

# Tutorial Big Data CLEI 2019



## Ciudad de Panamá, Panamá

Por: Edwin Montoya  
[emontoya@eafit.edu.co](mailto:emontoya@eafit.edu.co)

---

Inspira Crea Transforma

UNIVERSIDAD  
**EAFIT**<sup>®</sup>

# Edwin Montoya



Ingeniero de Sistemas

Doctor Ingeniero en Telecomunicación

Jefe Depto Sistemas

Profesor: Big Data, Cloud Computing, Text Mining

Investigador: Big Data, E-learning con Big Data e IA

Comité de Maestría en Ciencia de Datos y Analítica



\* Participó en la formulación y consolidación del Centro de Excelencia en Big Data y Analytics

\* Comité técnico y Junta Directiva

\* Investigador Adjunto

# Experiencia en Big Data

- Varios cursos formales en Pregrado y Maestría en Big Data
- Varios cursos en Educación continua en Tecnologías Big Data.
- 7 cursos de tecnologías Big Data a nivel nacional dentro del programa de Ciudadano de Datos con Caoba, 4 ciudades en Colombia.
- 5 proyectos de maestría recientes relacionados con Big Data
- Asesorías y Consultorías: Modelo de Referencia para Big Data a Terpel (Distribuidor de Petróleo)
- Proyectos de Investigación en Caoba

## Laboratorios

- <https://github.com/edwinmontoya/tutorialbigdatalei2019.git>
- Se tendrá acceso a una serie de Recursos en Nube:
  - AWS EMR:
    - <http://emr1.emontoya.ml:8888> (hue)
    - <http://emr2.emontoya.ml:8888> (si necesita)
    - <http://emr3.emontoya.ml:8888> (si necesita)
      - User: admin Password: Clei2019\*
  - <http://emr1.emontoya.ml:8890> (zeppelin)
  - <http://emr2.emontoya.ml:8890> (si necesita)
  - <http://emr3.emontoya.ml:8890> (si necesita)
    - Sin autenticación

## Recursos Gratuitos para Big Data

- **Cuenta FreeTier o AWS Educate en Amazon.**
  - Servicios RDS/Mysql (base de datos relacional)
  - Servicio EMR (Elastic MapReduce)
  - Servicio S3 (Almacenamiento por objetos)
- **Microsoft Azure FreeTier o Azure for Students:**
  - <https://azure.microsoft.com/en-us/free/students/>
  - <https://visualstudio.microsoft.com/dev-essentials/>
- **Databricks Community**
  - Cluster gratis restringido para Spark + Jupyter Notebooks
- **Máquinas virtuales Sandbox de:**
  - Hortonworks:
    - <https://www.cloudera.com/downloads/hortonworks-sandbox.html>
  - Cloudera:
    - [https://www.cloudera.com/downloads/quickstart\\_vms/5-13.html](https://www.cloudera.com/downloads/quickstart_vms/5-13.html)

# Algunas iniciativas en Colombia

Inspira Crea Transforma

UNIVERSIDAD  
**EAFIT**<sup>®</sup>



[Inicio](#) | [¿Qué hacemos?](#) | [¿Qué es CAOBA?](#) | [Contacto](#)



## Resultados científicos de los proyectos de la ALIANZA CAOBA 2018

Resultados científicos de los proyectos d...  
Ver más tarde Compartir

EnVivo Universidad EAFIT

IA  
INTELIGENCIA ARTIFICIAL CIENCIA VS. FICCIÓN

# ¿Qué es CAOBA?

## Centro de Excelencia y apropiación en Big Data y Data Analytics

Primer acuerdo público – privado que promueve el Big Data y la analítica en Colombia.

Nace de la convocatoria  
Colciencias-MinTIC 2015

En Operación desde 2016



# ¿Quiénes somos?

Participan:



