

UD 1- Xestión de sistemas de almacenamento e ecosistemas big data.

Introdución

Índice

Introdución.....	3
A “cara oculta” da intelixencia artificial.....	3
Perfís profesionais.....	4
Perfís profesionais.....	5
Perfís profesionais.....	6
Analista de datos (“data analyst”).....	6
Perfís profesionais.....	7
Científico/a de datos (“data scientist”).....	7
Perfís profesionais.....	8
Enxeñeiro/a de datos (“data engineer”).....	8
Perfís profesionais.....	9
Enxeñeiro/a de aprendizaxe automático “machine learning engineer”.....	9
Que é BIG DATA?.....	10
As 3 V do Big Data.....	11
... ou as 5 V do Big Data.....	12
... ou as 7 V.....	12
Ordes de magnitude en Big Data.....	13
Características do Big Data.....	17

Introdución

A “cara oculta” da intelixencia artificial

- A meirande parte da infraestrutura que arrodea á intelixencia artificial corresponde ó que chamamos big data.

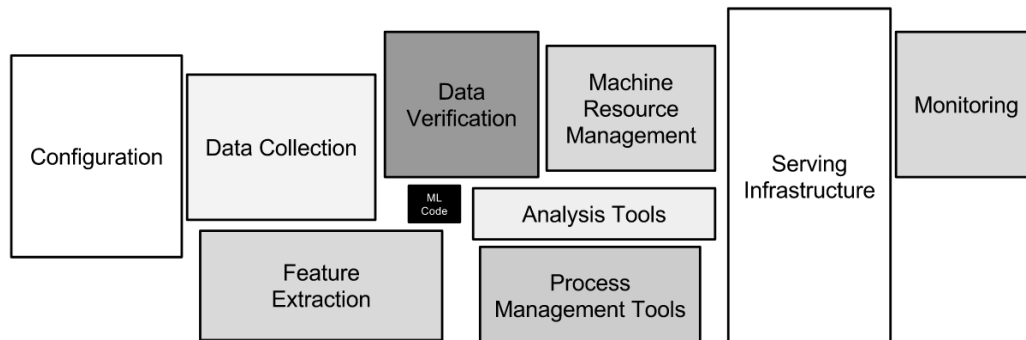
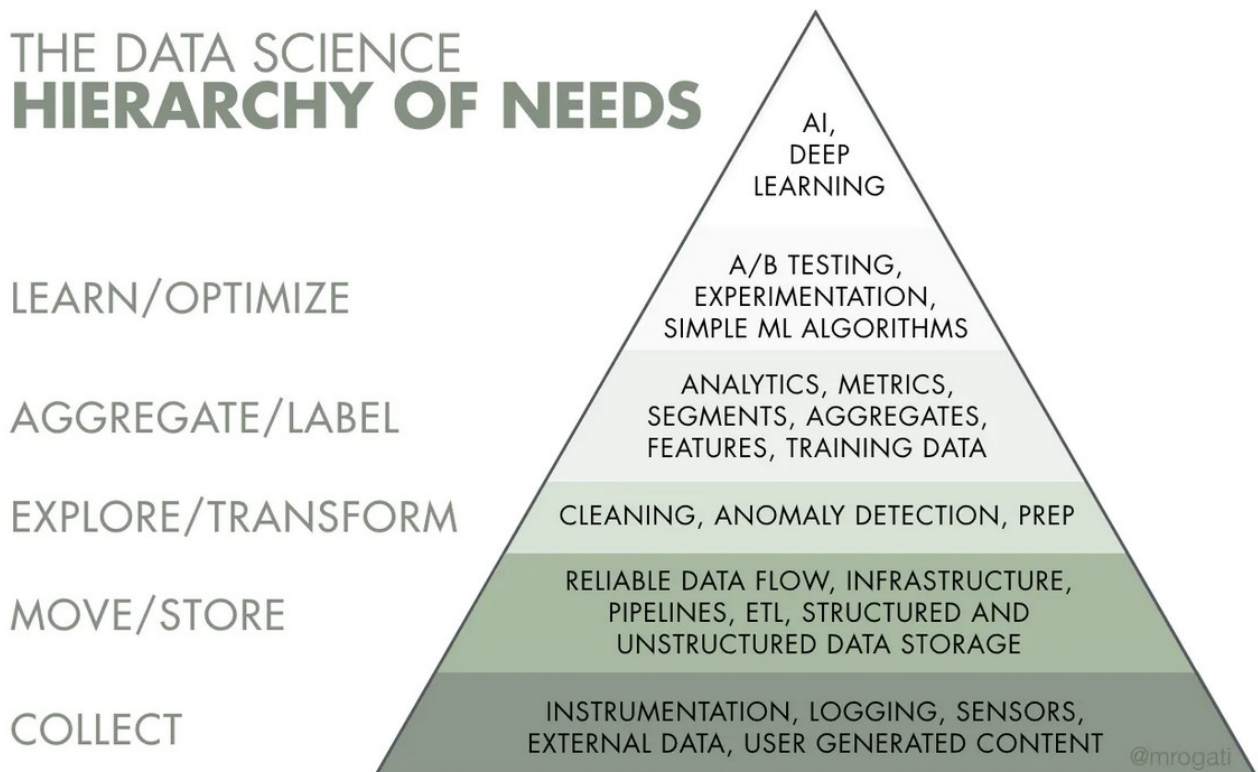


