



# Universidad de Santander

Personería Jur. 810 de 12/03/96 Min.Educación

**UDES**

VIGILADA MINEDUCACIÓN



# Tópicos administrativos I

## Inteligencia de Negocios

### *y minería de Datos*

### *(usando Power Bi, R, Python y Knime)*

Msc- MST - Especialista en sistemas de información- [ALFREDO A. DIAZ CLARO](#)



# **Horas de Reunión Programadas**



| Unidad de aprendizaje                                                      | Horario                                                            |
|----------------------------------------------------------------------------|--------------------------------------------------------------------|
| MÓDULO I. Machine Learning y la inteligencia de negocios.                  | <ul style="list-style-type: none"><li>• 8 a.m. a 10 a.m.</li></ul> |
| MÓDULO II. Estadística Descriptiva e Inteligencia de Negocios con Power Bi | <ul style="list-style-type: none"><li>• 10 a 12 m.</li></ul>       |
| MÓDULO III. Introducción al machine Learning y algoritmo supervisado       | <ul style="list-style-type: none"><li>• 2 pm. A 4 p.m.</li></ul>   |
| MÓDULO IV. Introducción a algoritmos no supervisados                       | <ul style="list-style-type: none"><li>• 4 p.m. a 6 p.m.</li></ul>  |



**Universidad  
de Santander**  
Personería Jur. 810 de 12/03/96 Min.Educación  
VIGILADA MINEDUCACIÓN  
**UDES**

**LA CALIDAD  
NOS UNE**

# Análisis de transacciones

Fuente: Big Data. La revolución de los datos masivos. Pág. 77

## El poder de los datos



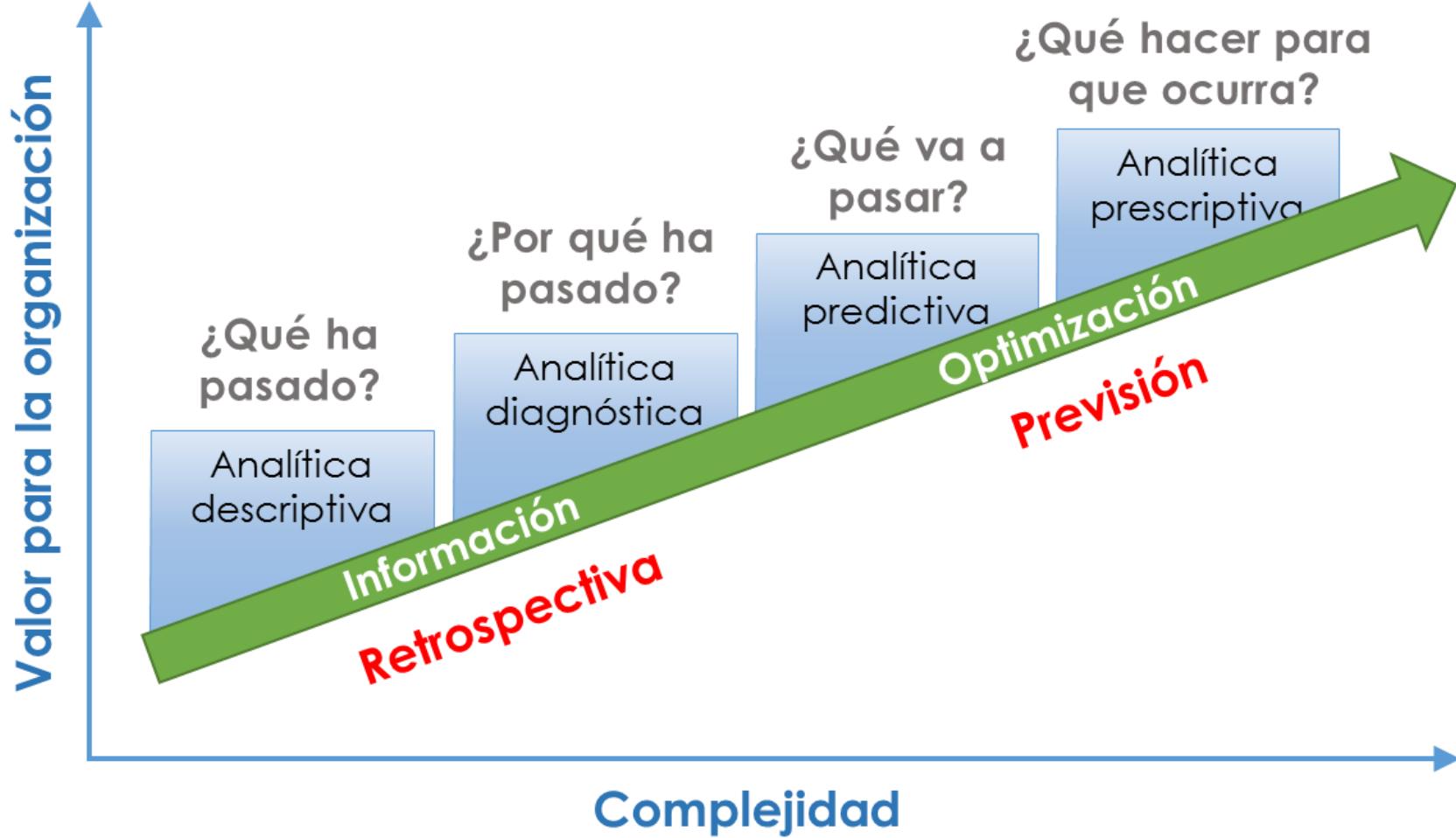
Y descubriremos las diferencias encontraremos diferencias entre B.I., Big Data, Machine Learning, minería de datos, feature engineering , estadística, Ciencia de datos, Data Lake, Datawarehouse, Data Scraping, data wrangling (similar a ETL)...



Descriptive Analytics  
Traditional Business Intelligence  
Group Tendencies  
Dashboards



Predictive Analytics  
Machine Learning  
Individual Predictions  
Predictive Applications



# Ejemplo de BI en linea

Haciendo click en la imagen ingresa a los demos de BI.

The screenshot shows a web browser displaying the Stratebi website at [www.stratebi.com/demos](http://www.stratebi.com/demos). The page features a navigation bar with links like Nosotros, Casos de éxito, Trabajo, Contacto, TodoBI blog, Documentos, Enlaces de interés, and Acceder. Below the navigation is a main menu with categories: INICIO, SOLUCIONES DE NEGOCIO, TECNOLOGÍAS, CLIENTES, PORTFOLIO, DEMOS, FORMACIÓN, BI OPEN SOURCE, and a search bar.

Below the menu, there are six examples of BI dashboards:

- TABLERO FUTBOLERO (LIGA ESPAÑOLA)**: A dashboard showing the ranking of football teams in the Spanish League. It includes a chart titled "Ranking de equipos" and a map of Madrid with team locations.
- SMARTCITY MADRID**: A dashboard for Smartcity Madrid, featuring a map of the city with various data points and icons.
- BIG DATA**: A dashboard illustrating Big Data use cases, specifically focusing on Social Networks. It shows "Real Time" data processing and social media icons.
- REAL TIME DASHBOARD**: A dashboard titled "STRATEBI - BIG DATA" showing real-time data from a map of Spain.
- E-commerce Análisis**: A dashboard for E-commerce analysis, displaying consumption reports, maps, and charts.
- facebook DASHBOARD COMENTARIOS**: A dashboard for Facebook comments, showing statistics and geographical distribution.

The bottom of the screen shows the Windows taskbar with various application icons.