El futuro  
es de todos

DNP  
Departamento  
Nacional de Planeación



El futuro digital  
es de todos

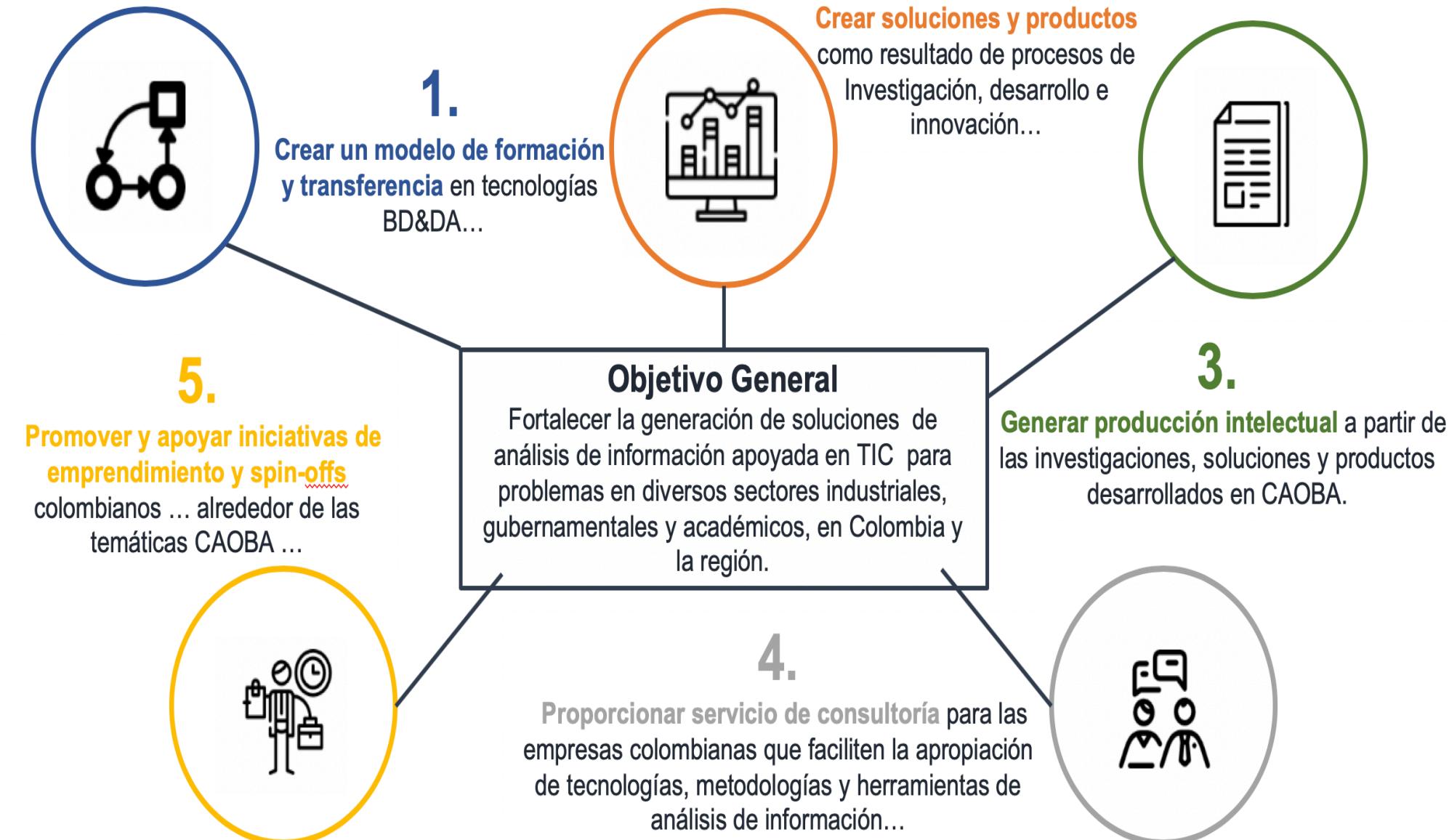
MinTIC



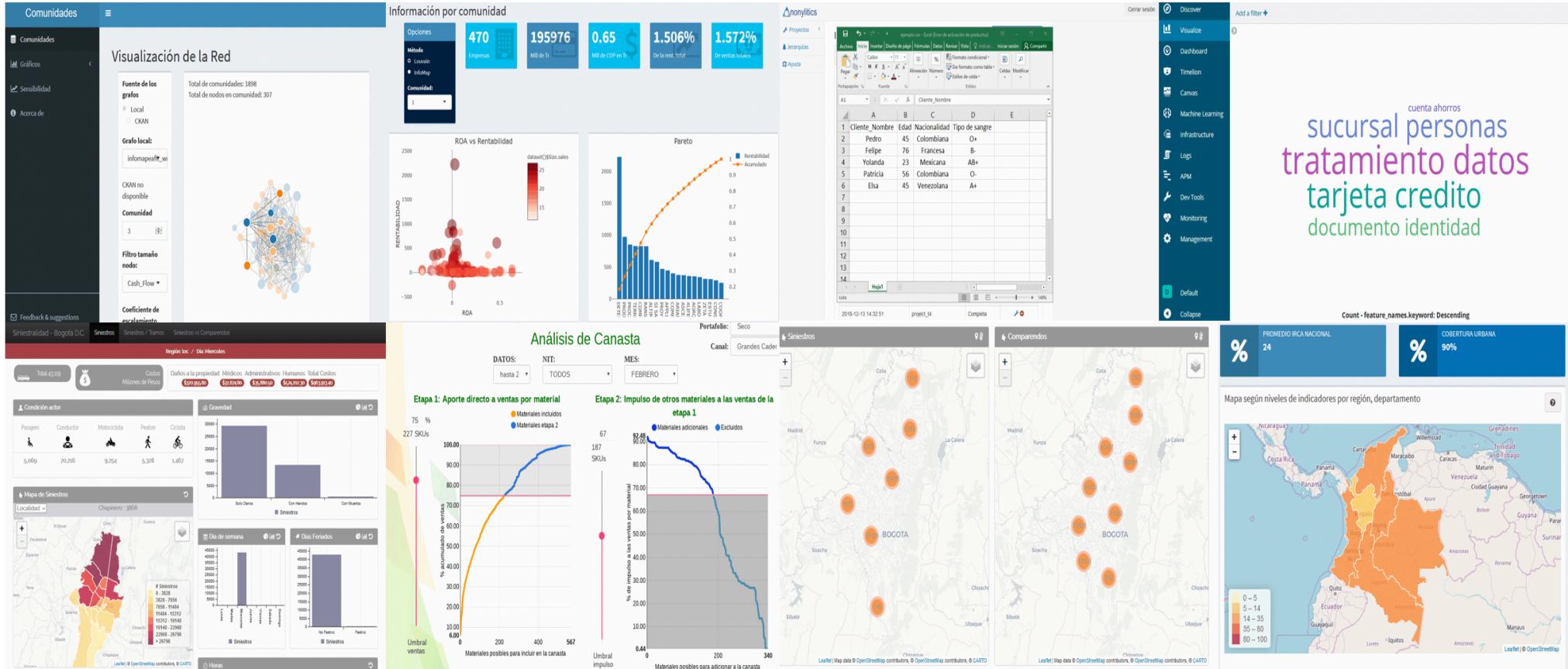
El conocimiento  
es de todos

Colciencias

# Objetivos CAOBA

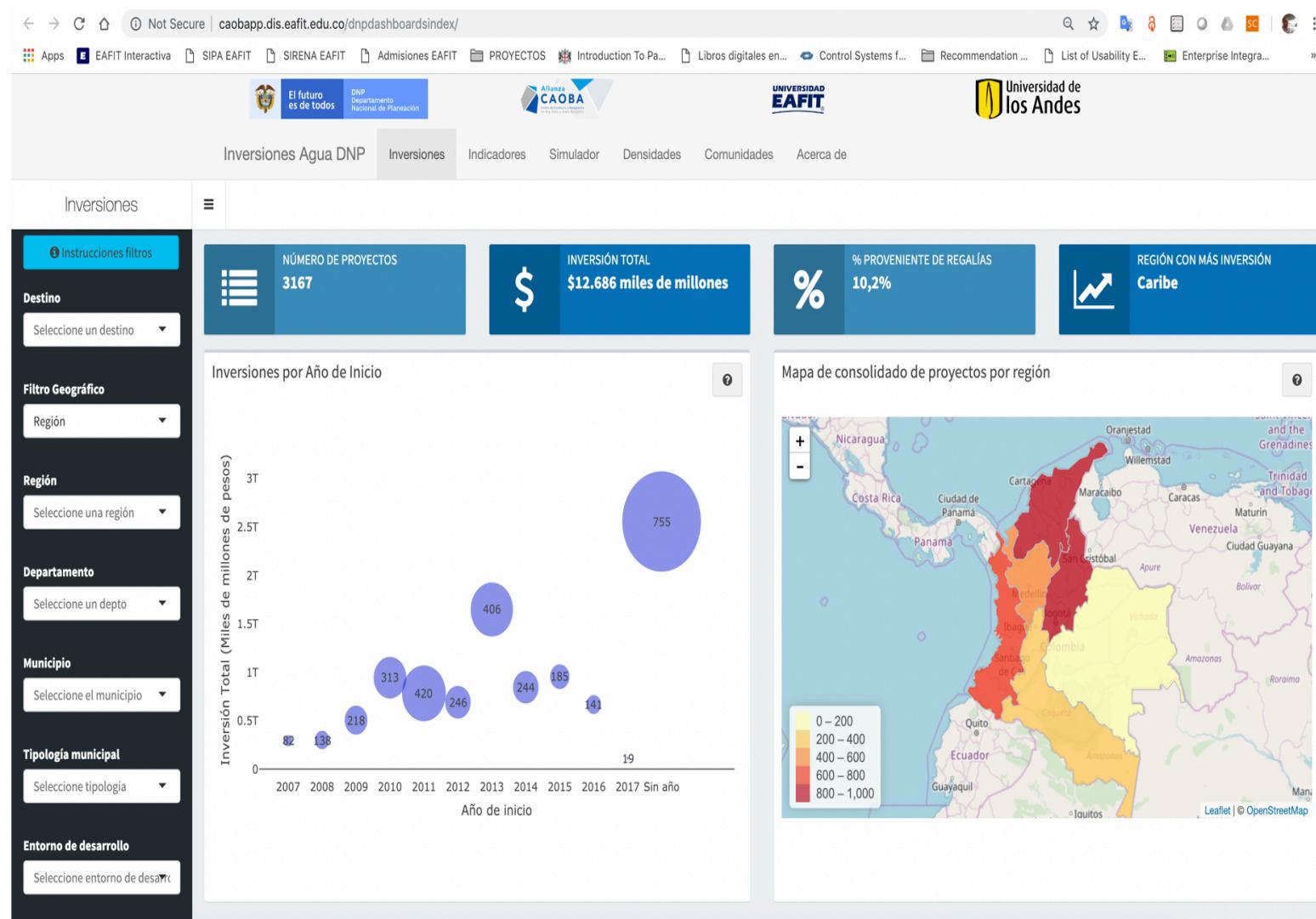


# Resultados de la investigación aplicada



# INVERSIONES AGUAS DNP

Cuantificación del efecto de las inversiones en el sector de agua potable en la prestación del servicio de agua.