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

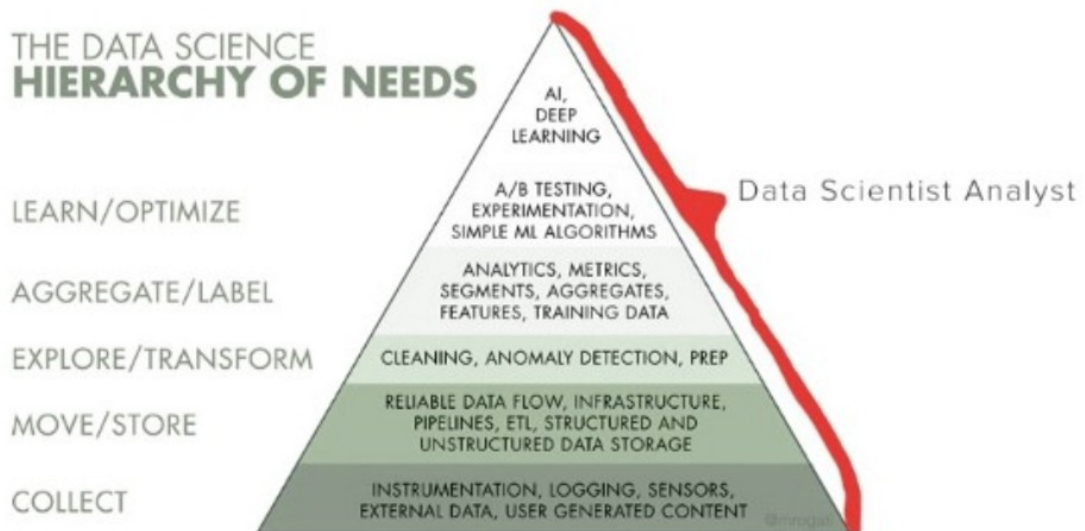
Fonte: <https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf>



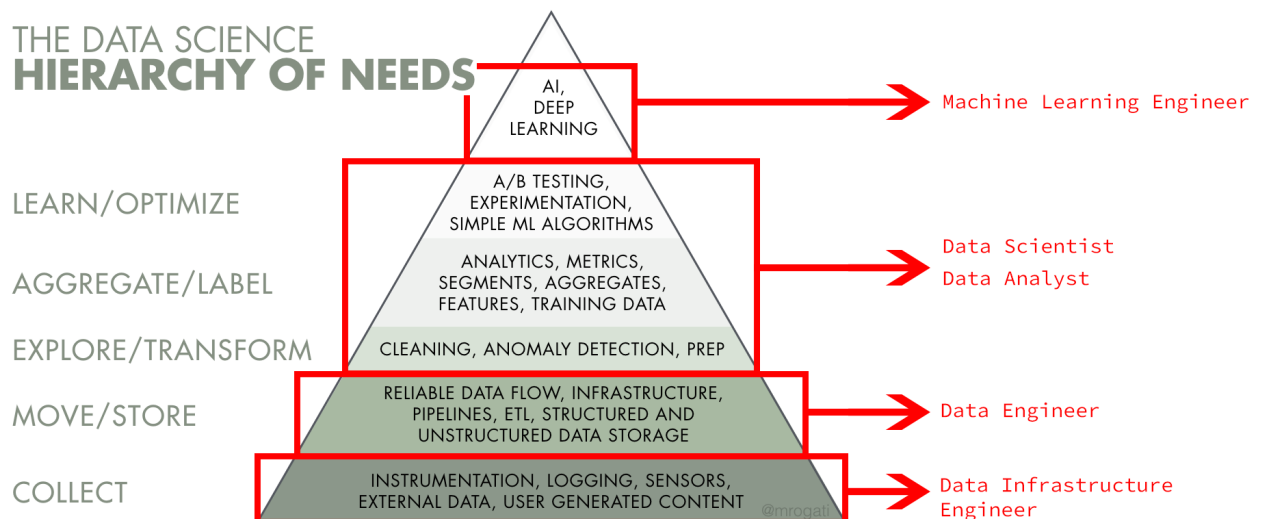
Fonte: <https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007>