# Ejemplo en Power Bi

Retail Analysis Sample PBIX - Power BI Desktop

Archivo Inicio Vista Modelado Ayuda Herramientas de objetos visuales Formato Datos y detalles

Cortar Copiar Copiar formato Pegar Obtener datos Recientes Especificar datos Editar consultas Actualizar Nueva página Nuevo objeto visual Botones pregunta Cuadro de texto Imagen Formas Desde Marketplace De archivo Cambiar tema Temas Administrar relaciones Nueva medida Nueva columna Nueva medida rápida Cálculos Compartir

Store Sales Overview

This Year Sales por Chain

10 New Stores

104 Total Stores

Total Sales Variance by FiscalMonth and District Manager

DM Allan Guinot, Andrew Ma, Annelie Zubair, Brad Sutton, Carlos Grilo, Chris Gray, Chris McGurk, Tina Lassila, Valery Ushakov

This Year Sales por PostalCode y Store Type

Store Type: New Store (Blue), Game Store (Red)

Total Sales Variance %, Sales Per Sq Ft y This Year Sales por District y District

PostalCode: LI - 01, LI - 02, LI - 03, LI - 04, LI - 05, FD - 01, FD - 02, FD - 03, FD - 04

Legend: Store Type (New Store, Game Store)

Ubicación: PostalCode

Ubicación: Store Type

Ubicación: Latitud

Ubicación: Longitud

Tamaño: This Year Sales

Visualizaciones > Campos >

Buscar

Sales

District

BusinessUnitID

District

DistrictID

DM

DM\_Pic

DM\_Pic\_fl

Item

Store

Time

VIGILADA MINEDUCACIÓN

unab

POSGRADOS

ESTADÍSTICAS

Información sobre herramientas...

Info Overview District Monthly Sales New Stores +

# Ejemplo en Knime

KNIME Analytics Platform

File Edit View Node Help

100% Quick Access

KNIME Explorer X LOCAL (Local Workspace) EXAMPLES (knime@hub.knime.com) Caoba curso Ejemplos Alfredo

Workflow Coach X Recommended Nodes

Node Repository X IO Manipulation Views Analytics DB Other Data Types Structured Data Scripting Tools & Services KNIME Labs Workflow Control Workflow Abstraction Reporting

0: Predictor Abandono seguros X Welcome to KNIME Analytics Platform

Entrenamiento de un predictor de baja en contratos de seguro  
Este ejemplo se construye sobre PMML para un modelo de fidelización usando Decision Tree algorithm.

Interactive Histogram (local) Statistics Box Plot Scatter Plot (local)

Excel Reader (XLS) Datos del contrato Joiner Node 99 Node 100 Node 101 Node 102 Dec Tree I

File Reader Integraremos las tablas Number To String (deprecated) Cambiamos las variables numéricas de 1/0 a string Color Manager Marcar abandono Partitioning 80% vs. 20% Equal Size class

Description X Predictor A... Title Training a Churn Predictor Description This workflow is an example of how to build churn prediction using a Decision Tree algor Tags No tags have been added yet. Links Churn Prediction (Website) Creation Date 2015-4-23 Author knime KNIME Hub Search X

Outline X Console X KNIME Console Log file is located at: C:\Users\Adiaz\knime-workspace\.metadata\knime\knime.log WARN File Reader 0:91 The file 'file:/D:/Users/adiaz/Documents/Machine%20Learnig%20seguros/contratos.xls' does not exist. WARN Excel Reader (XLS) 0:98 The file 'D:/Users/adiaz/Documents/Machine Learnig seguros/contratos.xls' does not exist. WARN Excel Reader (XLS) 0:107 The file 'D:/Users/adiaz/Documents/Machine Learnig seguros/datos.xls' does not exist.

```
graph LR; FR[File Reader] --> J[Joiner]; J --> DS[Datos del contrato]; DS --> N99[Node 99]; N99 --> N100[Node 100]; N100 --> N101[Node 101]; N101 --> N102[Node 102]; N102 --> DTI[Decision Tree I]; N99 --> NTS[Number To String]; NTS --> CM[Color Manager]; CM --> P[Partitioning]; P --> ES[Equal Size]; ES --> CLS[class]; NTS --> CV[Cambiamos las variables numéricas de 1/0 a string]; CV --> M[Marcar abandono]; M --> P;
```

# Ejemplo en Python

C:\Users\Adiaz\Documents\adiaz\Documents\Analisis multivariado\Especializacion Analitica\ejemplos en Python\histogramas.py - Sublime Text (UNREGISTERED)

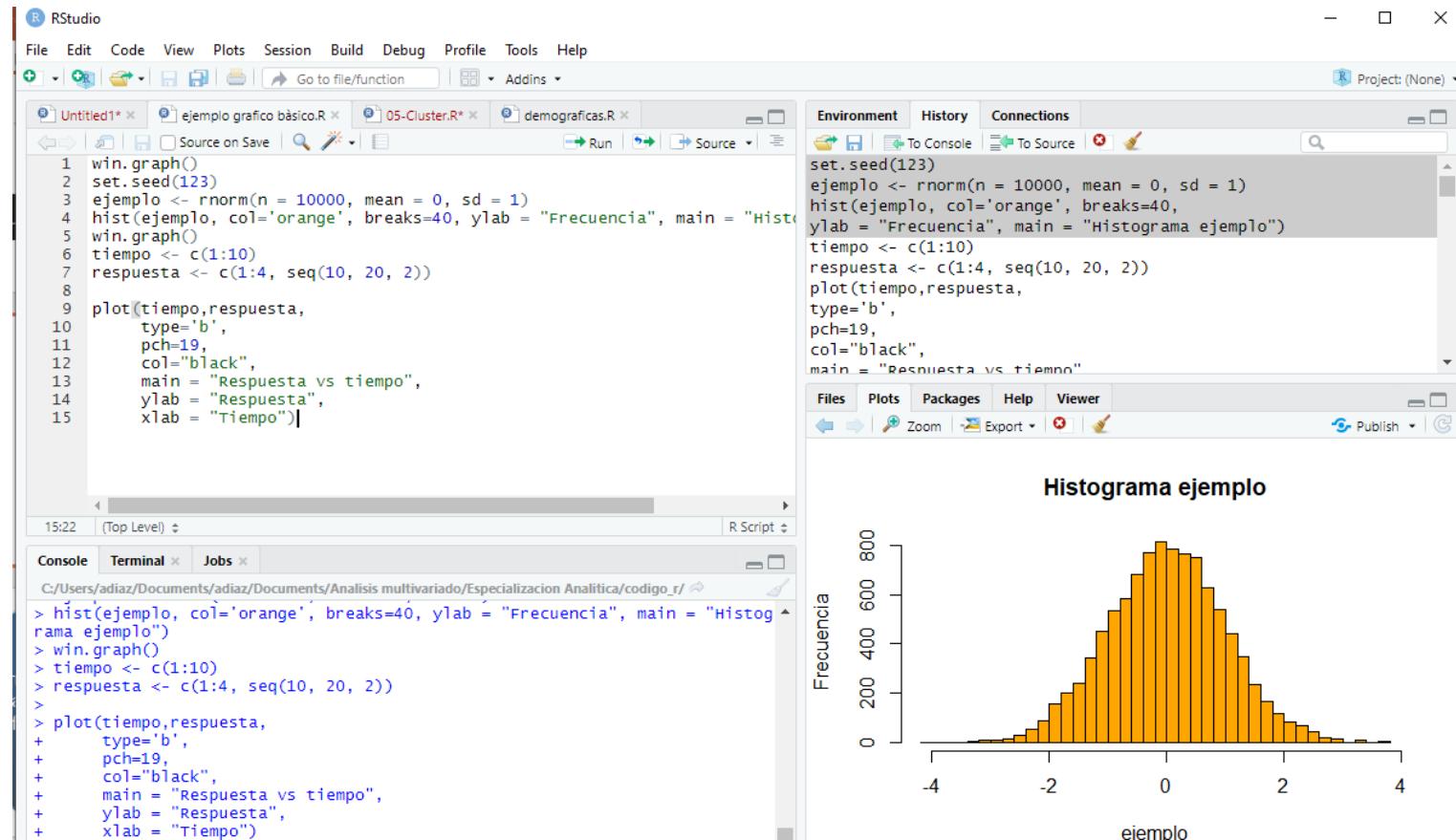
```
File Edit Selection Find View Goto Tools Project Preferences Help
```

labels.py    entry.py    textoboton.py    calculadora1.py    cuestionario.csv    estadisticaPython.py    \*REPL\* [python]    histogramas.py