## CENTROS PARA LA 4a Revolución Industrial



El Centro es una plataforma para discutir cuestiones éticas, valores y regulación de tecnologías de la Cuarta Revolución Industrial, tales como Internet de las cosas (IoT) y la inteligencia artificial / aprendizaje automático. Los proyectos involucran a múltiples partes interesadas, incluyendo a los reguladores, para desarrollar marcos de políticas que puedan ser aplicados a través de industrias y fronteras nacionales.

<https://es.weforum.org/focus/centro-para-la-cuarta-revolucion-industrial>

13 de diciembre de 2017

## 200 ciudadanos podrán formarse en analítica de datos y TI con la convocatoria de Científicos de Datos

Hasta el 30 de enero de 2018, profesionales y tecnólogos en ingeniería, administración, economía, matemáticas, estadística o carreras afines, podrán postularse a esta convocatoria, que busca fortalecer la productividad y competitividad de la industria TIC nacional.



La convocatoria de Científicos de Datos del Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC), Colciencias y el Centro de Excelencia y Apropiación en Big Data y Data Analytics (CAOBA), busca la formación y certificación de 200 ciudadanos colombianos como Citizen Data Scientists, en un curso de 70 horas de duración.

### Información para

Ciudadano

Industria

Emprendedores

Sector Académico

Sector Gobierno

### Categorías VUTIC

Registros

Espectro Radioeléctrico

Radiodifusión Sonora

Servicios Postales

RABCA:  
Radioaficionados y  
Banda Ciudadana

Verificación IMEI /  
Autorización Venta  
Terminales Móviles



**300 Profesionales**

se formarán en una de las habilidades necesarias para afrontar los retos que trae la Cuarta Revolución Industrial.

Formación en **Ciencia de Datos**

**Inspira Crea Transforma**

UNIVERSIDAD  
**EAFIT**<sup>®</sup>

# RutaN – Medellín – Centro regional para la 4a Rev Industrial



**MinTIC**

Ministerio de Tecnologías  
de la Información y las Comunicaciones

Grandes esfuerzos en promoción, capacitación, regulación, etc para la industria 4.0



EN MEDELLÍN SE INAUGURÓ EL  
**CENTRO PARA LA CUARTA  
REVOLUCIÓN INDUSTRIAL**

[CONOCE MÁS](#)

**Medellin, sede del Centro para LA de la  
4a Rev industrial**

<https://www.rutanmedellin.org/es/cuarta-revolucion-industrial>

# Algunos programas de Especialización y Maestría

Inspira Crea Transforma

UNIVERSIDAD  
**EAFIT**<sup>®</sup>



**Especialización en  
Inteligencia  
Artificial**

**COMING SOON**

Área Curricular de Ingeniería de Sistemas e Informática  
Facultad de Minas

Recuerdame Recuérdame

Universidad Nacional de Colombia

UnalMedellin



Estudiar en EAFIT Academia Investigación Bienestar y cultura International Proyección Acerca de EAFIT

## Maestría en Ciencias de los Datos y Analítica

Inicio La maestría ▾ Estructura académica Contacto



<http://www.eafit.edu.co/maestria-ciencias-datos-analitica>



Universidad de los Andes Colombia

Facultad de Ingeniería Departamento de Ingeniería Industrial

Inicio Departamento Programas Académicos Investigación Servicios Nuestro equipo Boletines

Usuario Contraseña Recuérdame Identificarse

Maestría en Inteligencia Analítica para la Toma de Decisiones (Analytics)

Código SNIES: 104198

MAESTRÍA EN INTELIGENCIA ANALÍTICA PARA LA TOMA DE DECISIONES (Analytics)

Home Contenido Objetivo

<https://industrial.uniandes.edu.co/es/programas-academicos/maestria/maestria-en-inteligencia-analitica-para-la-toma-de-decisiones>

Inspira Crea Transforma



<https://www.icesi.edu.co/facultad-ingenieria/maestria-en-ciencia-de-datos>

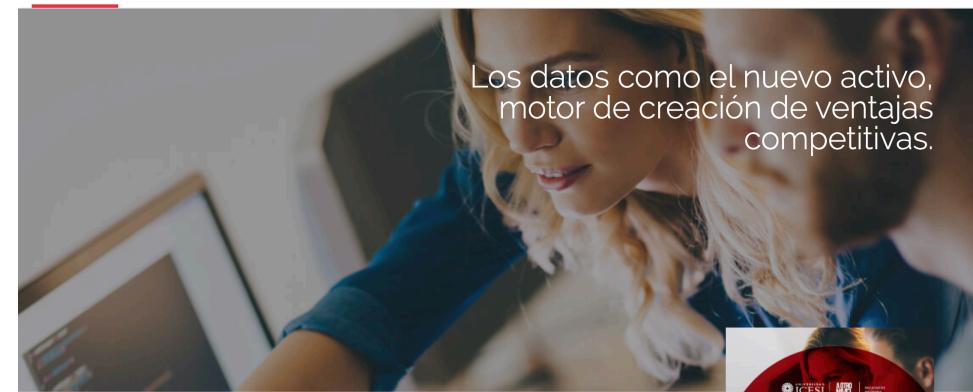
## Maestría en Ciencia de Datos

Facultad de Ingeniería

LA MAESTRÍA RAZONES PLAN DE ESTUDIO ADMISIÓN

APLICAR

Los datos como el nuevo activo, motor de creación de ventajas competitivas.



## MAESTRÍA EN ANALÍTICA PARA LA INTELIGENCIA DE NEGOCIOS

SNIES 105238



"Nuestra tradición es construir el futuro"

El programa de Maestría en Analítica para Inteligencia de Negocios es una iniciativa interdisciplinaria para formar consultores en análisis de datos para las organizaciones que apoyen la toma de decisiones estratégicas.

El Programa tiene una mezcla única de formación en:  
Capacidades analíticas y de ciencia de los datos.  
Capacidades de gestión de información.  
Capacidades administrativas, organizacionales y de comunicación.