Perfís profesionais

Sometimes, being a Data Scientist in a company could look like that:



- Actualmente existe un conxunto de perfís que traballan en contornas de intelixencia de negocio, análise de datos, ciencia de datos e outras áreas relacionadas, que non están claramente definidos.




The data science hierarchy of needs - Created by [Monica Rogati](#). Fonte: <https://towardsdatascience.com/data-engineer-vs-data-scientist-bc8dab5ac124>

Perfís profesionais

Nuevos perfiles para procesar, entender y visualizar datos


BIG DATA ARCHITECT

- Desarrolla, construye, prueba y administra arquitecturas Big Data.
- Crea esquemas para el data management para integrar, mantener, centralizar y proteger las fuentes de datos.




BIG DATA ENGINEER

- Desarrolla, construye, prueba y mantiene software como bases de datos o procesadores a gran escala.
- ETL (Extraction, Transformation and Load)




DATA SCIENTIST

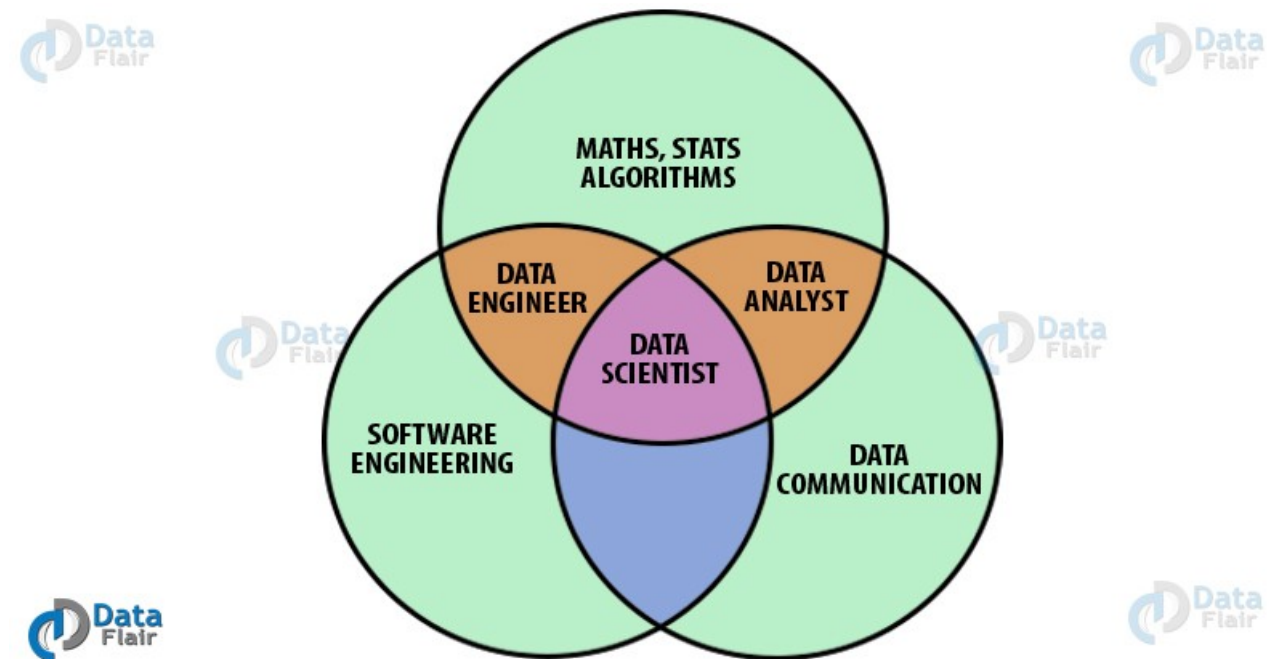
- Detecta patrones y analiza datos dentro de una organización para extraer el mayor valor de los mismos.
- Visualización y comunicación.