Figure 1

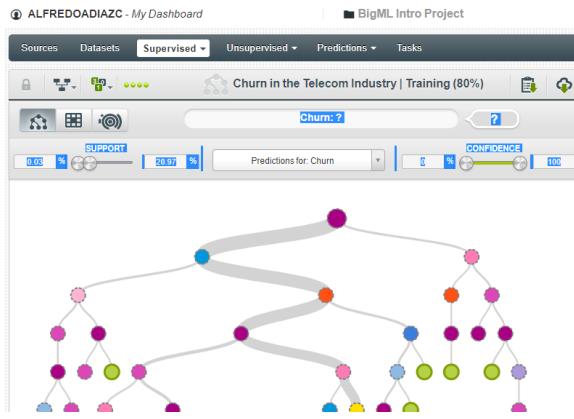
A histogram showing a normal distribution. The x-axis ranges from -0.3 to 0.3 with major ticks every 0.1. The y-axis ranges from 0 to 4 with major ticks every 1. The histogram consists of blue bars. A smooth red curve is overlaid on the histogram, peaking at approximately x=0 with a value of about 4.2. The distribution is symmetric and centered around zero.

```
1 import matplotlib.pyplot as plt #.importando matplotlib
2 import seaborn as sns #.importando seaborn
3 import numpy as np #.importando numpy
4 #.parametros estéticos de seaborn
5 sns.set_palette("deep", desat=.6)
6 sns.set_context(rc={"figure.figsize": (8, 6)})
7 mu, sigma = 0, 0.1 #.media y desvio estandar
8 s = np.random.normal(mu, sigma, 1000) #.generando datos aleatorios
9 #.histograma de distribución normal.
10 cuenta, cajas, ignorar = plt.hist(s, 30, density=True)
11 normal = plt.plot(cajas, 1/(sigma * np.sqrt(2 * np.pi)) * np.exp(-0.5 * ((cajas - mu)**2 / (sigma**2))), linewidth=2, color='r')
12 .....np.exp(-0.5 * ((cajas - mu)**2 / (sigma**2)))
13 .....linewidth=2, color='r')
14 plt.show()
```



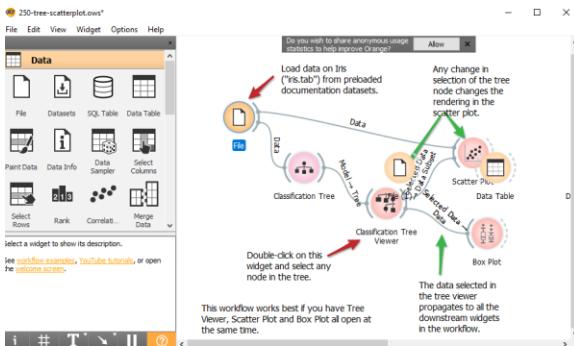
[https://hub.knime.com/knime/space/Examples/04\\_Analytics/02\\_Statistics/01\\_Simple\\_Example\\_with\\_Statistics](https://hub.knime.com/knime/space/Examples/04_Analytics/02_Statistics/01_Simple_Example_with_Statistics)





# Otras herramientas

- BigML
- ORANGE





# Qué es Big Data'



# The Internet in Real Time

[Tweet](#) [Share](#) 586 [Guardar](#) [G+](#)

By the time you finish reading this sentence, there will have been 219,000 new Facebook posts, 22,800 new tweets, 7,000 apps downloaded, and about \$9,000 worth of items sold on Amazon... depending on your reading speed, of course. Now that the Internet is widely available, just one second of global online activity is jam-packed full of events, from communication with others to data storage to entertainment options galore.

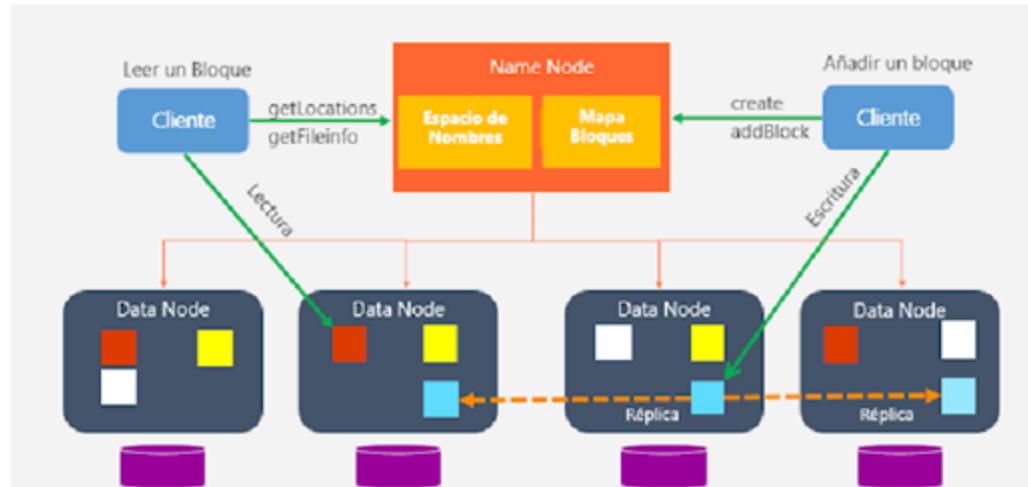
For example, in the amount of time you've been on this page, this is how much data has already passed through the Internet.

249,600

Activar Windows  
Ve a Configuración para activar Windows.

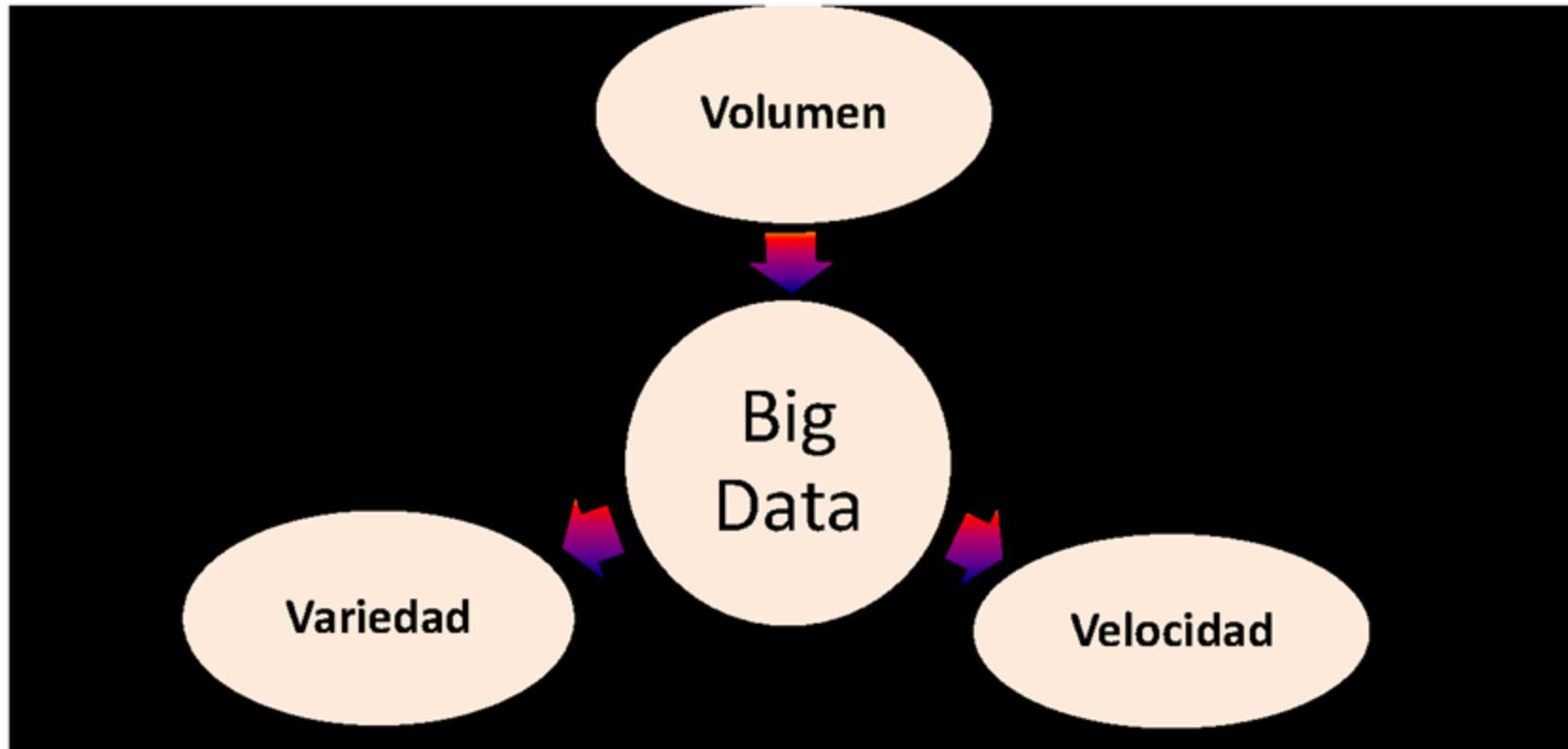


# Big Data



“Big Data” son datos cuyo volumen, diversidad y complejidad **requieren nueva arquitectura, técnicas, algoritmos y análisis** para gestionar y extraer valor y conocimiento oculto en ellos.

