https://www.vitutor.com/estadistica/descriptiva/a_10.html

- La media es independiente de las amplitudes de los intervalos
- La media es muy sensible a las puntuaciones extremas
- La media no se puede calcular si hay un intervalo con una amplitud indeterminada

Adicional a la media aritmética simple también existe:

- Media Aritmética ponderada: Esta se calcula asignándole a cada clase un peso, y obteniendo un promedio de los pesos, teniendo estos pesos valores diferentes, la diferencia cuando calculamos la media aritmética es que a todos los pesos se las da un mismo valor.

$$\overline{X}_w = \frac{\sum_{i=1}^{k} x_i w_i n_i}{\sum_{i=1}^{k} w_i n_i}$$

- Media Geométrica: La media geométrica de un conjunto de datos es el resultado de multiplicarlos entre si y aplicar la N-enésima raíz.

$$\overline{x} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots x_n}$$

Como puede observarse no utiliza la suma como la media aritmética sino que utiliza la multiplicación.

La mediana

Una variable estadística es

La moda

Una variable estadística es

Cuartiles

Los cuartiles son los 3 valores de la variable que dividen a un conjunto de datos ordenados en 4 partes iguales[7].

Q1, Q2 y Q3 determinan los valores correspondientes al 25%, 50%, y 75% de los datos.

Para calcularlos se debe primero ordenar los datos de menor a mayor y luego buscar el lugar que ocupa cada cuartil mediante la expresión:

[7] https://www.vitutor.com/estadistica/descriptiva/a_11.html

$$\frac{k \cdot N}{4},\ k = 1,2,3$$

Y si estamos ante una tabla de frecuencias acumuladas se deberá utilizar.

$$Q_k = L_i + \frac{\frac{k \cdot N}{4} - F_{i-1}}{f_i} \cdot a_i$$

Percentiles

El percentil es una medida de posición usada en estadística que indica, una vez ordenados los datos de menor a mayor, el valor de la variable por debajo del cual se encuentra un porcentaje dado de observaciones en un grupo de observaciones[8].

Los percentiles son los 99 valores que dividen la serie de datos en 100 partes iguales y proporcionan los valores correspondientes al 1%, al 2%... y al 99% de los datos[9].

P50 coincide con la mediana.

$$\frac{k \cdot N}{100},\quad k = 1,2,..99$$

$$P_k = L_i + \frac{\frac{k \cdot N}{100} - F_{i-1}}{f_i} \cdot a_i$$

[8] https://es.wikipedia.org/wiki/Percentil

[9] https://www.vitutor.com/estadistica/descriptiva/a_13.html

Para efectos de estudio en este material también revisaremos los deciles, los cuales corresponden a los 9 valores que dividen la serie de datos en 10 partes iguales, los deciles nos dan los valores correspondientes al 10%, al 20% y al 90% de los datos[10].

Vale la pena mencionar que el decil 5 (D5) corresponde a la mediana.

La fórmula para calcular los deciles es:

$$\frac{k \cdot N}{10}, \quad k = 1, 2, ..9$$

Y para frecuencias acumuladas

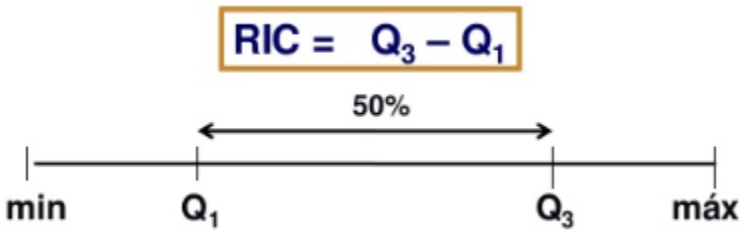
$$D_k = L_i + \frac{\frac{k \cdot N}{10} - F_{i-1}}{f_i} \cdot a_i$$

Rango

El rango representa la diferencia entre el valor máximo y el valor mínimo de un conjunto de datos. El rango nos muestra la distribución de los valores en una serie. Si el rango es un número muy alto, entonces los valores de la serie están bastante distribuidos. En cambio, si se trata de un número pequeño, quiere decir que los valores de la serie están muy cerca entre sí[11].

$$R = x_{(k)} - x_{(1)}$$

Por otro lado, se le llama rango intercuartílico o rango intercuartil, a la diferencia entre el tercer y el primer cuartil de una distribución.



[10] https://www.vitutor.com/estadistica/descriptiva/a_12.html

[11] https://es.wikihow.com/calcular-el-rango-estad%C3%ADstico

Varianza

La varianza es la media aritmética del cuadrado de las desviaciones respecto a la media de una distribución estadística[12].

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{N}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{N}$$

Algunas propiedades básicas de la varianza se numeran a continuación:

- La varianza siempre será un valor positivo o cero
- Si a todos los valores de la variable se les suma un número la varianza no cambia
- Si todos los valores de las variables se multiplican por un número, la varianza queda multiplicada por el cuadrado de dicho número.
- La varianza al igual que la media es un índice muy sensible a las puntuaciones extremas
- En los casos en los que no se puede hallar la media, tampoco se podrá encontrar la varianza
- La varianza no viene expresada en las mismas unidades que los datos, ya que las desviaciones están elevadas al cuadrado.

Desviación Estándar

La desviación estándar es la medida de dispersión más común, que indica qué tan dispersos están los datos con respecto a la media. Mientras mayor sea la desviación estándar, mayor será la dispersión de los datos[13].

El símbolo σ (sigma) se utiliza frecuentemente para representar la desviación estándar de una población, mientras que s se utiliza para representar la desviación estándar de una muestra. La variación que es aleatoria o natural de un proceso se conoce comúnmente como ruido.

[12] https://www.vitutor.com/estadistica/descriptiva/a_15.html

[13] https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/data-concepts/what-is-the-standard-deviation/

En palabras simples, la desviación estándar es la raíz cuadrada de la varianza.

$$\sigma = \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{N}} \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{N}}$$

Conocidas la media y la desviación estándar se puede proceder a calcular el coeficiente de variación.