BIG DATA | Retos y desafíos del Big Data



Fonte: <https://www.coursera.org/learn/impacto-datos-masivos/> Curso MOOC de la UAB en Coursera Big Data: el impacto de los datos masivos en la sociedad actual BY-NC-SA.



Fonte: <https://data-flair.training/blogs/data-scientist-vs-data-engineer-vs-data-analyst/>

Perfís profesionais

Analista de datos (“data analyst”)

- Usa linguaxes de consulta (“**query languages**”) para recuperar e manipular información.
- Usa filtros e manipulacións de datos para reorganizar e limpa-los datos (“**data wrangling**”).
- Usa **técnicas estatísticas** descriptivas para producir.
- Colabora na **adquisición de requisitos** do proxecto.
- Comunica os resultados usando **informes e/ou visualizacións**.

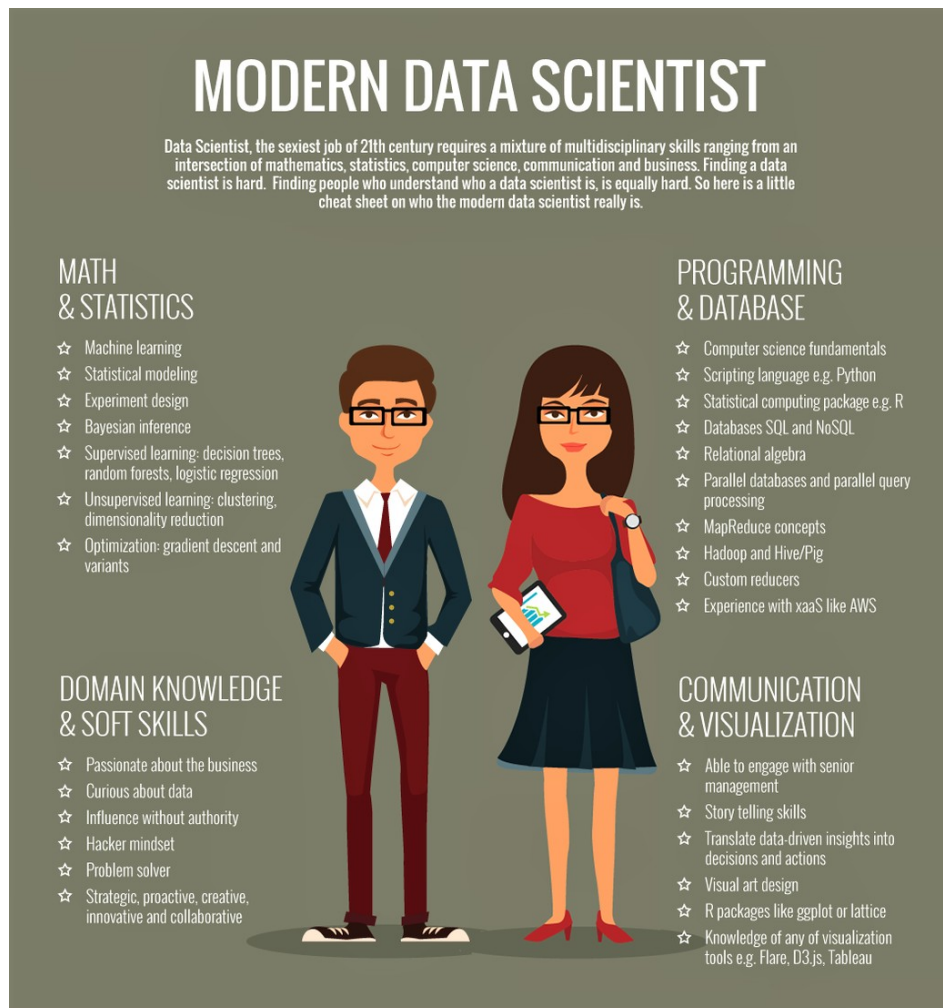


Panel de control. Fonte: <https://docs.microsoft.com/es-es/power-bi/create-reports/service-dashboards>