<https://www.javeriana.edu.co/maestria-analitica-para-inteligencia-de-negocios>

Inspira Crea Transforma

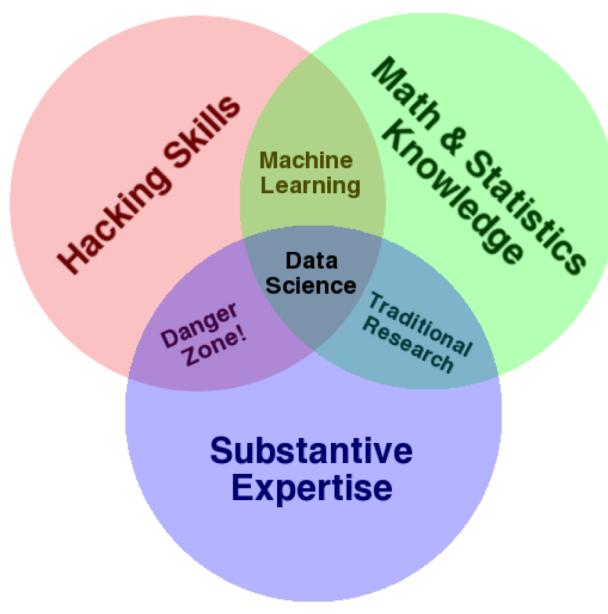
UNIVERSIDAD  
**EAFIT**<sup>®</sup>

# Perfiles profesionales en Big Data y Analítica



Inspira Crea Transforma

UNIVERSIDAD  
**EAFIT**<sup>®</sup>

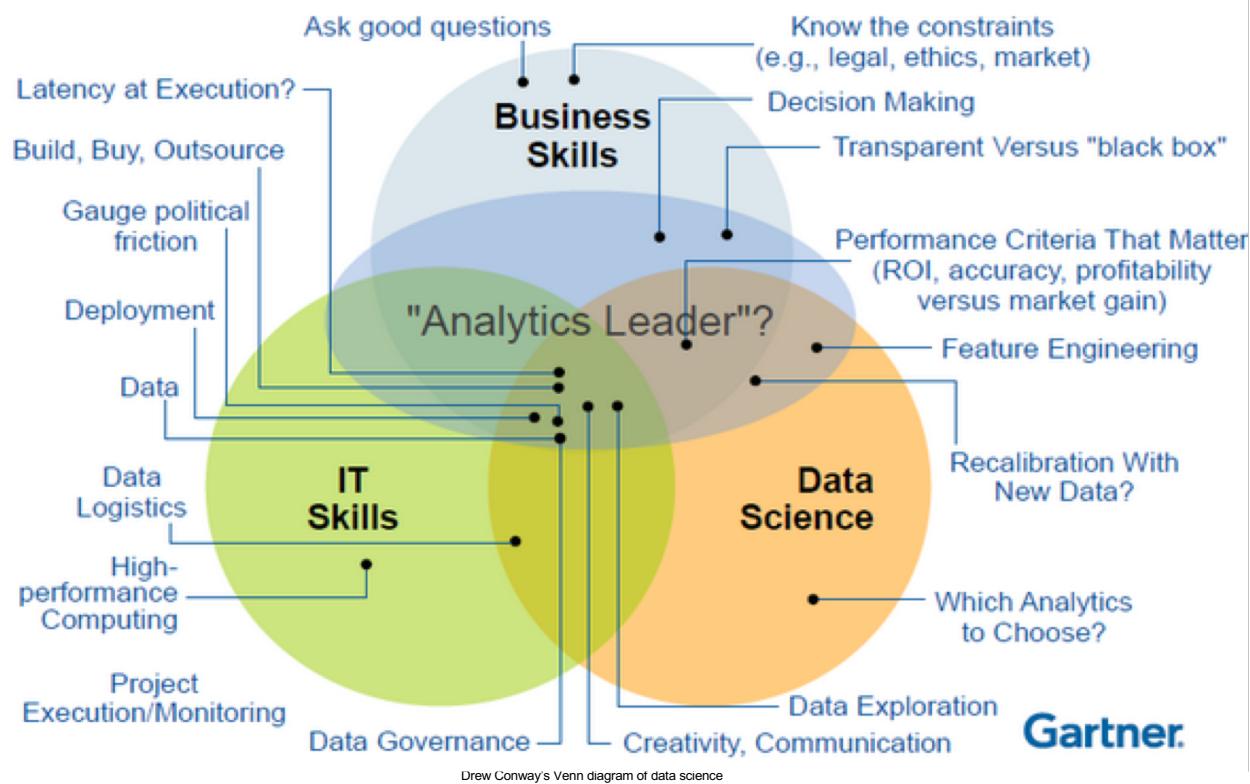


Drew Conway's Venn diagram of data science

Inspira Crea Transforma

UNIVERSIDAD  
**EAFIT**<sup>®</sup>

## Driving the Success of Data Science Solutions: Skills, Roles and Responsibilities ...



**Data Scientist**  
also known as Data Managers, statisticians.



A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.

**Skills:** Mathematics, Programming, Communication



**Will use programmes such as:**  
SQL, Python, R

**Data Engineers**  
also known as database administrators and data architects.



They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.

**Skills:** Programming, Mathematics, Big data



**Will use programmes such as:**  
Hadoop, NoSQL, and Python

**Data Analysts**  
also known as business Analysts.



They typically help people from across the company understand specific queries with charts.

**Skills:** Statistics, Communication, Business knowledge



**Will use programmes such as:**  
Excel, Tableau, SQL

**DATA**  
Engineer

**DATA**  
Scientist

DataCamp  
Learn Data Science By Doing

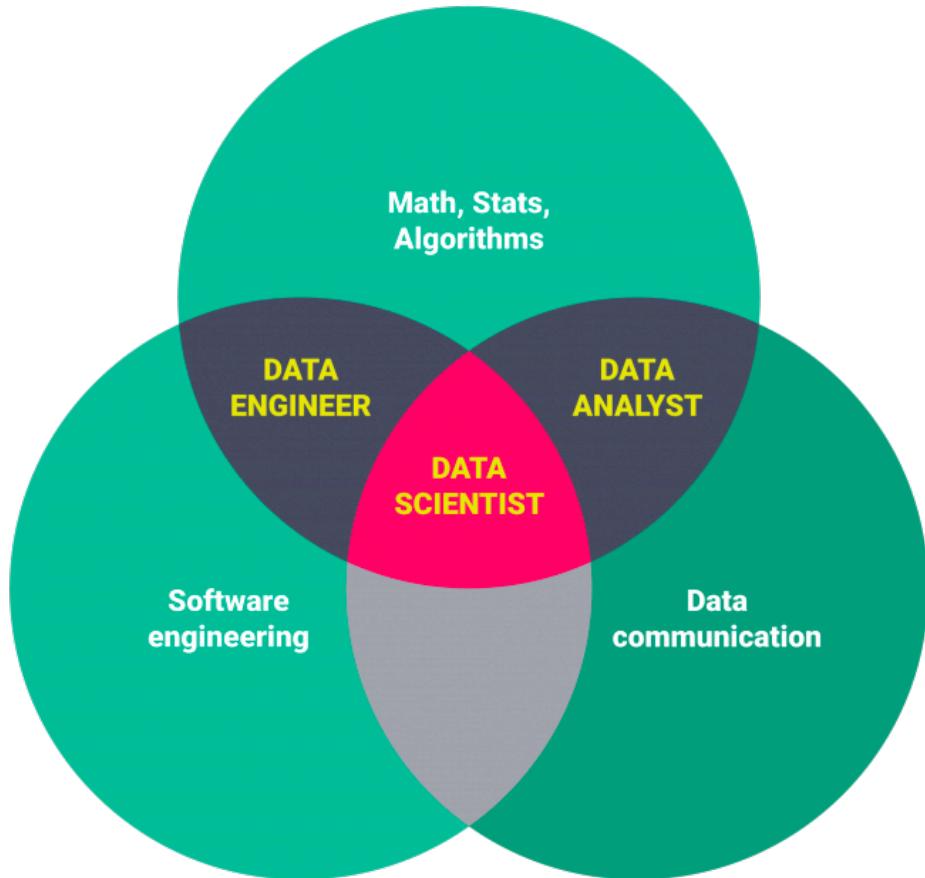
Develops, constructs, tests, and maintains architectures. Such as databases and large-scale processing systems.