El coeficiente de variación, también denominado como coeficiente de variación de Spearman, es una medida estadística que nos informa acerca de la dispersión relativa de un conjunto de datos. Su cálculo se obtiene de dividir la desviación típica entre el valor absoluto de la media del conjunto y por lo general se expresa en porcentaje para su mejor comprensión[14].

El coeficiente de variación se utiliza para comparar conjuntos de datos pertenecientes a poblaciones distintas. Si atendemos a su fórmula, vemos que este tiene en cuenta el valor de la media. Por lo tanto, el coeficiente de variación nos permite tener una medida de dispersión que elimine las posibles distorsiones de las medias de dos o más poblaciones.

$$C.V = \frac{\sigma}{\bar{X}}$$

Otros valores y estadísticos.

- Valor P: El valor P es una medida de la fuerza de la evidencia en sus datos en contra de la hipótesis nula. Por lo general mientras más pequeño sea el valor P, más fuerte será la evidencia para rechazar la hipótesis nula. Tradicionalmente el valor P se compara con valores menores que 0.05 o 0.01, dependiendo del campo de estudio[15].

- Valor T: Un valor t es el resultado de una prueba estadística. El valor se encuentra en la distribución t de Student que es apropiado para los grados de libertad. La ubicación especifica la probabilidad de obtener el valor t por casualidad. Si la probabilidad es menor que el nivel de significación, el resultado se juzga que es estadísticamente significativo[16].

  En estadística, una prueba t de Student, prueba t de estudiante, o Test-T es cualquier prueba en la que el estadístico utilizado tiene una distribución t de Student si la hipótesis nula es cierta. Se aplica cuando la población estudiada sigue una distribución normal pero el tamaño muestral es demasiado pequeño como

[14] http://economipedia.com/definiciones/coeficiente-de-variacion.html

[15] http://www.sinestetoscopio.com/el-famoso-valor-p/

[16] https://www.ibm.com/support/knowledgecenter/es/SS4QC9/com.ibm.solutions.wa_an_overview.2.0.0.doc/t_value.html

para que el estadístico en el que está basada la inferencia esté normalmente distribuido, utilizándose una estimación de la desviación típica en lugar del valor real. Es utilizado en análisis discriminante[17]. Generalmente un valor T es aceptable si es mayor que +2 y menor que -2.

La Distribución Normal

La distribución normal es la más importante de todas las distribuciones de probabilidad. Es una distribución de variable continua cuyo rango es del menos infinito al más infinito.
Fue descubierta por Gauss al estudiar los errores en las observaciones astronómicas.

La popularidad se debe a tres razones principales:

- La gran cantidad de fenómenos reales que se pueden modelizar con esta distribución.
- Muchas de las distribuciones de uso frecuente tienden a aproximarse a la distribución normal bajo ciertas condiciones.
- En virtud del teorema central del límite, todas aquellas variables que puedan considerarse causadas por un gran número de pequeños efectos tienden a distribuirse con una distribución normal.

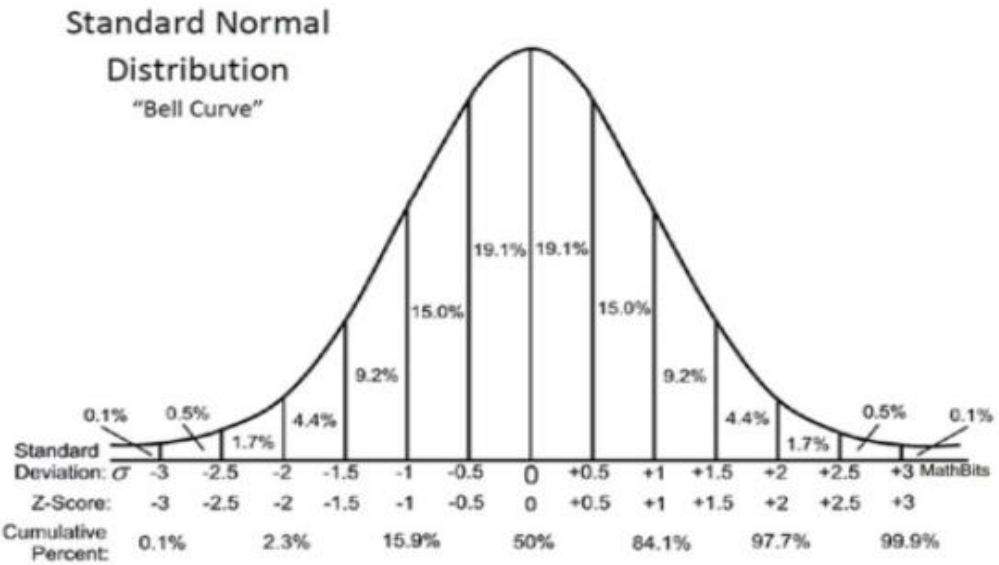


Figura 4-4 Grafica de la función de densidad de la distribución normal

La distribución de probabilidad esta expresada de la siguiente manera:

$$f(X)=\left(\frac{1}{\sigma_X\sqrt{2\pi}}\right)\exp\left(\frac{-(X-\mu_X)^2}{2\sigma_X^2}\right)=\left(\frac{1}{\sigma_X\sqrt{2\pi}}\right)e^{\left(\frac{-(X-\mu_x)^2}{2\sigma_x^2}\right)}$$

[17] https://es.wikipedia.org/wiki/Prueba_t_de_Student

Una característica distintiva de una distribución normal es la probabilidad (o densidad) asociado con segmentos específicos de la distribución. La distribución normal en la figura está dividida en los intervalos más comunes (o segmentos): uno, dos, y tres desviaciones estándar de la media[18].