Perfís profesionais

Científico/a de datos (“data scientist”)

- Analista que comprende e captura as necesidades da organización con respecto ós datos (“**data needs**”).
- Usa ferramentas estatísticas, de intelixencia artificial e de **Machine Learning** para realizar predicións ou identificar patróns en grandes conxuntos de datos (“**datasets**”).
- Investiga e soluciona preguntas do negocio que soen requirir **modelos predictivos, prescritivos** ou incluso cognitivos.
- É capaz de **mellora-la precisión e o rendemento** de algoritmos Machine Learning.
- Utiliza mecanismos potentes de comunicación (“**storytelling**”) para comunica-los resultados obtidos.

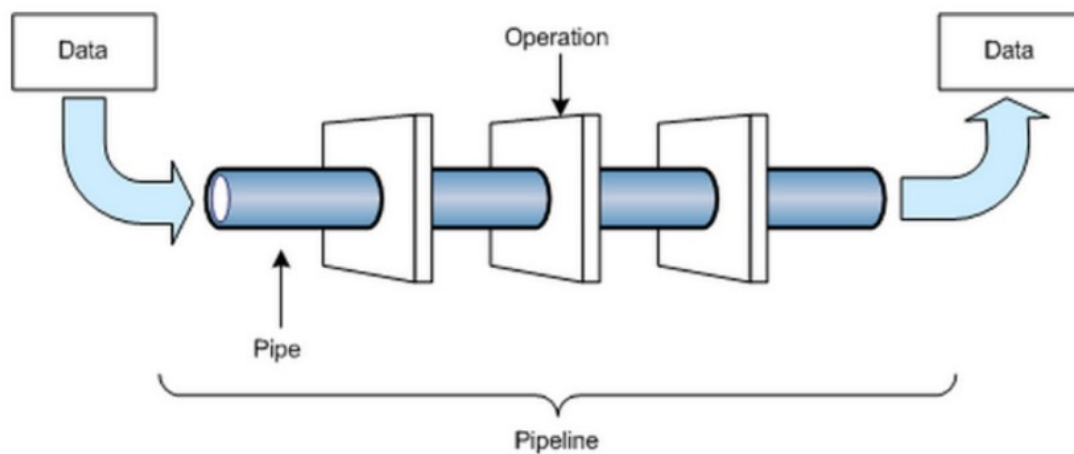


Fonte: <https://medium.com/metadatos/quiero-ser-cient%C3%ADfico-de-datos-pero-c%C3%B3mo-lo-consigo-5faec19d53ea>

Perfís profesionais

Enxeñeiro/a de datos (“data engineer”)

- Deseña e mantén a **Arquitectura de Datos** da organización.
- Deseña e mantén as plataformas de procesamento de datos a gran escala (“**big data platforms**”).
- Deseña e mantén **data pipelines**, xestionando o rendemento e posibles erros.
- Produce recomendacións para mellora-la calidade dos datos e a eficiencia no seu uso (“**data governance**”).
- Fai as **inxestas dos datos en cru e os procesa para que poidan ser utilizados** nas análises.



Pipeline created from raw data to end results data. Fonte: <https://towardsdatascience.com/data-engineer-vs-data-scientist-bc8dab5ac124>

Perfís profesionais

Enxeñeiro/a de aprendizaxe automático “machine learning engineer”

- Colabora co/a científico/a de datos na **creación de modelos ML**.
- Encárgase da posta en produción, monitorización de modelos ML (“**operacionalización**”).
- Verifica que os algoritmos ML en produción cumpren seus obxectivos e **adáptaos a cambios** na definición do problema a resolver ou nos datos de entrada.
- Presente especialmente en **organizacións grandes**, con procesos ML ben desenvolvidos.