**Big Data:** Volumen, Velocidad, Variedad,  
Veracidad y Valor



# Big Data



- Hace referencia a grandes conjuntos de datos que por la velocidad a la que se generan; las múltiples fuentes y formatos; y la capacidad de procesos.
- Problema: Escalabilidad de grandes cantidades de datos
  - Exploración 50 TB en 1 nodo @ 50 MB/sec = 12 días
- Solución → Divide-Y-Vencerás
  - Exploración en un clúster de 1000 nodos = 16 minutos



HUE

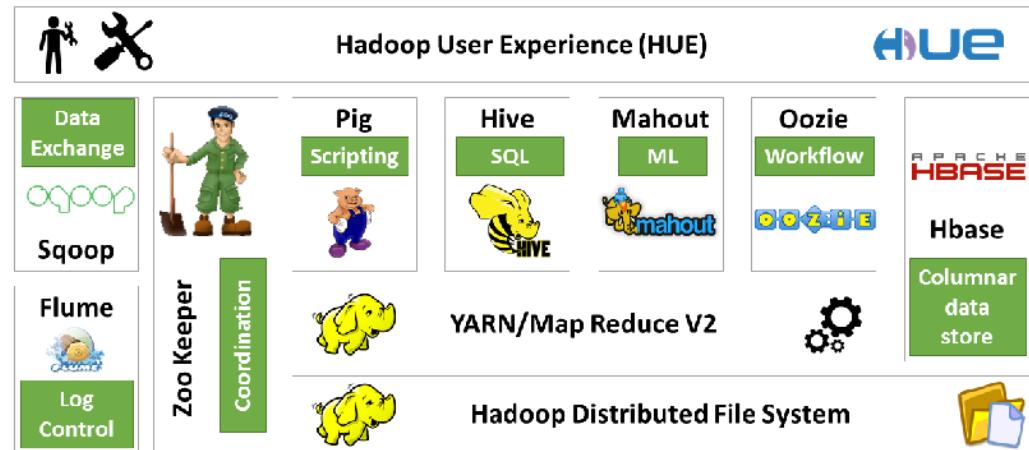
APACHE  
HBASE

Hbase

Columnar  
data  
store

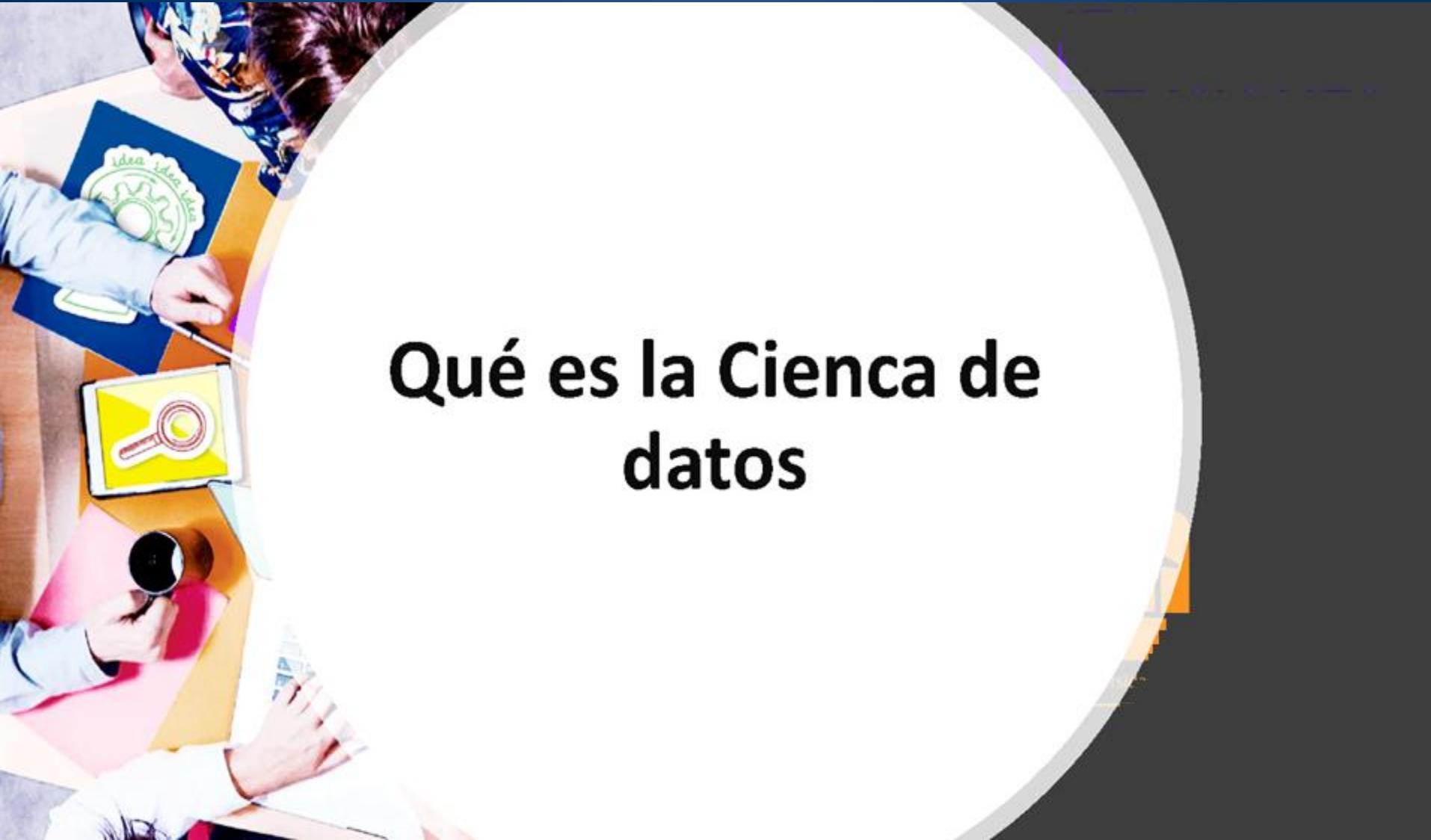


## The Apache Hadoop Stack



# Herramientas de BigData

# Qué es la Cienca de datos



# Ciencia de Datos

- Ámbito del conocimiento que engloba las habilidades asociadas al análisis inteligente de datos, incluyendo Big Data



# Analítica de datos

- **Analítica:** Cualquier proceso orientado a datos que provee hallazgos (*insight*)
- **Analítica de negocios (*Business analytics*)**

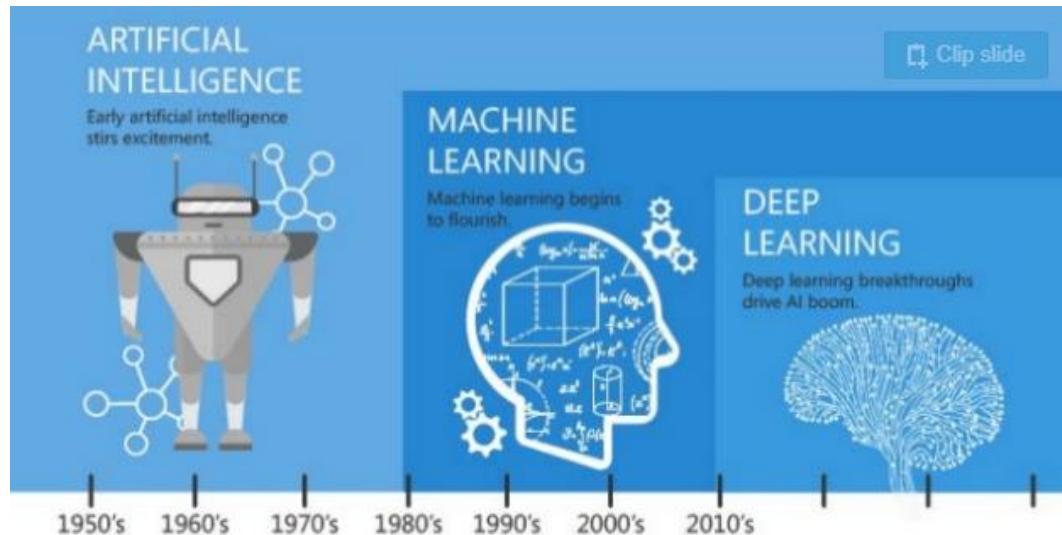
Aprovecha todas las formas de analítica para lograr los objetivos del negocio.