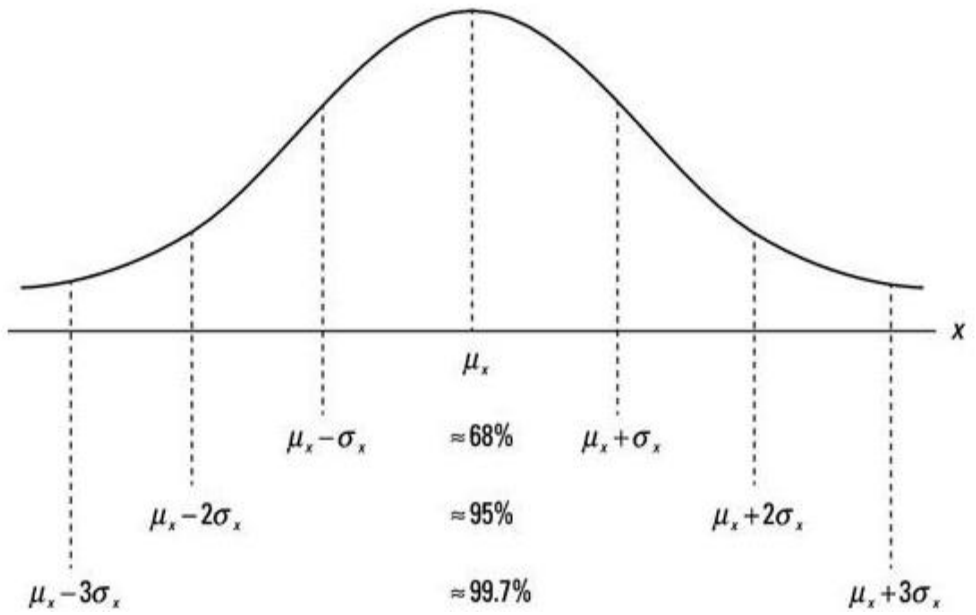


Figura 4-5 Intervalos más comunes de la distribución normal

La distribución normal en la figura está dividida en los intervalos Con una variable aleatoria normalmente distribuida, aproximadamente el 68 por ciento de las mediciones están dentro de una desviación estándar de la media, el 95 por ciento son dentro de dos desviaciones estándar, y el 99,7 por ciento son dentro de tres desviaciones estándar.

Estandarización de Datos

La estandarización de datos es importante ya que en la mayoría de casos nos encontraremos que dentro de un mismo set de datos, los atributos tienen diferente naturaleza, origen y forma de medición, en otras palabras, si los datos no son estandarizados estos no serían comparables, por ejemplo, supongamos que estamos analizando una muestra de la población para determinar va a adquirir cierto tipo de producto que se lanzara al mercado, en la base de datos seguramente encontraremos la edad de la persona, su género, un rango salarial, cantidad de miembros de su familia, frecuencia de visita al supermercado y por supuesto la variable objetivo.

Como puede observarse la cantidad y naturaleza de cada variable es totalmente diferente, algunas son referente a tiempo tal como la edad, otras están en dólares, otros cuentan personas, etc.

[18] http://maniqui.ru/educacin-y-lenguas/economa-y-finanzas/econometra/16125-reconociendo-las-variables-habituales-distribucin.html

Para solventar esto, es apropiado aplicar transformaciones a nuestros datos en análisis y así garantizar la exactitud del modelo.

- Escalamiento por base decimal: Se basa en la transformación
    - $X' = X/(10 \wedge h)$, h es el parámetro que determina la intensidad del escalamiento que se aplicara, el valor transformado estará en el rango [-1,1]
- Mínimos y máximos: esta transformación se basa en el mínimo y máximo del set de datos en análisis y su salida siempre se espera en el rango [-1,1]
- Índice Z: Esta transformación se basa en el uso de la media y la desviación estándar de la variable a analizar
    - $X' = \frac{X-\mu}{\sigma}$, si la distribución es normal o cercana a esta, esta transformación devolverá valores en el rango [-3,3]

Vale pena señalar que, la estandarización no cambia la forma de la distribución de los datos, si bien es cierto al aplicar la transformación, la media se sitúa en cero y la desviación estándar cambia a uno, la curva que describe la distribución no cambia.

Por último, para cerrar este tema, hay que mencionar que una normalización es apropiada para distribuciones unimodales y más aún si es una distribución simétrica.

Capítulo 5

Análisis Exploratorio

Para la mayoría de los estudiantes la estadística es un tema misterioso donde operamos con números por medio de fórmulas que no tienen sentido (Graham).

AED (Análisis Exploratorio de Datos)

Independientemente de la complejidad de los datos disponibles y del procedimiento estadístico que se tenga intención de utilizar, una exploración minuciosa de los datos previa al inicio de cualquier análisis posee importantes ventajas que un analista no puede pasar por alto[19].

Una exploración minuciosa de los datos permite identificar entre otras cosas:

- Posibles errores (datos mal introducidos, respuestas mal codificadas, etc.)
- Valores extremos (valores que se alejan demasiado del centro)
- Pautas extrañas en los datos (valores que se repiten demasiado o que no aparecen nunca, etc.)
- Variabilidad no esperada
- Etc.

La finalidad del AED es examinar los datos previamente a la aplicación de cualquier técnica estadística. De esta forma se consigue un entendimiento básico de los datos y de las relaciones existentes entre las variables analizadas.