Data Scientist vs. Machine Learning Engineer vs. Data Engineer

	Data scientist	Machine learning engineer	Data engineer
What they do	Build models that help business get better insights and make predictions from their data	Automate ML processes and make models work in a production environment.	Build, test and maintain data pipelines; provide ML models with quality data.
Skill set	<ul style="list-style-type: none">✓ Knowledge of math and statistics✓ Decision making and data optimization skills✓ High proficiency in SQL✓ Scripting skills (R/Python)	<ul style="list-style-type: none">✓ Solid programming background✓ Data science skills✓ Knowledge of math and statistics✓ Rapid prototyping skills✓ Good problem-solving-skills✓ Proficiency in deep learning frameworks	<ul style="list-style-type: none">✓ Scripting skills (Linux commands)✓ Knowledge of databases✓ Knowledge of cloud technologies✓ Proficiency in SQL✓ Data modelling skills✓ ELT development skills
Tools used	Python, R, Pandas, Jupyter notebooks, SQL	Python, PyTorch, TensorFlow, cloud services	SQL, Oracle, Hadoop, Amazon S3, Python



Fonte: <https://www.altexsoft.com/blog/machine-learning-engineer/>

Que é BIG DATA?

- Os **macrodatos**, tamén chamados **datos masivos**, **intelixencia de datos**, **datos a gran escala** ou **Big Data** é un termo que fai referencia a conxuntos de datos tan grandes e complexos que precisan de aplicacións informáticas non tradicionais de procesamento de datos para tratalos adecuadamente

<https://es.wikipedia.org/wiki/Macrodatos>

- Gartner**, definiu o Big Data como: “Big data is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.”
 - Podemos, pois, definir Big Data como a capacidade de procesar ...
 - ... grandes **volumes** de datos,...
 - ... de tipos moi distintos,...
 - ... a moi baixo custe,...
 - ... e a gran **velocidade**.

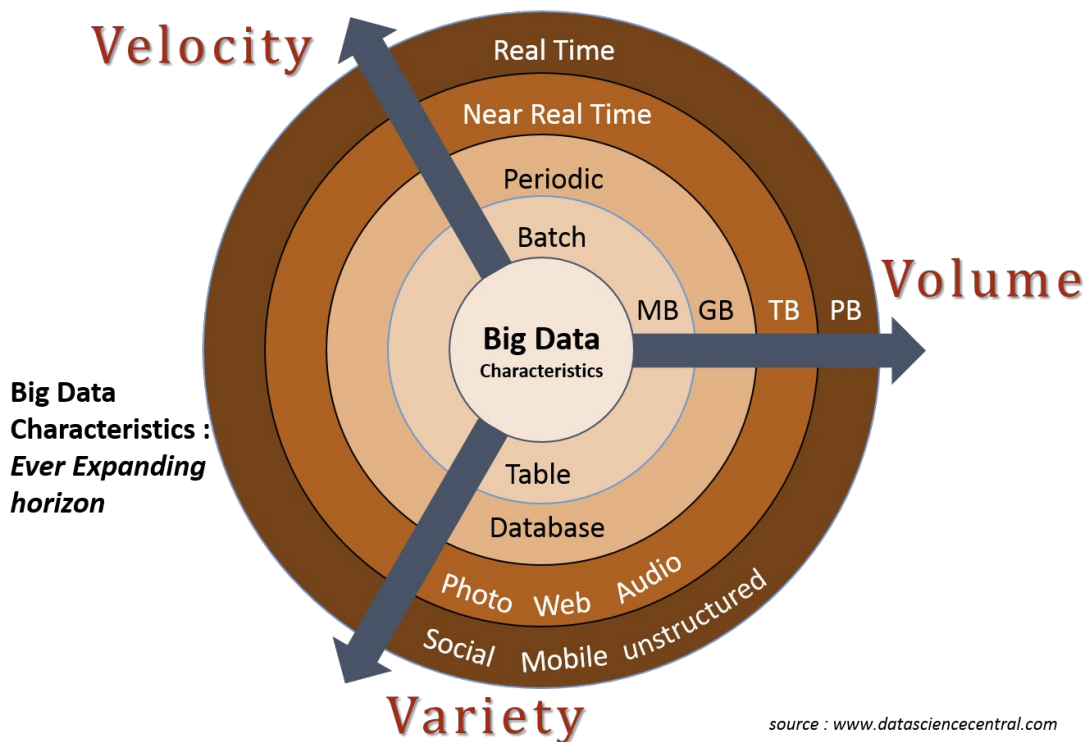


Fonte: <https://www.analyticsteps.com/blogs/top-10-big-data-technologies-2020>

- Veremos no curso que hoxe, poder tratar inmensas cantidades de datos conséguese con ...
 - ... computación distribuída ...
 - ... e almacenamento distribuído ...
- A solución as limitacións que existían foi “**levar a computación ós almacéns dos datos**”.