Condiciones:

1. Relevante para el negocio
2. *Insight* accionable
3. Medición del desempeño y del valor



# Ciencia de datos la Inteligencia Artificial, Machine Learning y Deep Learning



- AI components:
  - Computer Vision
  - Language Processing
  - Creativity
  - Etc;
- Machine learning:
  - Classification
  - Clustering
  - Neural Network
  - Etc.
- Revolution in ML:
  - Deep learning



A large, semi-transparent white circle covers the right side of the slide, containing the main title text. The background features a collage of images related to science and education, including a person holding a globe, a hand holding a tablet with a magnifying glass icon, and a person wearing a lab coat and safety glasses looking through a telescope.

# Herramientas para Data Science

# Herramientas

## Lenguajes de Programación



python



Clojure

pandas



hadoop

## Visualización

D3js



mongoDB

Microsoft  
SQL Server 2012

## Machine Learning



orange  
g  
CleverData  
BigData Prediction



RAPID MINER  
KNIME

## Procesamiento de Datos



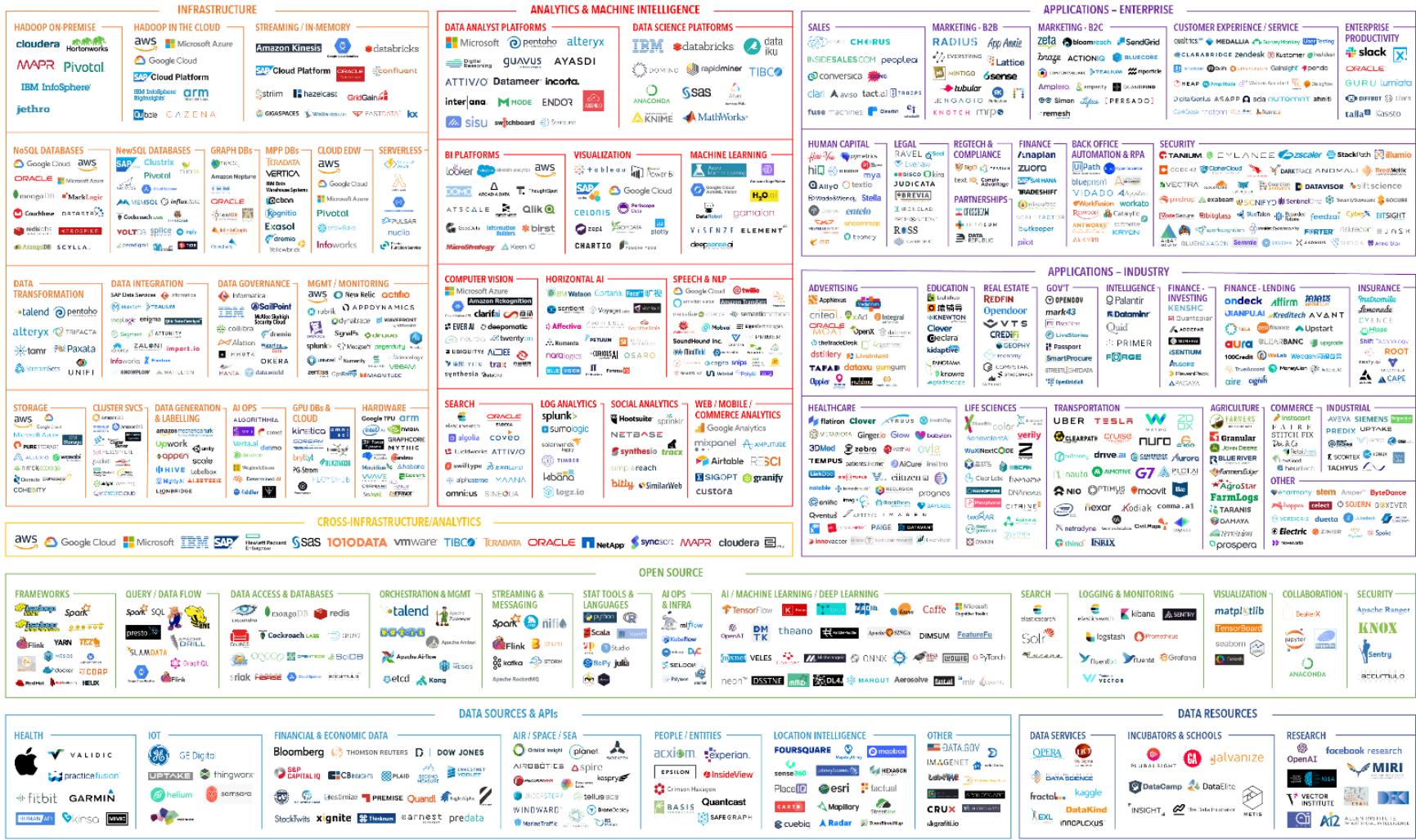
Cassandra

JHepWork



Neo4j

DATA & AI LANDSCAPE 2019



July 16, 2019 - FINAL 2019 VERSION

© Matt Turck (@mattturck), Lisa Xu (@lisaxu92), & FirstMark (@firstmarkcap)

mattturck.com/data2019



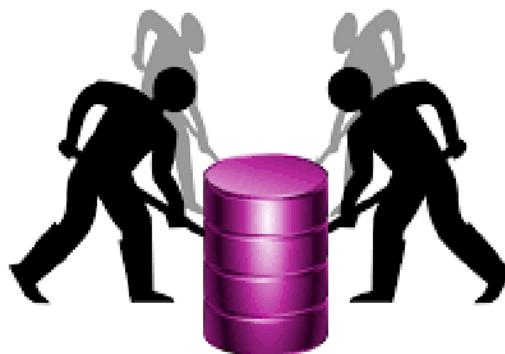


## *La minería de datos*



# *Minería de datos*

- **¿Qué es minería de datos?**
- Minería de datos es un medio para desarrollar inteligencia de negocios a partir de los datos que una organización recoge, organiza y almacena. En esencia, es una herramienta de analítica de negocios.
- La minería de datos es el proceso de búsqueda de información valiosa, y su posterior análisis, en una gran base de datos, *data warehouse*, *data mart* u otras fuentes voluminosas de información (textos y Web), que se utilizará en la gestión de la toma de decisiones.
- La minería de datos, en esencia, puede realizar dos operaciones:
  - 1) predecir tendencias y comportamientos;
  - 2) identificación de patrones desconocidos con anterioridad.



# **Herramientas de software de minería de datos**

Existen numerosos fabricantes de software que construyen herramientas de minería de datos tanto de software propietario como de código abierto.

## **Software propietario**

- SPSS. Clementime
- SAS (Enterprise Miner)
- IBM (Intelligent Miner)
- Megaputer (PolyAnalyst)
- Microsoft SQL Server Data Mining
- Teradata Warehouse Miner
- Oracle Data Mining
- Statsoft. Statistica, prueba gratuita 30 días
- Knime
- Rapidminer

# **G.I.G.O.**



**"GARBAGE IN, GARBAGE OUT"**



# *Machine Learning*

# Clasificación de Machine Learning

## Aprendizaje Supervisado

- Modelos Predictivos.
- La máquina aprende explícitamente.
- Predice el futuro a partir de datos históricos.
- Resuelve problemas de clasificación y regresión.