Cleans, massages and organizes (big) data. Performs descriptive statistics and analysis to develop insights, build models and solve a business need.

A cartoon illustration shows two people at desks. On the left, a man with a beard and headphones works on a laptop; a red coffee cup sits on his desk. On the right, a woman with glasses and a yellow top works on a laptop; there are books and a small potted plant on her desk. Above them, several floating data visualization icons (charts, graphs, and a neural network diagram) represent the work of both roles.

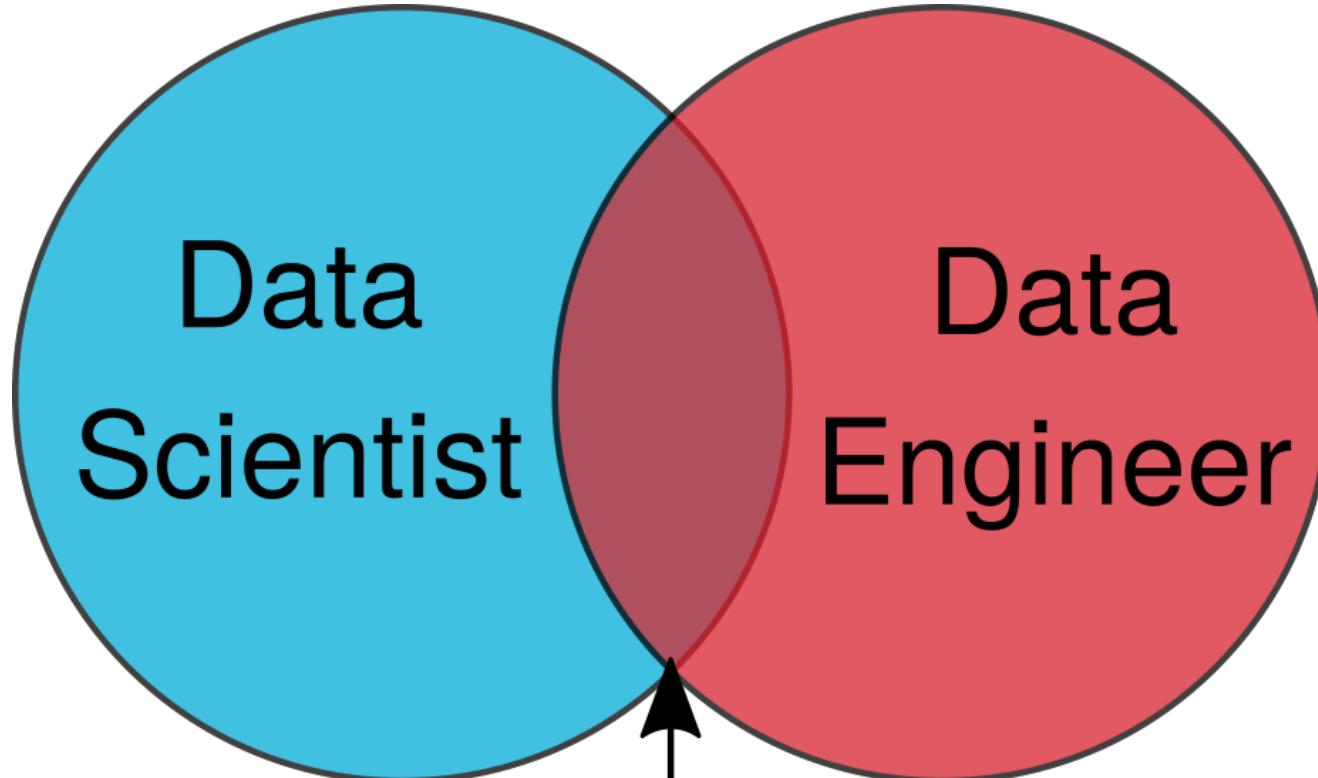
Inspira Crea Transforma