As 3 V do Big Data

- **Volume:** o big data, coma o seu propio nome indica, é sinónimo de grandes cantidades de datos, é a súa principal característica, xeralmente no rango de Terabytes ou Petabytes,
- **Velocidade:** analizar gran cantidade de información require moito tempo, pero a tecnoloxía permite analizar e procesar grandes cantidades de datos (de sensores, actualizacións frecuentes, tempo real...) con maior rapidez, incluso en tempo real. As solucións van dende o procesamento de datos mediante colas (en batch) ata o procesamento en streaming ou incluso en tempo real.
- **Variedade:** a procedencia dos datos é moi variable, non só SXBD, tamén texto, imaxes, audio, vídeo... Pero ademais, para o manexo deses datos non estruturados, o Big Data, proporciona solucións para a análise deses datos.



Fonte: <https://learncuriously.wordpress.com/2019/07/28/big-data-for-beginners/>

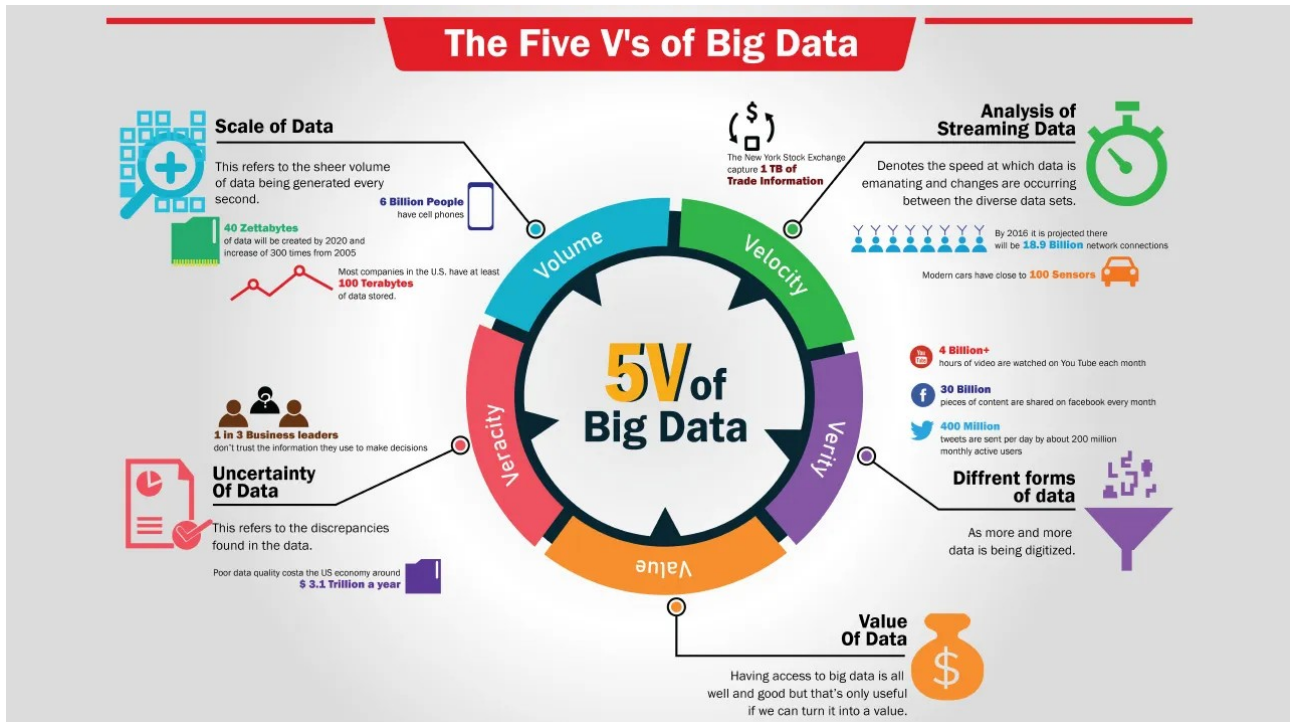
Big Data é a mellor opción cando

- se van tratar moitos datos, moi grandes
- os datos tómanse a gran velocidade, streamings
- os datos non están estruturados

Pola contra, non ten moito sentido executar en sistemas de Big Data, aplicacións que requiran de moita memoria RAM para o seu procesamento.

... ou as 5 V do Big Data...