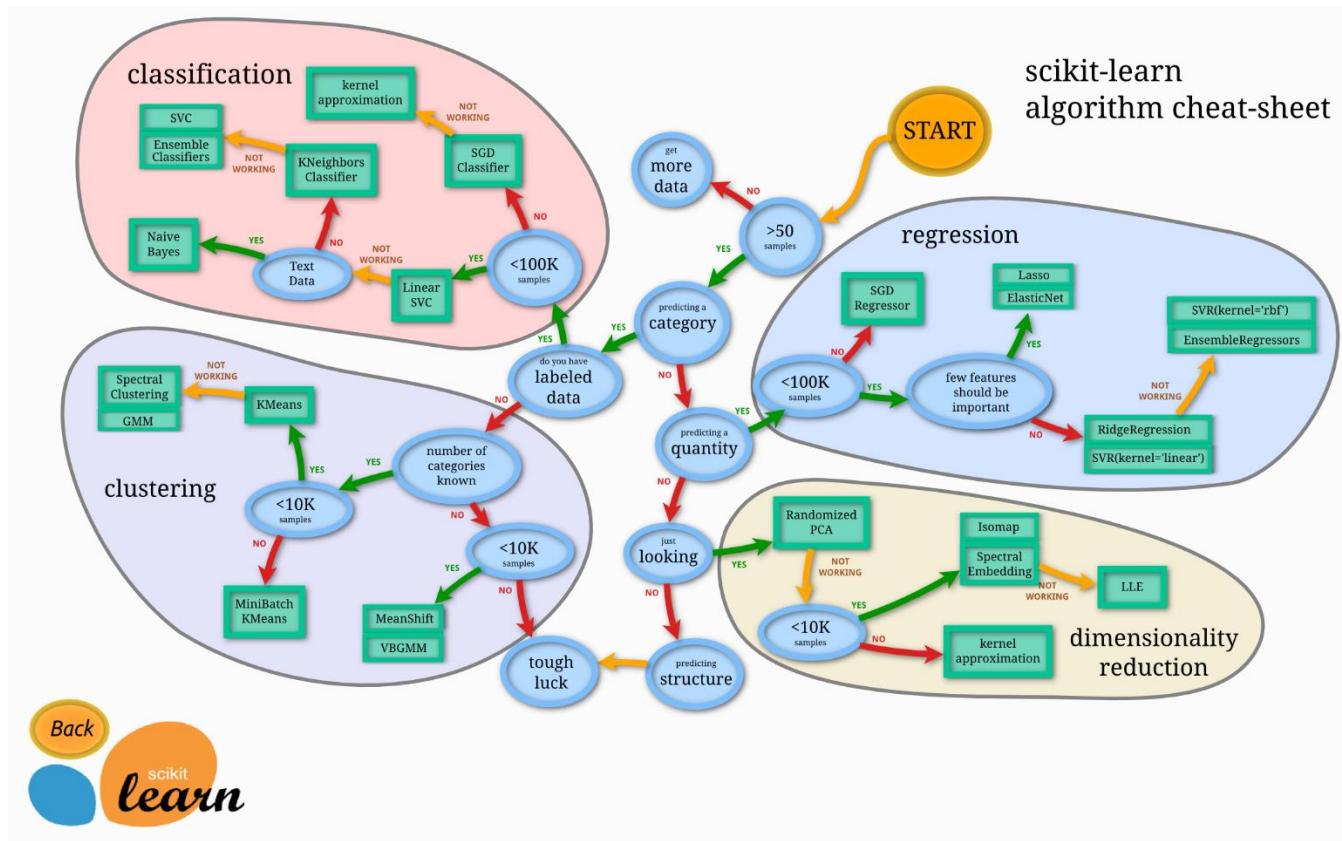
## Aprendizaje No Supervisado

- Modelos Descriptivos.
- La máquina entiende los datos.
- La evaluación es cualitativa o indirecta.
- No realiza predicciones, encuentra algo específico.

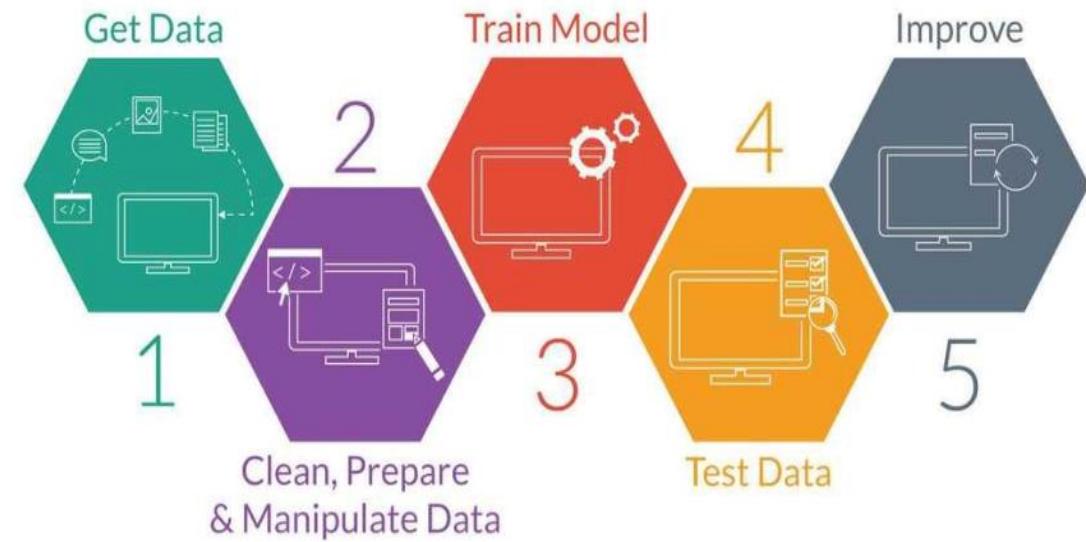
## Aprendizaje Reforzado

- Un enfoque de la IA
- Aprendizaje basado en los hallazgos.
- La máquina aprende a como actuar en un determinado entorno.
- Maximiza los hallazgos.

# Seleccionar el estimador correcto



# Flujo de Machine learning



# Conjunto de datos (dataset)

- Es la materia prima del sistema de predicción. Es el histórico de datos que se usa para entrenar al sistema que detecta los patrones. El conjunto de datos se compone de instancias, y las instancias de factores, características o propiedades.
- <https://archive.ics.uci.edu/ml/datasets.php>

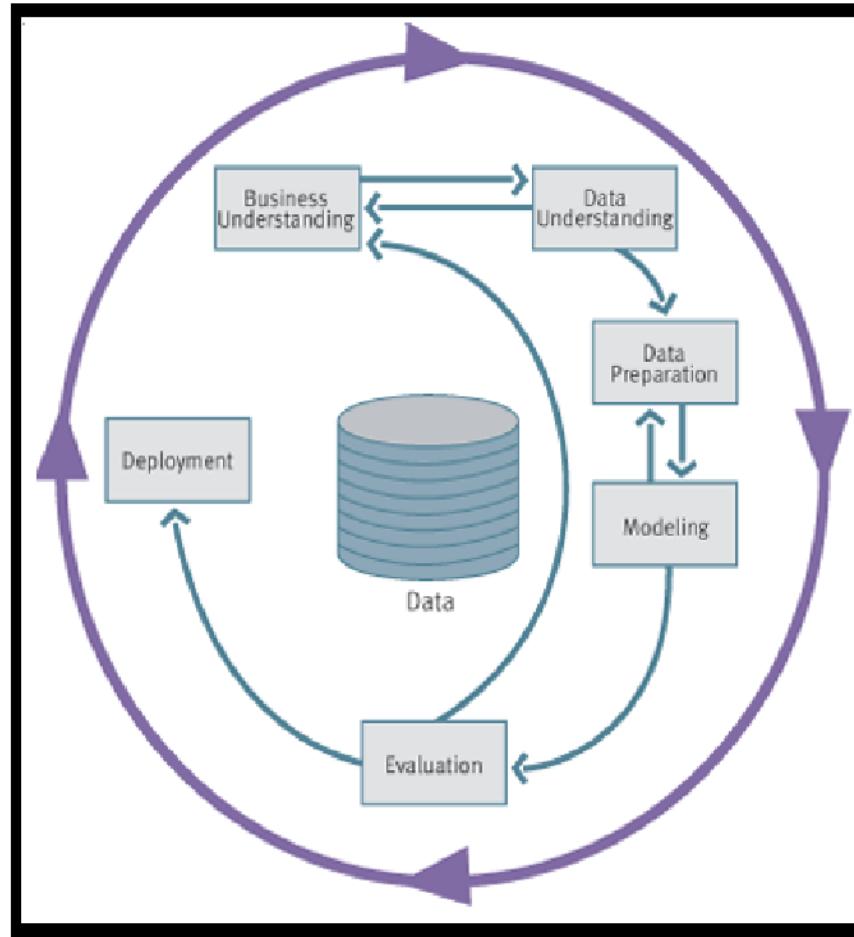


# *Metodologías de Ciencias de datos De Crisp a Asum*

# CRISP -DM

- Para proyectos de **implementación de minería de datos, CRISP-DM. Cross Industry Standard Process for Data Mining**