UNIVERSIDAD  
**EAFIT**<sup>®</sup>



Inspira Crea Transforma

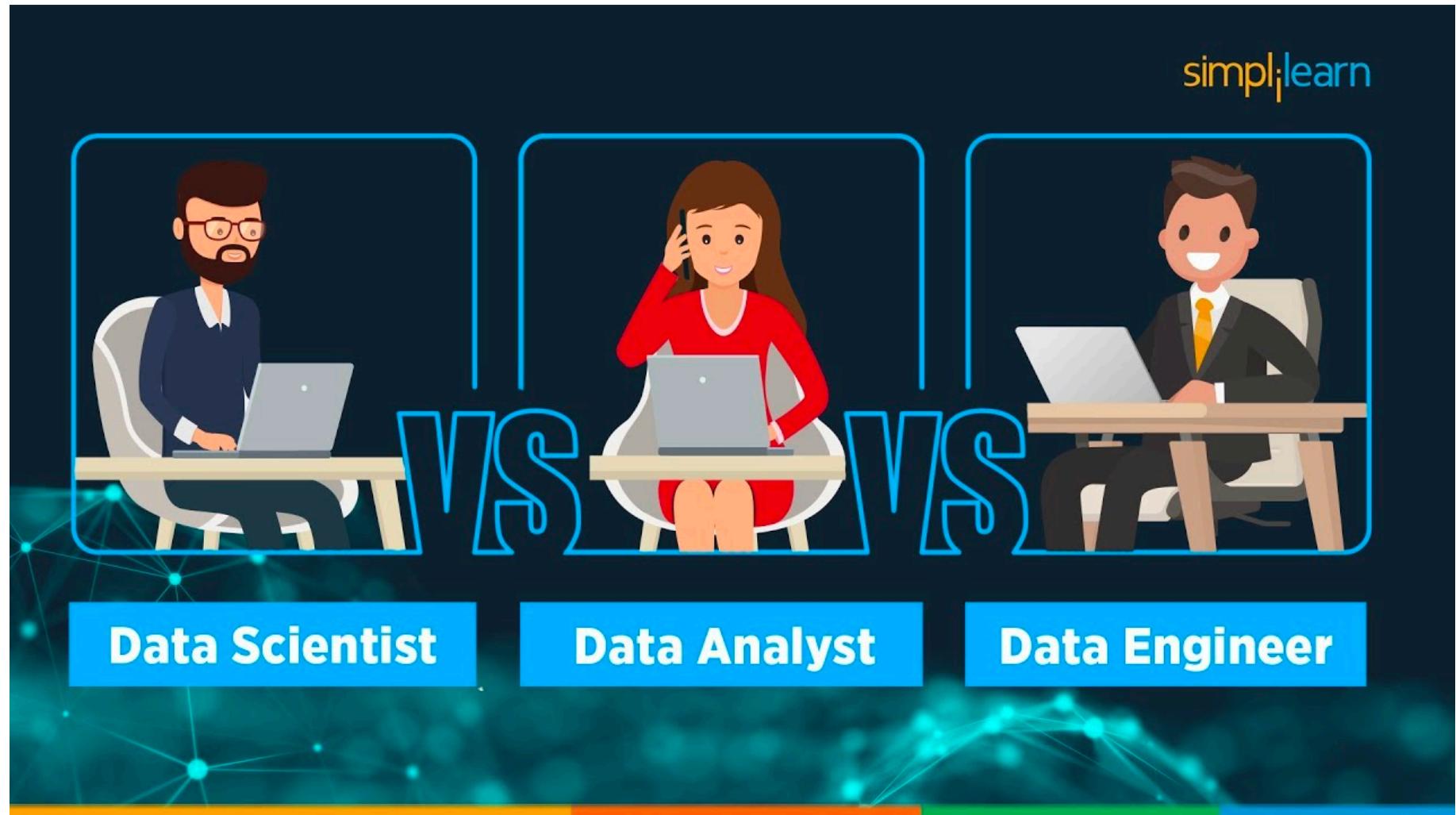
UNIVERSIDAD  
**EAFIT**<sup>®</sup>



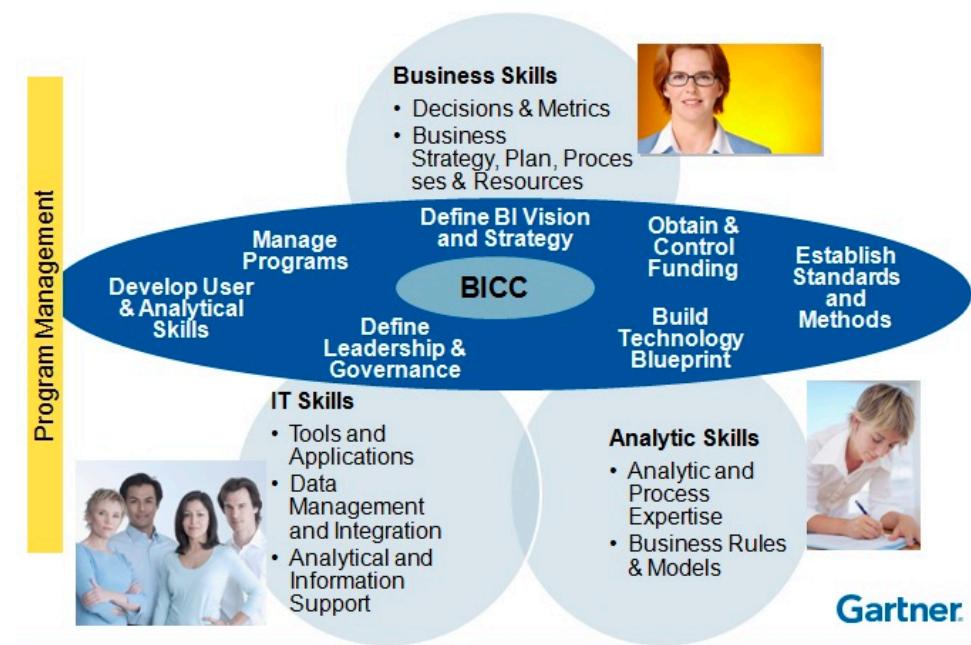
Big Data

Inspira Crea Transforma

UNIVERSIDAD  
**EAFIT**<sup>®</sup>



# BICC de Gartner



Inspira Crea Transforma

UNIVERSIDAD  
**EAFIT**<sup>®</sup>

# Conclusión

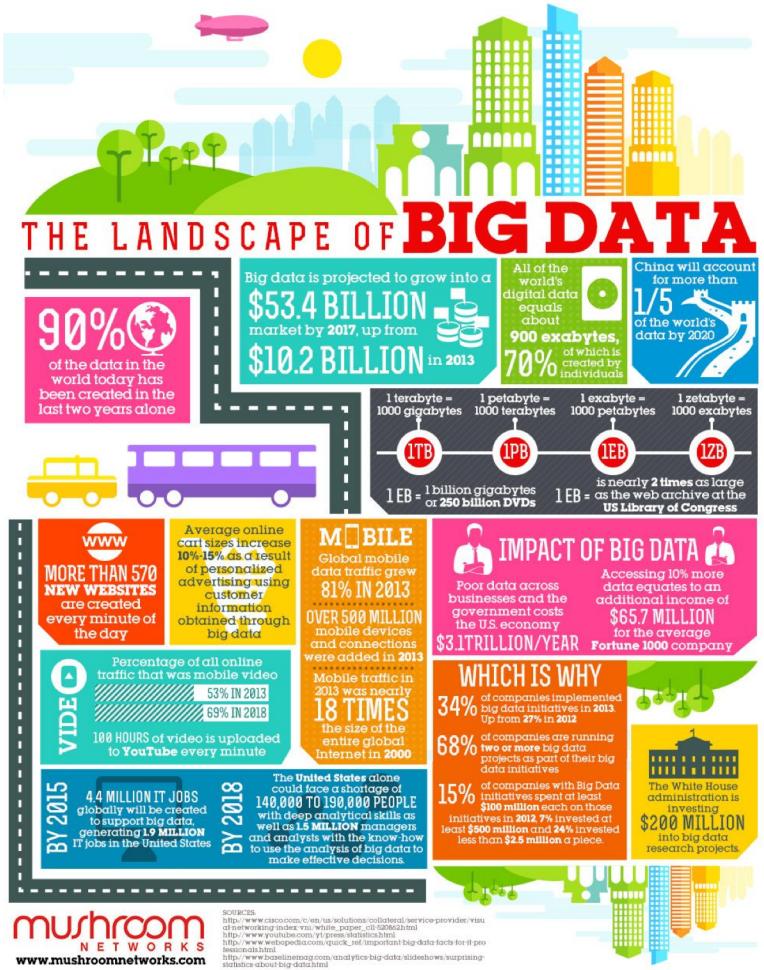
- Una sola persona NO puede manejar todos las habilidades
- Se requiere un trabajo multi e inter disciplinario
- ¿Tú como quieras jugar en este mundo del Big Data?
  - Desde la Tecnología
  - Desde la Ciencia de los datos
  - Desde el Negocio