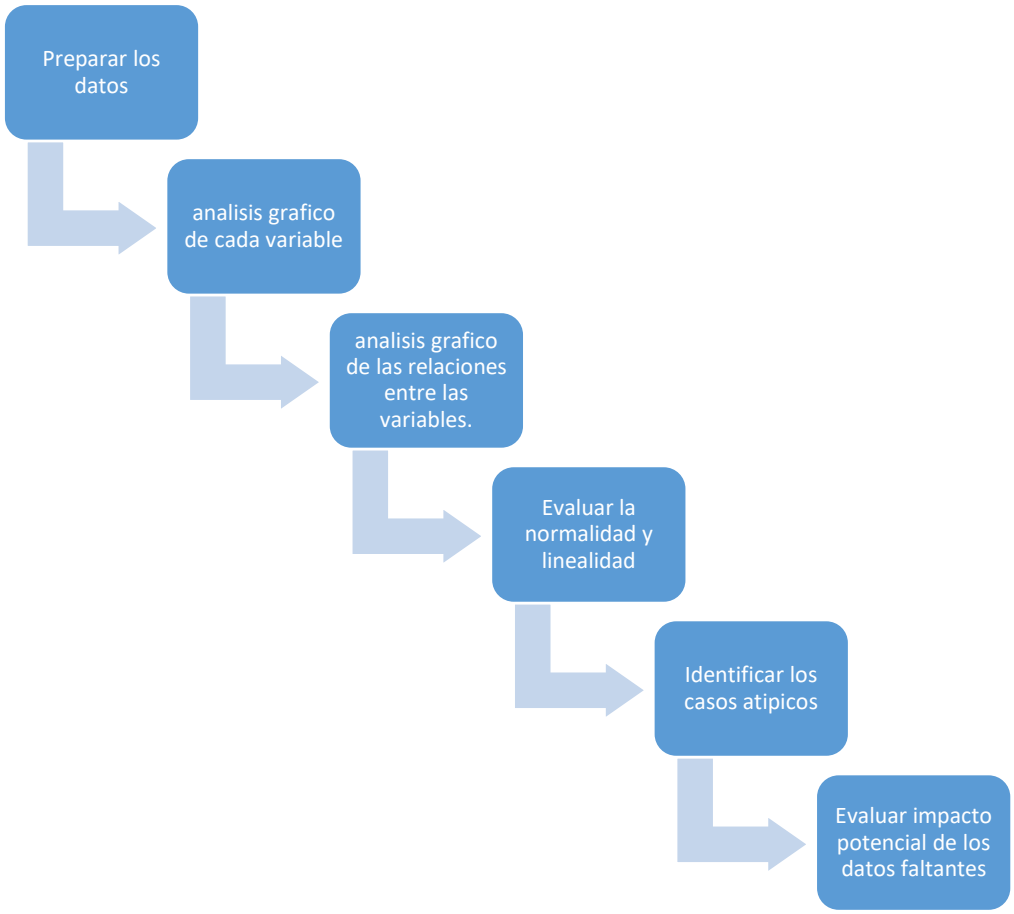
El AED proporciona métodos sencillos para organizar y preparar los datos, detectar fallos en el diseño y recogida de los datos, tratamiento y evaluación de datos ausentes, identificación de casos atípicos y comprobación de los supuestos subyacentes en la mayor parte de las técnicas multivariantes.

El examen previo de los datos es un paso mandatorio, que lleva tiempo, y que habitualmente se descuida por parte de los analistas de datos. Las tareas implícitas en dicho examen pueden parecer insignificantes y sin consecuencias a primera vista, pero son una parte esencial de cualquier análisis estadístico[20].

[19] http://www.ugr.es/~fmocan/MATERIALES%20CURSO/Exploratorio.pdf

[20] https://ciberconta.unizar.es/leccion/aed/ead.pdf

Las etapas para el AED son:



.

Figura 5-1 Etapas del AED

La inmensa mayoría de los paquetes estadísticos permite realizar manipulaciones de los datos previas a un análisis de los mismos. Algunas operaciones útiles son las siguientes:

- Combinar conjuntos de datos de archivos distintos
- Seleccionar sub-conjunto de datos
- Dividir el archivo en varias partes
- Transformar variables
- Ordenar casos
- Agregar nuevos datos o variables

- Eliminar datos o variables
- Guardar datos o resultados.

Tal como se ve en la gráfica inicial, una vez organizados los datos, el segundo paso es prácticamente realizar un análisis gráfico y numérico de cada una de las variables del problema con el fin de tener una idea inicial de la información contenida en el conjunto de datos.

Este tipo de análisis se puede realizar tomando como referencia la siguiente tabla:

Escala de Medida	Tipo de Gráfica	Medidas Tendencia Central	Medidas de Dispersión
Nominal	Diagrama de barras Diagrama de líneas Diagrama de sectores	Moda	
Ordinal	Boxplot	Mediana	Rango Intercuartilico
Intervalo	Histograma	Media	Desviación
Razón		Media Geométrica	Coefic. De Variación.

Por otro lado, las variables cualitativas son parte importante de este tipo de análisis, generalmente los datos correspondientes a variables cualitativas se agrupan de manera natural en diferentes categorías o clases y se cuenta el número de datos que aparece en cada una de ellas.

Se suelen representar mediante diagramas de barras, pastel o líneas.

Una vez realizado el análisis multidimensional el siguiente paso consiste en analizar la existencia de posibles relaciones entre variables, dicho estudio puede llevarse a cabo desde una óptica bidimensional o multidimensional.

Las tres situaciones generales que pueden darse son:

- Ambas variables son cualitativas
- Ambas variables son cuantitativas
- Una variable es cualitativa y otra cuantitativa

El caso más complejo de analizar es el segundo en donde se tienen 2 variables numéricas, por lo que el primer paso es representar gráficamente el fenómeno mediante un diagrama de dispersión.

Por otro lado si lo que queremos es investigar si los datos siguen a la distribución normal se puede utilizar el grafico de tipo QQ-Plot.