- .1 Comprensión del negocio
2. Comprensión de los datos
3. Preparación de los datos
4. Modelado (construcción de modelo)
5. Evaluación
6. Despliegue (desarrollo)

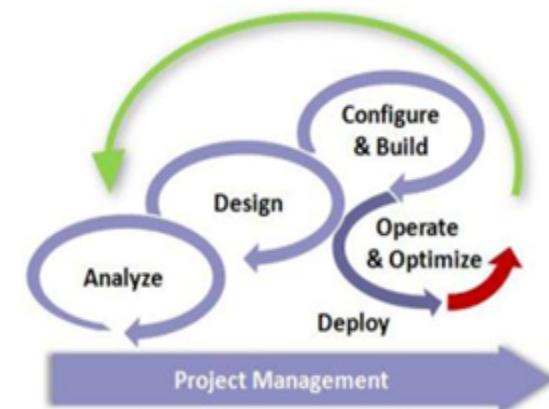


ORGANIZACIÓN UNIÓN

# Proceso de minería de datos (de CRISP-DM a Asum)

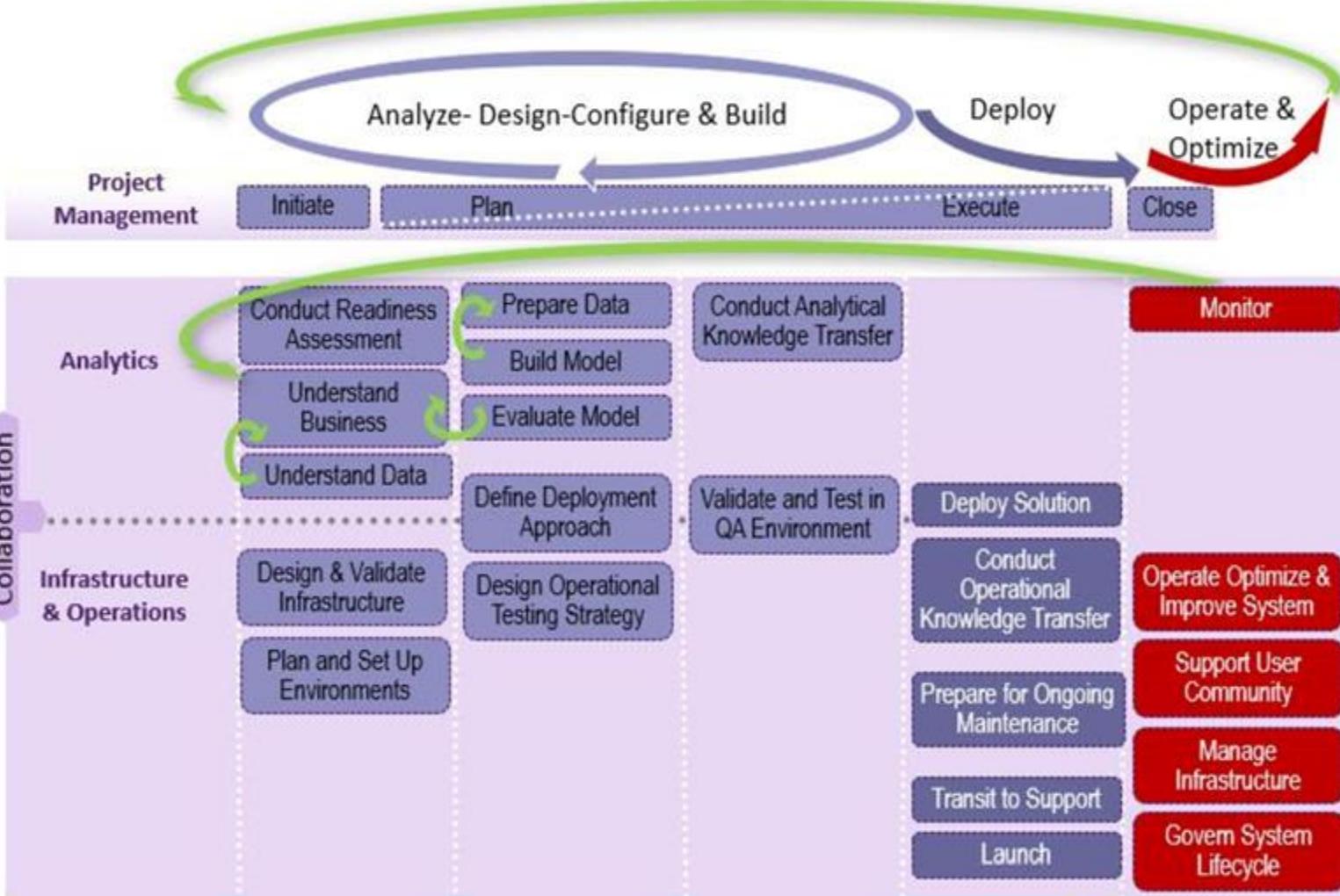
En 2015, IBM Corporation, uno de los impulsores tradicionales de CRISP-DM, planteó una nueva metodología methodology llamada [Analytics Solutions Unified Method for Data Mining/Predictive Analytics](#) (ASUM-DM) que extiende CRISP-DM, y es parte de la metodología general ASUM (Analytics Solutions Unified Method) incorporada en los productos y soluciones analíticas de IBM. En un tiempo se verá si esta evolución tiene éxito en la comunidad de Data Science.

ASUM- Analytics Solutions Unified Method



[http://gforge.icesi.edu.co/ASUM-DM\\_External/index.htm#cognos.external.asum-DM\\_Teaser/deliveryprocesses/Analyze-Design-Configure%20Build\\_4E20B18D.html?proc=0eKIHlt6EeW\\_y7k3h2HTng&path=0eKIHlt6EeW\\_y7k3h2HTng,0eHEyVt6EeW\\_y7k3h2HTng](http://gforge.icesi.edu.co/ASUM-DM_External/index.htm#cognos.external.asum-DM_Teaser/deliveryprocesses/Analyze-Design-Configure%20Build_4E20B18D.html?proc=0eKIHlt6EeW_y7k3h2HTng&path=0eKIHlt6EeW_y7k3h2HTng,0eHEyVt6EeW_y7k3h2HTng)

Figure 1: Analytics Solutions Unified Method (ASUM).  
Source: IBM Corporation.





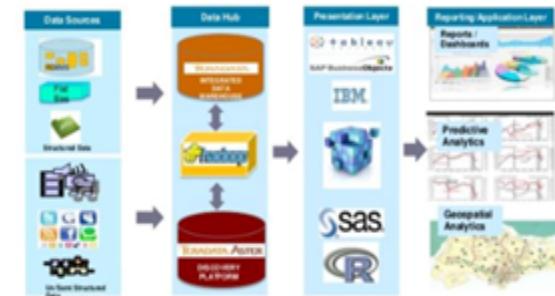
## *Inteligencia de Negocios*

# ¿Qué es *inteligencia de negocios*?

Es un proceso interactivo para explorar y analizar información estructurada sobre un área, normalmente almacenada en un *data warehouse*, para descubrir tendencias o patrones, a partir de los cuales derivar ideas y extraer conclusiones, incluye la comunicación de los descubrimientos y la realización de los cambios.

Reúne datos, tecnología, análisis y conocimiento humano para optimizar decisiones de negocio y conseguir el éxito de una empresa.

Componentes *data warehouse*- *data mart*, herramientas de analítica, *data mining* y herramientas de *reporting*, *quering*, *dashboard*, *scorecards* y herramientas de visualización y de descubrimiento.



# Tableros de control

Un cuadro de mando es un tipo de informe que proporciona una instantánea de alto nivel sobre la situación de una empresa u organización. Los cuadros de mandos proporcionan una representación visual del rendimiento que puede servir de resumen del progreso positivo o negativo que pueda tener una organización.