# ESTE TUTORIAL ESTA ENFOCADO A?

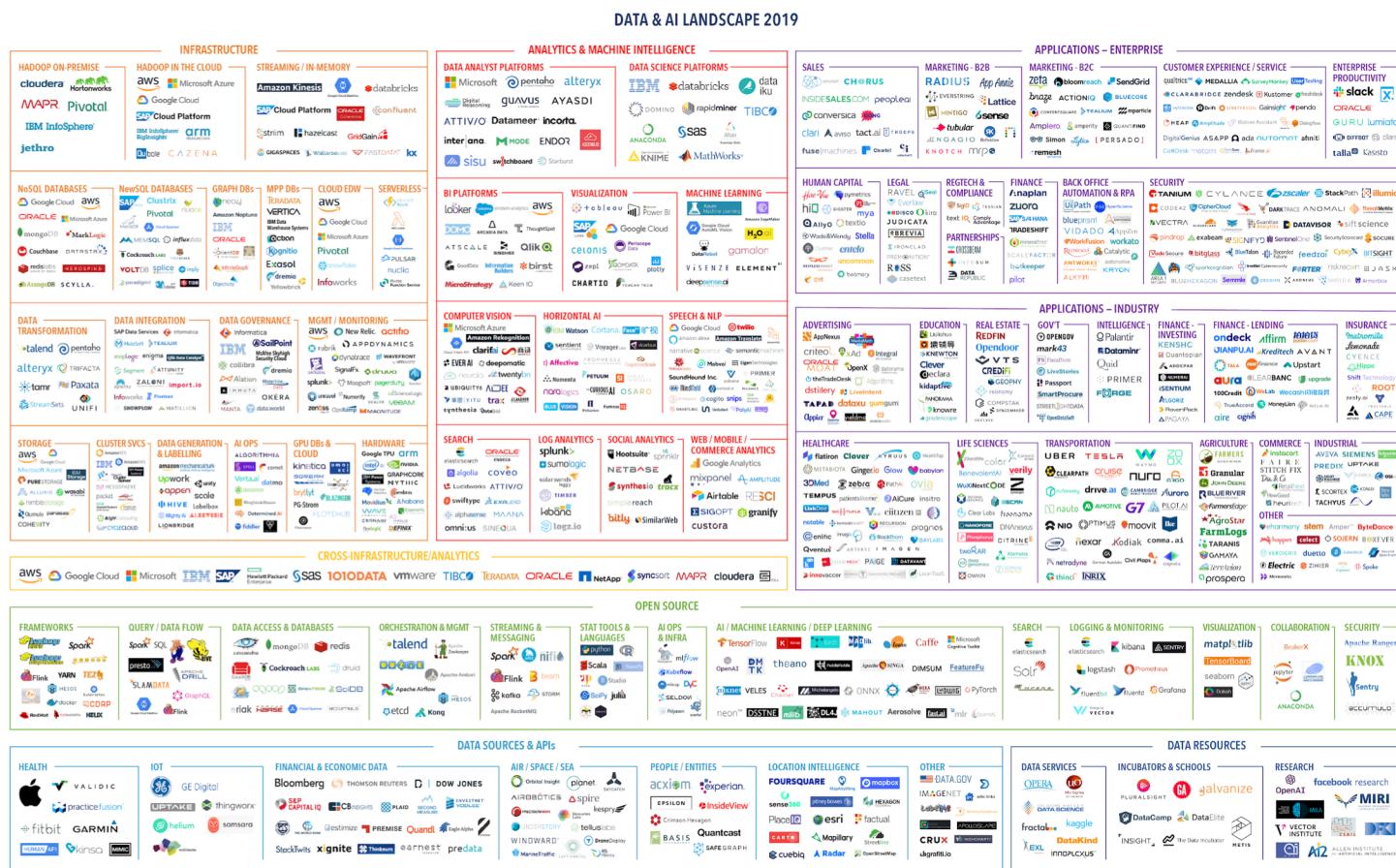
Ingeniería de Datos

Inspira Crea Transforma

UNIVERSIDAD  
**EAFIT**<sup>®</sup>



Inspira Crea Transforma



July 16, 2019 - FINAL 2019 VERSION

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

[mattturck.com/data2019](http://mattturck.com/data2019)

FIRSTMARK  
EARLY STAGE VENTURE CAPITAL

# Inspira Crea Transforma

**UNIVERSIDAD  
EAFIT**®

# 1. Introducción a Big Data

Inspira Crea Transforma

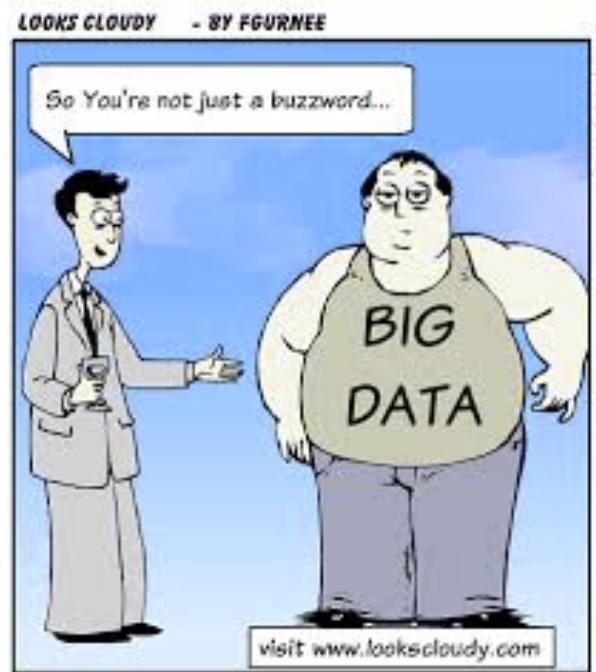


# Qué es Big Data

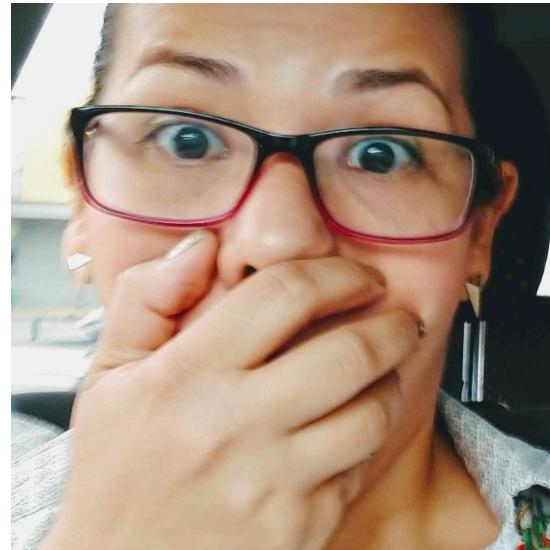
Es el área de conocimiento dentro de la informática que se ocupa de solucionar problemas de transporte, almacenamiento, procesamiento, análisis y aprovechamiento de datos caracterizados como Big Data, esto ocurre especialmente cuando las tecnologías tradicionales no son suficientes.

# ¿Qué es Big Data?

- Un Buzzword?



# ¿El tamaño si importa?



Porqué???

Inspira Crea Transforma

UNIVERSIDAD  
**EAFIT**<sup>®</sup>

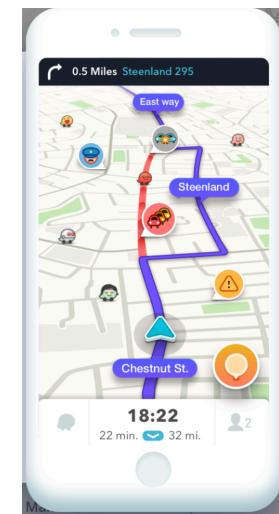
# Big (Grandes) o Many (Muchos)

- El colapso viene por la cantidad, NO tanto por el tamaño
- Ej:
  - Waze



Google Maps usa cerca de 0.6MB de data por hora.

Waze usa cerca de 0.23MB por hora



# Pero el tamaño si importa

- 151 millones de suscriptores, en más de 200 países.
- Donde los almacena?
  - En la Nube – S3
  - 60+ PB con 1500+ millones de objetos (algo así como películas)
  - Diario procesa más de 400 TB

NETFLIX



# Características Big Data

Las 5 Vs pueden ser utilizadas para diferenciar conjuntos de datos categorizados como “Big Data” de los demás conjuntos de datos.