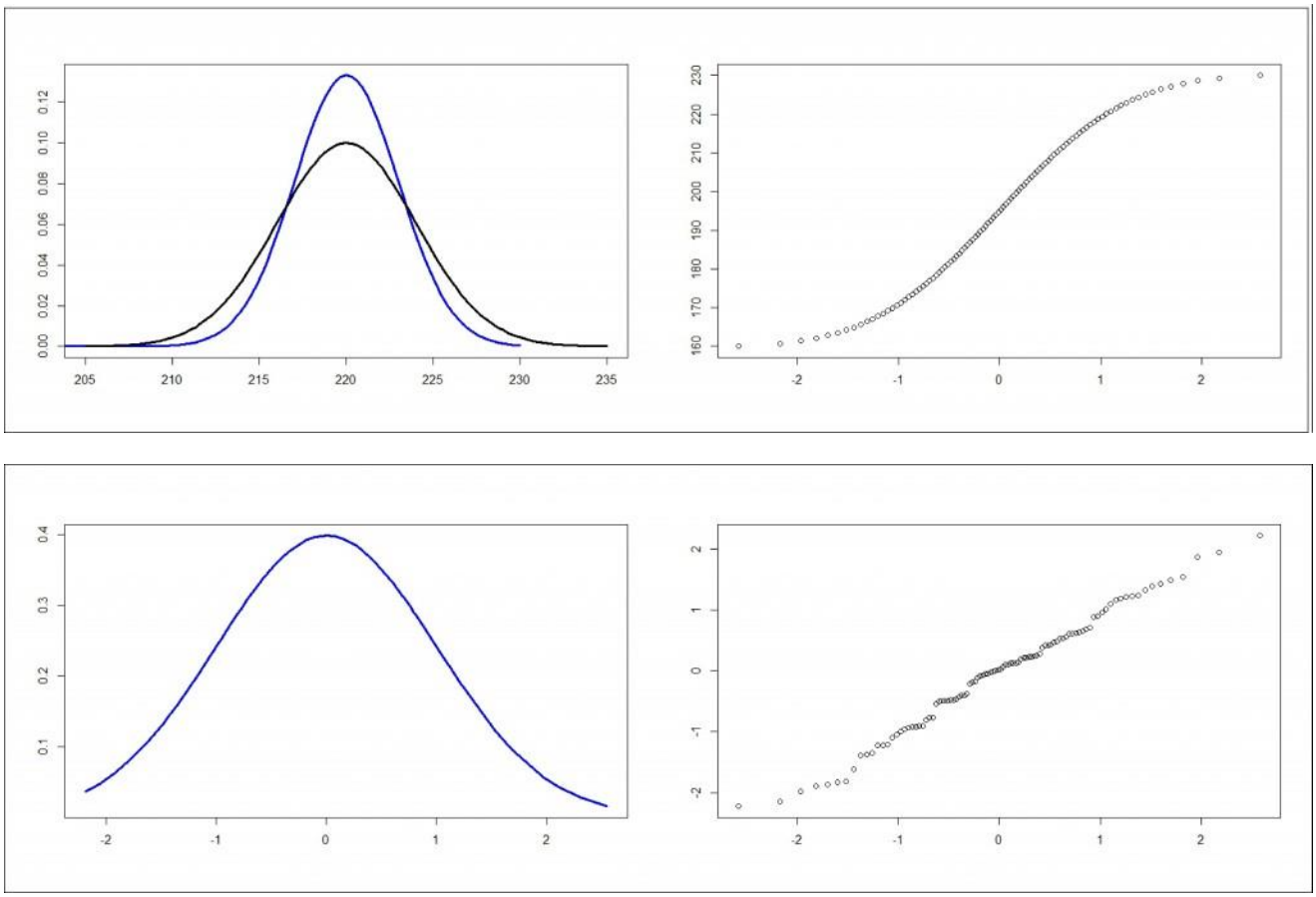


Figura 5-2 QQ Plot

Capítulo 6

Data Quality

Data sin revisar, pobremente organizada y sin actualizar es con frecuencia la responsable de malas decisiones, perdidas de oportunidades y altos costos.

Data Quality

Iniciaremos este capítulo hablando de Data Quality, la cual se puede definir como el conjunto de técnicas/metodologías para mantener la información de las organizaciones, completa, precisa, consistente, actualizada, única y valida[21].

La calidad de los datos es tarea tanto de quienes producen los datos así como de quienes la almacenan:

- Data Providers: Son la fuente de la data, es decir se encuentran en el punto inicial del proceso de creación de datos de calidad.

- Data recipients: Tienen la responsabilidad de almacenar la data en sus sistemas y garantizar la integridad de la misma y en sus procesos.

Para las organizaciones los datos constituyen su principal activo por lo que datos de mala calidad suelen ser muy costosos en tiempo y dinero y por ende la toma de decisiones se ve directamente afectada.

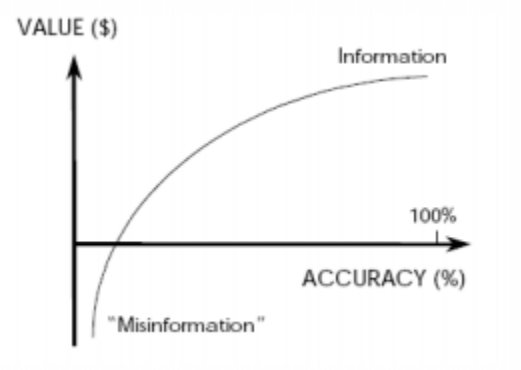


Figura 6-1 Relación DQ y Valor

[21] https://www.marketingdirecto.com/punto-de-vista/la-columna/data-quality-sergio-rodrigo

El desafío antes planteado requiere que las expectativas de Data Quality y las expectativas de negocio estén alineadas por lo que algunas preguntas a contestar son en este tipo de análisis son:

- ¿Cuantos o que datos hacen falta o están inutilizables?
- ¿Cuáles valores están en conflicto?
- ¿Qué registros están duplicados?
- ¿Qué relaciones de los datos hacen falta?

Por otro lado, las expectativas de negocio de cara a la calidad de los datos vienen expresadas con las siguientes preguntas:

- ¿Cómo/Cuanto ha disminuido el rendimiento debido a los errores en los datos?
- ¿Qué porcentaje del tiempo es dedicado a reprocesos debido a la calidad de los datos?
- ¿Cuánto valor en términos monetarios se está perdiendo debido a procesos de negocio que fallan por causa de la mala calidad de los datos?
- ¿Qué tan rápido podemos responder a las oportunidades de negocio dada la baja calidad de los datos?

Como se puede observar, la calidad de los datos tiene una alta importancia para las organizaciones y se hace todo lo posible para evitar el fenómeno del Basura Entra-Basura Sale.