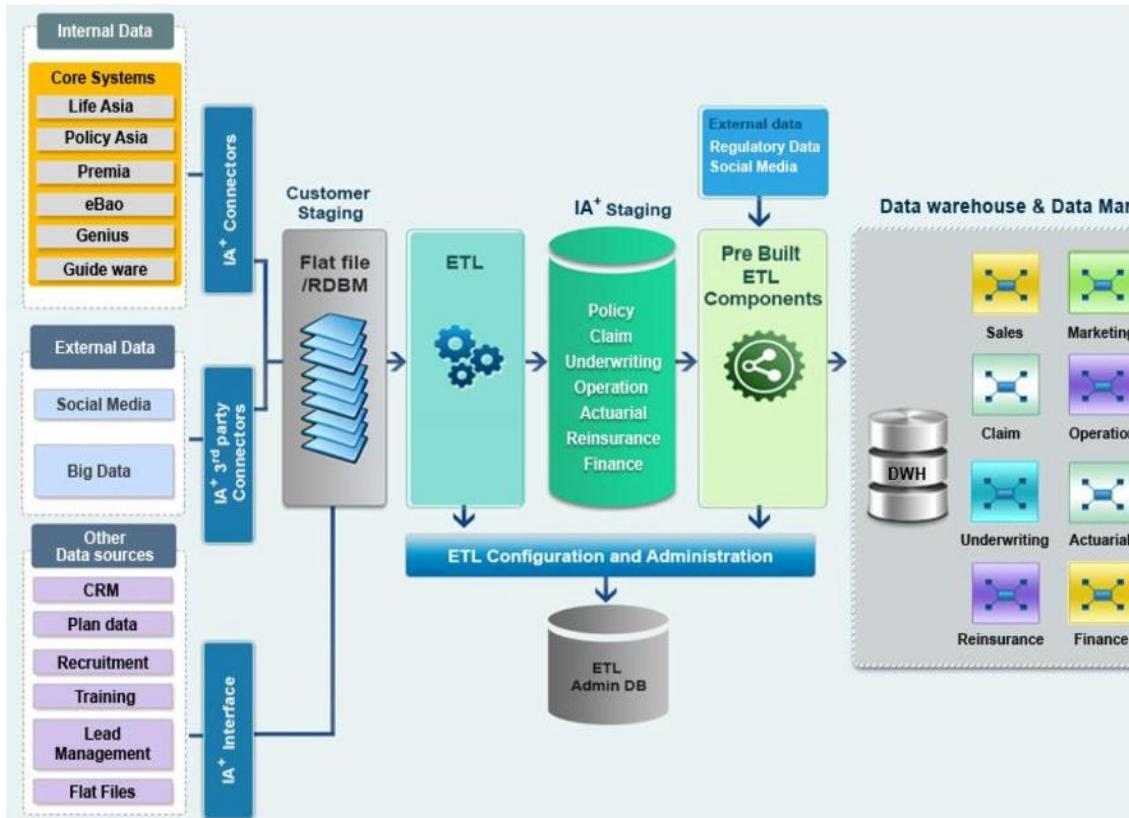


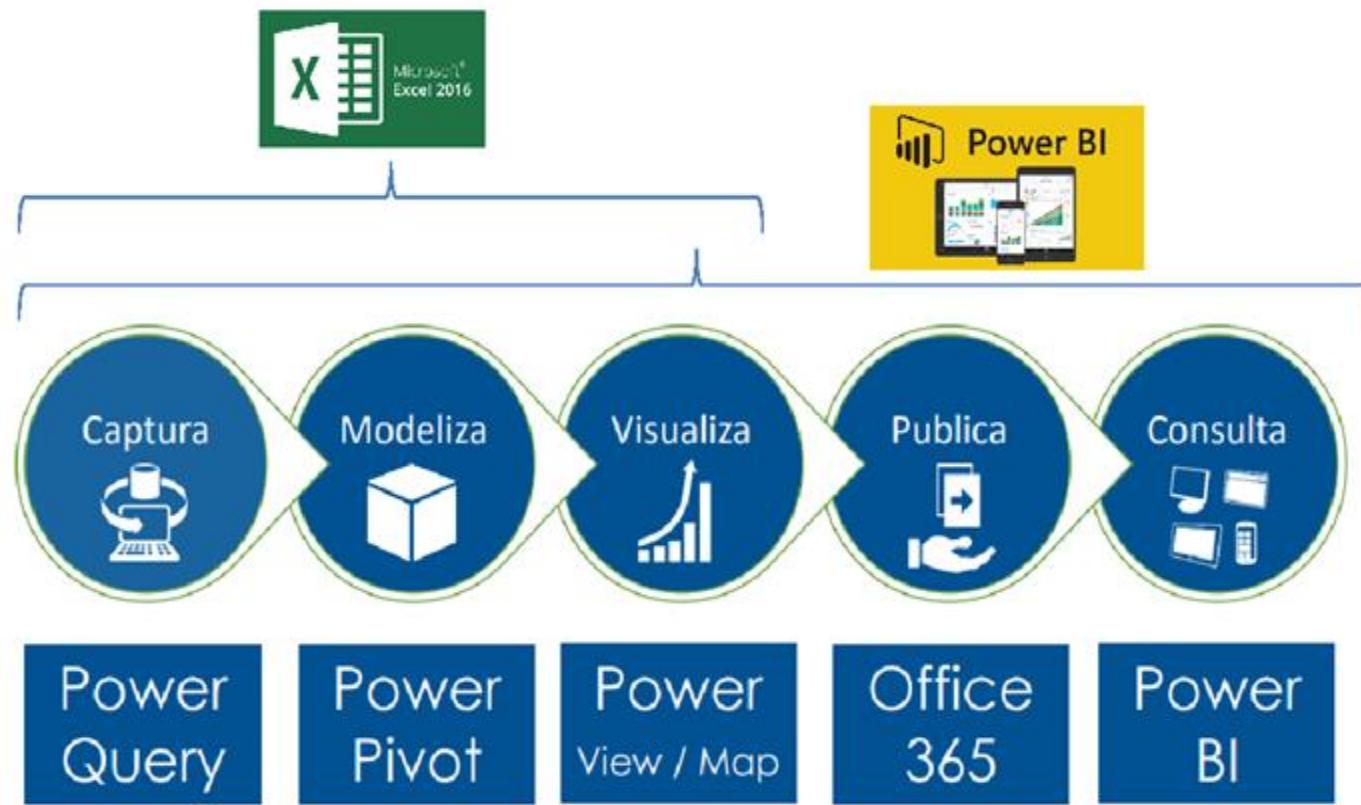
# *Componentes de arquitectura de un BI*

|                          | BI                                                                                                                                                                                 | Analítica de negocios                                                                                                                             |
|--------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|
| Responde a las preguntas | ¿Qué sucede?<br>¿Cuándo?<br>¿Quién?<br>¿Cuántos?                                                                                                                                   | ¿Por qué sucedió?<br>¿Ocurrió otra vez?<br>¿Qué pasaría si cambiáramos x?<br>¿Qué otras cosas dicen los datos que nunca se nos ocurrió preguntar? |
| Infraestructura          | Reporting (KPI, métrica) o interfaces de usuarios<br>Sistemas de monitoreo/alertas<br>Cuadros de mando (dashboard)<br>Balanced scorecards<br>OLAP<br>Consultas <i>ad hoc/query</i> | Análisis: estadístico-cuantitativo<br><i>Data mining</i><br>Modelamiento productivo<br>Análisis multivariado                                      |



# Arquitectura de procesos de BI





## *Microsoft Power BI*

- Microsoft Power BI es una evolución de las herramientas analíticas de Excel, con nuevas funciones y servicios que permiten visualizar datos, compartir información y colaborar de nuevas formas más sencillas e intuitivas.

**Uso de EXCEL**  
*1 Billón de usuarios de Office*



Descubrir      Analizar      Visualizar

**Colaboración con Office 365**  
*Entorno colaborativo en la nube*



+

Compartir      Buscar      Movilidad



# Universidad de Santander

**UDES**

Personería Jur. 810 de 12/03/96 Min.Educación

VIGILADA MINEDUCACIÓN

**LA CALIDAD  
NOS UNE**