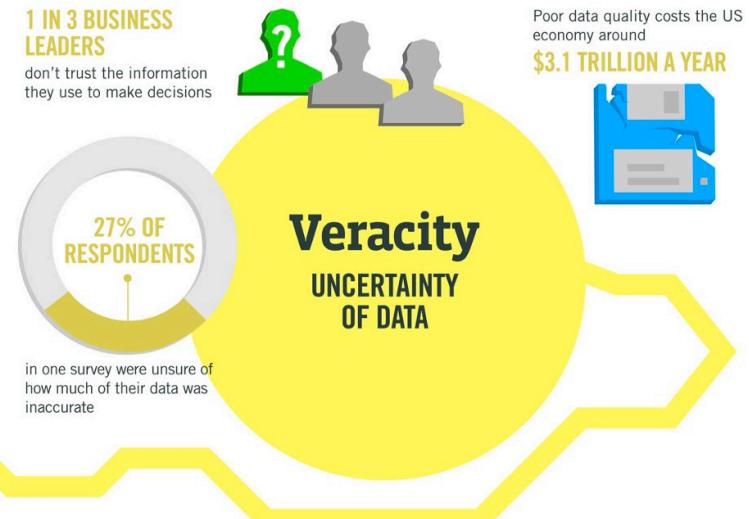
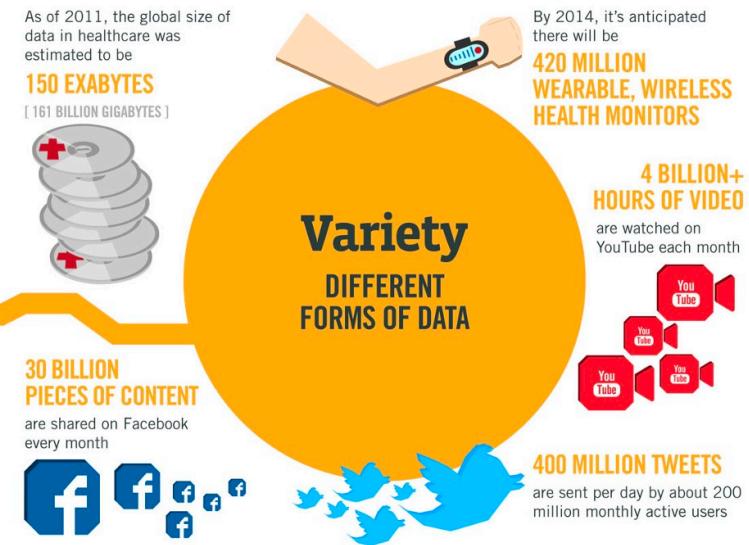
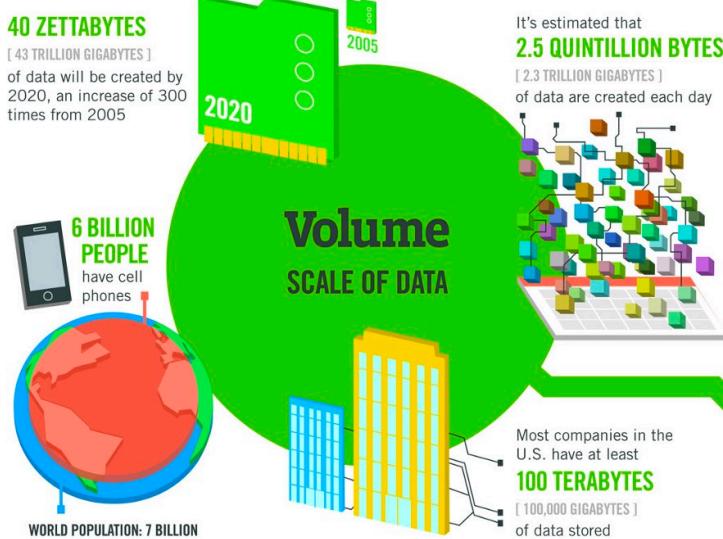
- Volumen
- Velocidad
- Variedad
- Valor
- Veracidad

# Tipos de datos Big Data

Los tipos de datos que una solución Big Data puede procesar son:

- Estructurados.
- No estructurados.
- Semi-estructurados.

# Big Data - V's



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Inspira Crea Transforma

# Cómo reconocer y abordar un caso Big Data

Antes de iniciar la implementación de una solución Big Data debe:

- Cubrir por lo menos 2 de las 3 primeras Vs, las otras 2 Vs no son negociables.
- Proporcionar beneficios claros y de valor para el negocio.
- Implementar mecanismos para la divulgación y entrenamiento Big Data.
- Implementar modelos de gobierno Big Data.
- Implementar mecanismos de integración con otras áreas como: operaciones, seguridad, etc.
- Proveer mecanismos para la gestión, trazabilidad y lineage de los datos.

# Big Data: Ciclo de vida Desde la Ingeniería de Datos

- **DATA:** Generación y Adquisición de datos
- **STORAGE:** Sistemas de almacenamiento masivo y distribuido
- **PROCESAMIENTO:** Análisis & Data Analytics
- **APLICACIONES:**
  - Negocios, Científicos, Económicos, ... para casi todos los dominios
  - Apps por: Datos estructurados, texto, web, multimedia, redes, móviles
  - Empresariales, IoT, Social Networks, Health, e-Science, Smart grid, etc.
  - Visualización

# Big Data: Ciclo de vida

- **DATA:** Generación y Adquisición de datos (ETL: Extract & Transform & Load)
  - **Generación:** Fuente
    - Datos empresariales operacionales - OLTP
    - Datos IoT, Redes Sociales
    - Datos científicos en muchos otros campos
  - **Adquisición:**
    - Datasets y/o real-time -> Colecciones de datos
    - Transmisión de datos (Inter-DCN & Intra-DCN)
    - Transformación de datos & Pre-procesamiento de datos
      - Integración
      - Limpieza
      - Eliminación de redundancia

# Big Data: Ciclo de vida

- **STORAGE:** Sistemas de almacenamiento masivos y distribuidos.
  - Datos transaccionales vs Datos para Analítica
  - CAP (Consistency & Availability & Partition Tolerance)
  - Almacenamiento para datos masivos
  - Sistemas de almacenamiento distribuidos
  - Mecanismos de almacenamiento:
    - Tecnologías tradicionales: DAS, SAN, NAS
    - File System (NFS, GFS, HDFS)
    - Database system (SQL & NoSQL)
      - Tradicionales -> SQL
      - Key-Value, Column-oriented, Document, etc, -> NoSQL
    - Repositorios, ej: tipo CKAN

# Big Data: Ciclo de vida

- **PROCESAMIENTO:** Análisis & Data Analytics (big data analytics)
  - Análisis de datos tradicional – OLAP-DWH
  - Métodos de analítica en Big Data
  - Arquitectura para análisis de datos Big Data
    - Análisis Real-Time vs Offline (batch)
    - Análisis a diferentes niveles
  - Data Mining & ML

# Big Data: Ciclo de vida

- **ANALÍTICA:** Procesamiento & Data Analytics

- Arquitectura para análisis de datos Big Data
  - Análisis Real-Time
  - Datos que cambian constantemente
  - Arquitecturas:
    - Clúster de procesamiento paralelo usando RDBMS tradicionales
    - Plataformas de computación basadas en memoria (ej: greenplum de EMC, HANA de SAP, Spark)
  - Offline (batch)
    - El tiempo de respuesta no es critico
    - Hadoop, Scribe, Kafka, Chukwa, etc.
  - Hibridas -> lamda

# Big Data: Ciclo de vida

- **APLICACIONES:**

- Evolución de las apps:
  - Comerciales
  - Científicas
  - Sociales