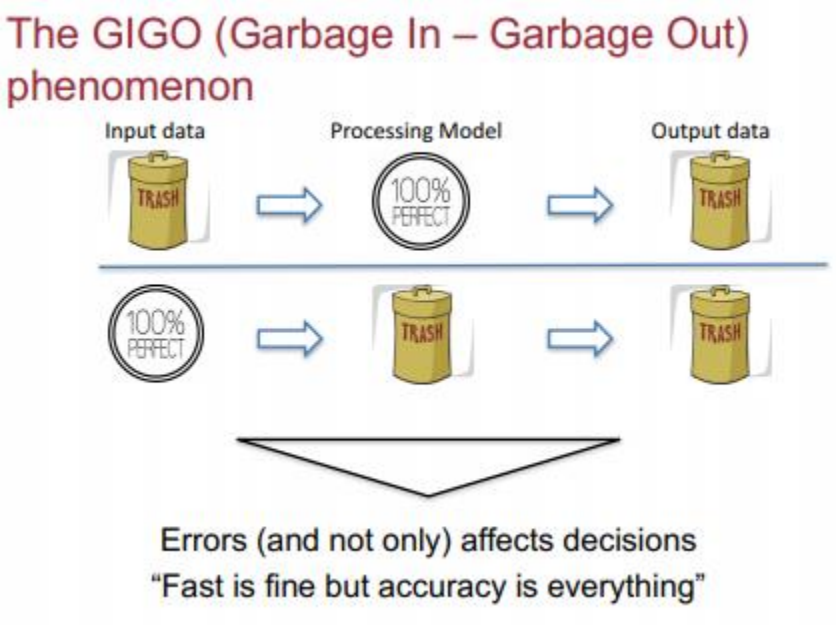


Figura 6-2 Fenómeno GIGO

Pero, ¿Por qué tenemos problemas con Data Quality?, existen diversas causas/etapas las cuales pueden agruparse en 3 grandes categorías.

- Producción de Datos:
    - Diferentes fuentes para hacer referencia a un mismo dato.
    - Entradas de datos con elementos subjetivos (ej. Typos)
    - Problemas sistemáticos en la captura de datos tales como asignación de códigos incorrectos a un catálogo.
- Almacenamiento:
    - Datos almacenados en diferentes formatos
    - Insuficientes formatos
- Uso de los datos:
    - Capacidad insuficiente de procesamiento y análisis de los datos.
    - Problemas de seguridad y acceso.

Para poder subsanar los problemas antes planteados, es necesario que se tome en cuenta la administración de Data Quality como parte fundamental de los procesos de negocio. Dicha administración, se puede dividir en cuatro grandes fases:

- Establecer las dimensiones de calidad.
- Evaluar las dimensiones de calidad.
- Análisis de los problemas de calidad
- Mejora continua de la calidad de la data.

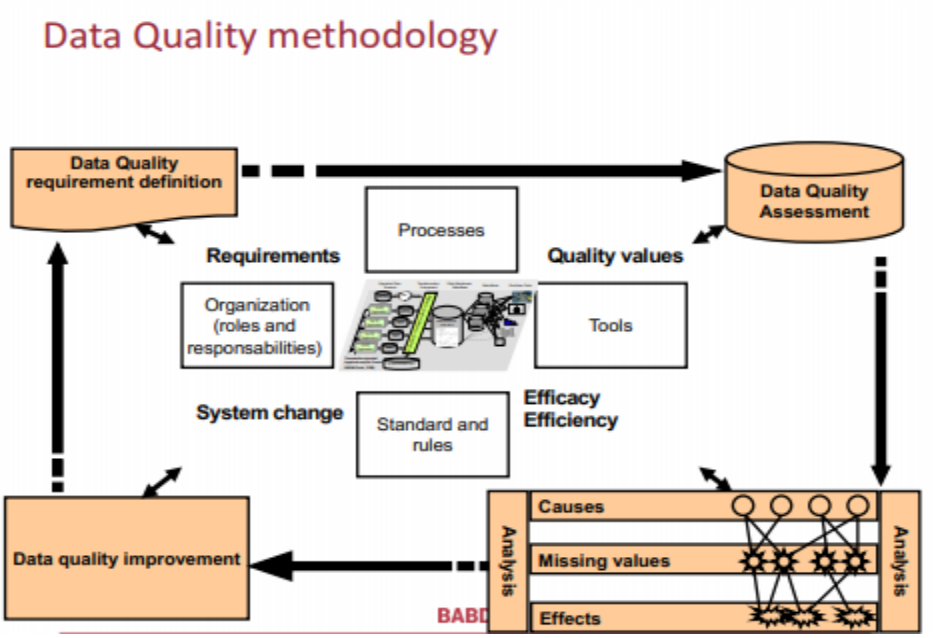


Figura 6-3 DQ Metodología

Ahora bien, existen más de 179 dimensiones para cuantificar el nivel de calidad de datos que se posee, sin embargo, esto no significa que todas deben medirse para poder dar una respuesta.

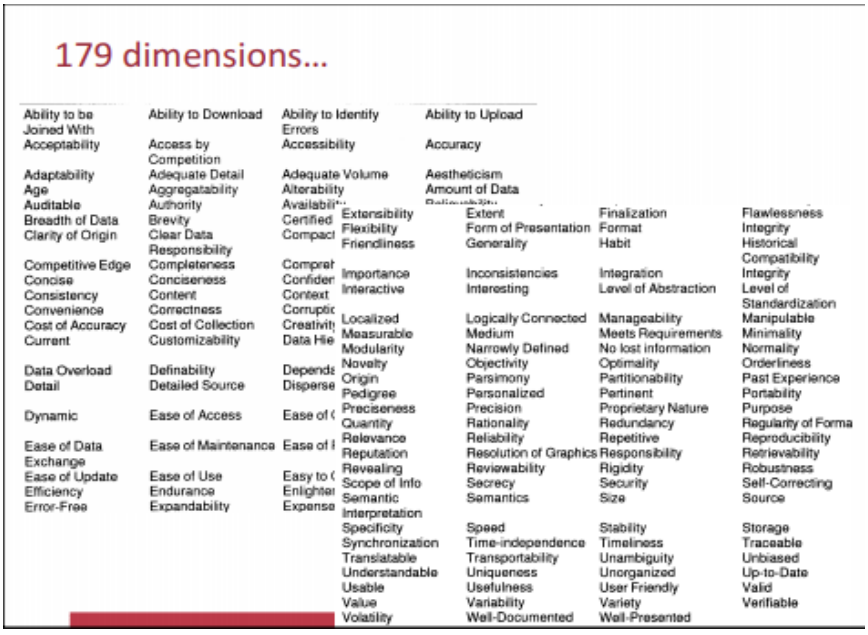


Figura 6-4 Dimensiones de Data Quality

Las dimensiones más utilizadas son las que propuso Richard Wang y Diane Strong en 1996:

Data Quality dimensions			
Intrinsic dimensions	**Contextual Dimensions**	**Representational Dimensions**	**Accessibility Dimensions**
Believability Accuracy Objectivity Reputation	Value-added Relevance Completeness Timeliness Appropriate amount of data	Interpretability Ease of understanding Representational Consistency Concise representation	Accessibility Access security

Figura 6-5 Propuesta en 1996 de Wang y Strong

Para efectos prácticos de los objetivos que persigue este documento, se tendrá un enfoque en 4 dimensiones principales:

- Exactitud: Es la medida que indica si la data esta correcta y confiable.
- Completitud: Es la medida que indica cuantos valores están completos, acá se le hace frente al problema de los valores nulos.
- Consistencia: Es la medida que indica el cumplimiento de las reglas semánticas y de negocio.
- Actualización: Es la medida que indica si los datos están lo suficientemente actualizados para una tarea en particular.

Algunos autores agregan una dimensión más con respecto a la industria[22]:

- Basada en estándares: Es la medida que indica que la data está conforme a los estándares de la industria

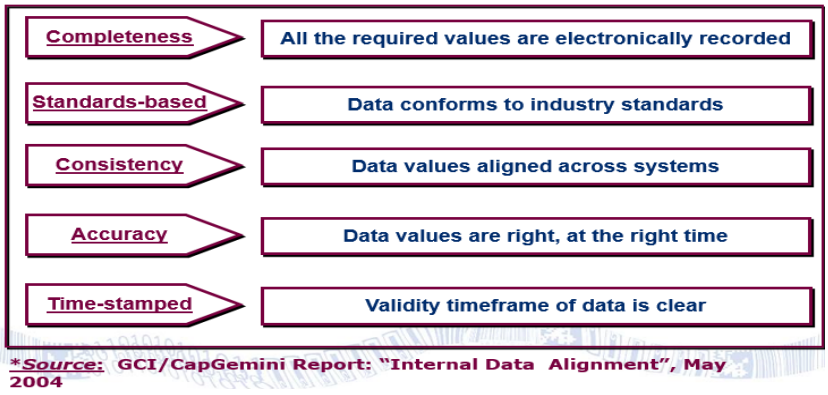


Figura 6-6 Definición de Data Quality

[22] https://www.gs1.org/services/data-quality/data-quality-framework

Pero, ¿Cómo se puede producir data de calidad? Veamos el siguiente diagrama que muestra los elementos claves para alcanzar este cometido.

Figura 6-7 Elementos para producir data de calidad

Al unificar todo lo antes planteado en la metodología de Data Quality, será posible plantear una estrategia de largo plazo que nos permitirá tener una mejor optimización de costos sostenible en el tiempo.

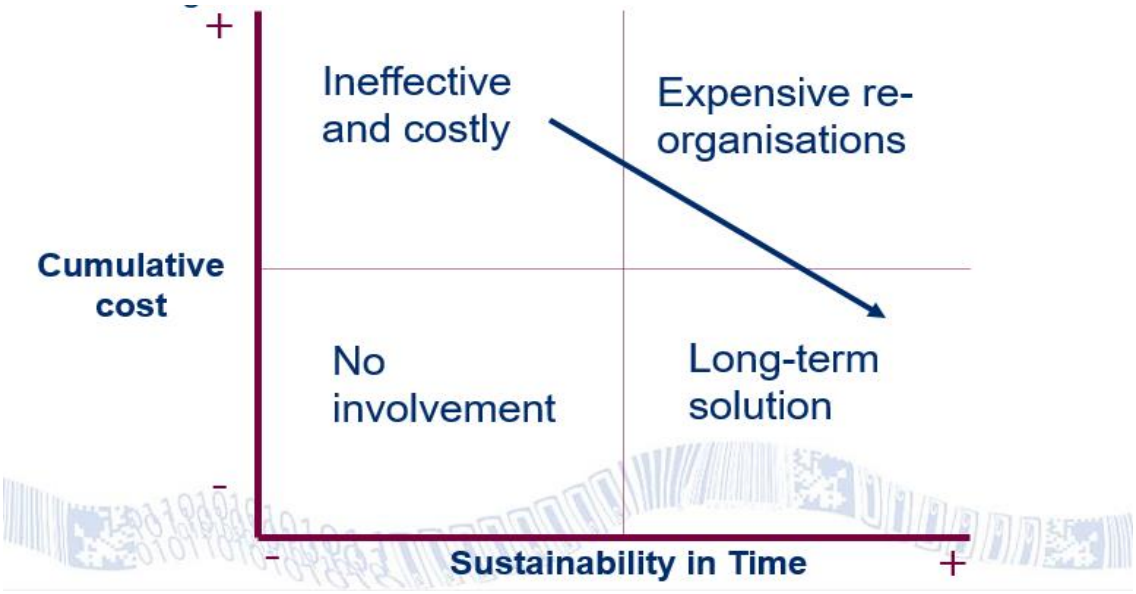


Figura 6-8 Sostenibilidad de DQ

Data Quality, un Proceso de 6 Pasos

A continuación se definen los 6 pasos básicos del proceso de Data Quality[23]:

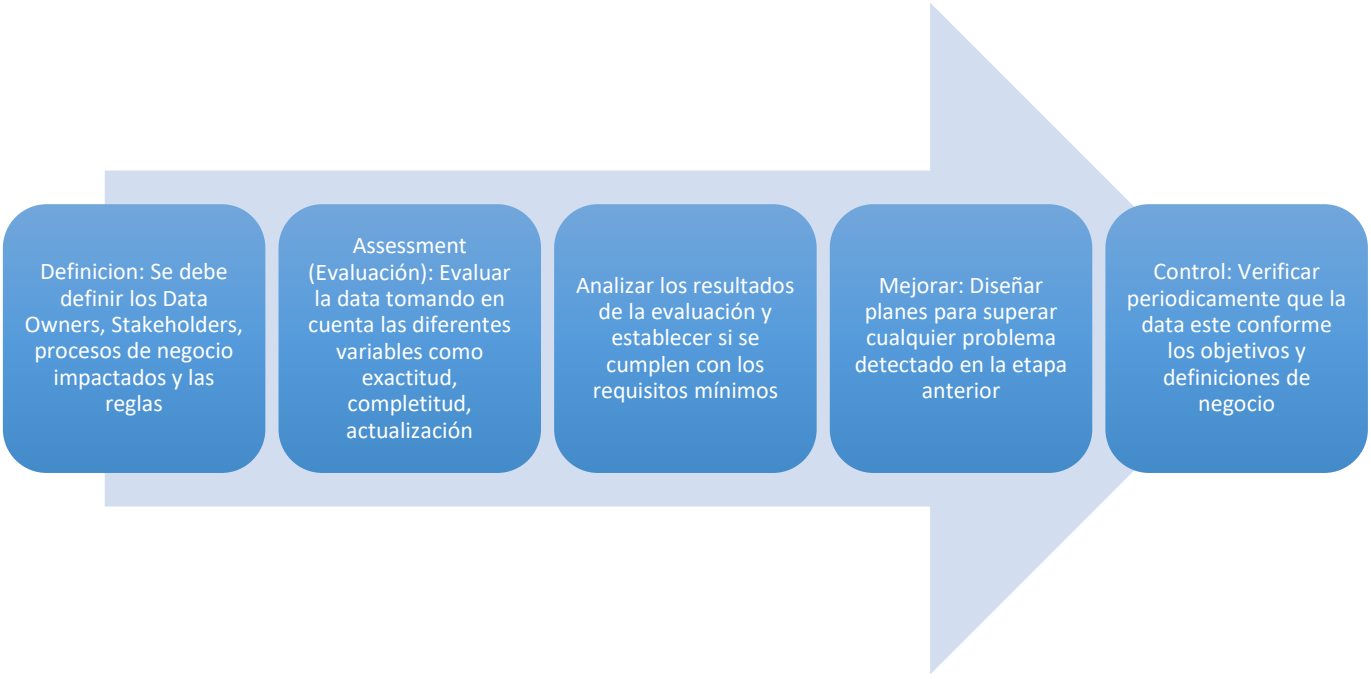


Figura 6-9 Pasos de DQ

Tipos de Errores

- Valores faltantes: un valor perdido corresponde a una ubicación en el dataset de la cual conocemos que tipo de dato esperamos, sin embargo su valor es desconocido. Por consiguiente, es imposible hacer un análisis estadístico si no tomamos una decisión sobre qué hacer con ese tipo de datos faltantes, las opciones son[24]:
    - Eliminar estas observaciones
    - Reemplazar con un valor por defecto
    - Reemplazar con el promedio, la moda, el mínimo o el máximo

  Esta decisión dependerá del tipo de variable, de su relevancia en el dataset, del contexto del análisis, entre otros elementos.

[23] http://www.dataversity.net/data-quality-simple-6-step-process/

[24] https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf

- Valores especiales: en las variables numéricas, un valor especial hace referencia a valores que no forman parte de los números reales R, en otras palabras hablamos de infinito positivo e infinito negativo.

- Valores outliers: Existe una cantidad de literatura enorme sobre este tema, la definición general de Barnett & Lewis nos dice que un valor outlier es una observación o grupo de observaciones que son inconsistentes con respecto a todo el grupo. Es importante mencionar que un outlier no es un error, pero es primordial detectarlos y dependiendo del contexto removerlos, o contemplarlos dentro del análisis.

  Una manera gráfica de determinar la existencia de dichos valores es mediante los gráficos Box-Plot

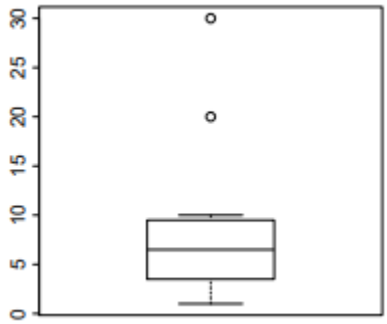


Figura 6-10 grafico Box-Plot

  Esta decisión dependerá del tipo de variable, de su relevancia en el dataset, del contexto del análisis, entre otros elementos.

- Valores con inconsistencias obvias: Una inconsistencia obvia ocurre cuando un registro contiene un valor o una combinación de valores que no pueden corresponder a una situación del mundo real. Por ejemplo la edad de una persona no puede ser negativa, un hombre no puede estar en estado de embarazo. Estos valores se tratan con reglas definidas junto a los stakeholders.

Bibliografía

- Big Data Analytics: Turning Big Data into Big Money
  by Frank J. Ohlhorst, November 2012

- Hadoop Essentials
  by Swizec Teller, April 2015

- Scalable Big Data Architecture: A Practitioner's Guide to Choosing Relevant Big Data Architecture
  by Bahaaldine Azarmi,2016

- Regression Analysis by Example, 4th Edition
  by Ali S. Hadi; Samprit Chatterjee,2006

- Basic Statistics for Trainers
  by Jean Houston Shore,2006

- A Framework for Analysis of Data Quality Research
  by Richard Y. Wang,1995

- An Introduction to Data Cleaning with R
  by Edwin de Jonge & Mark van der Loo,2013