- **Veracidade (fiabilidade):** hai que asegurar que os datos que se teñen sexan certos, do contrario os resultados tampouco serían correctos.
- **Valor:** como saber escolle-los datos a analizar que poidan aportar máis valor ós resultados.



Fonte: <https://morioh.com/p/ca19c6b8c0fe>

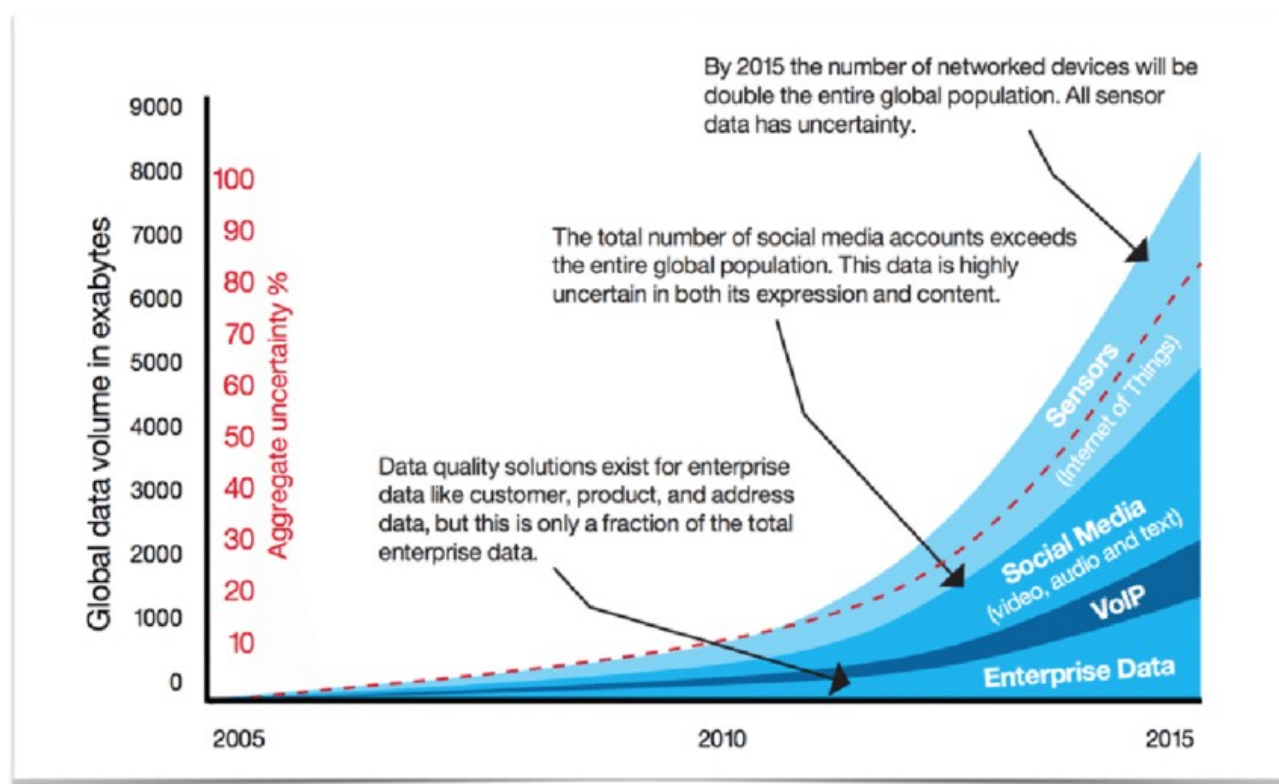
... ou as 7 V...

<https://www.iic.uam.es/innovacion/big-data-caracteristicas-mas-importantes-7-v/>

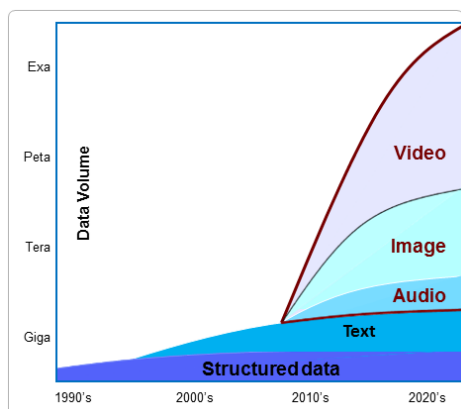
Ordes de magnitude en Big Data

Exemplos de Big Data:

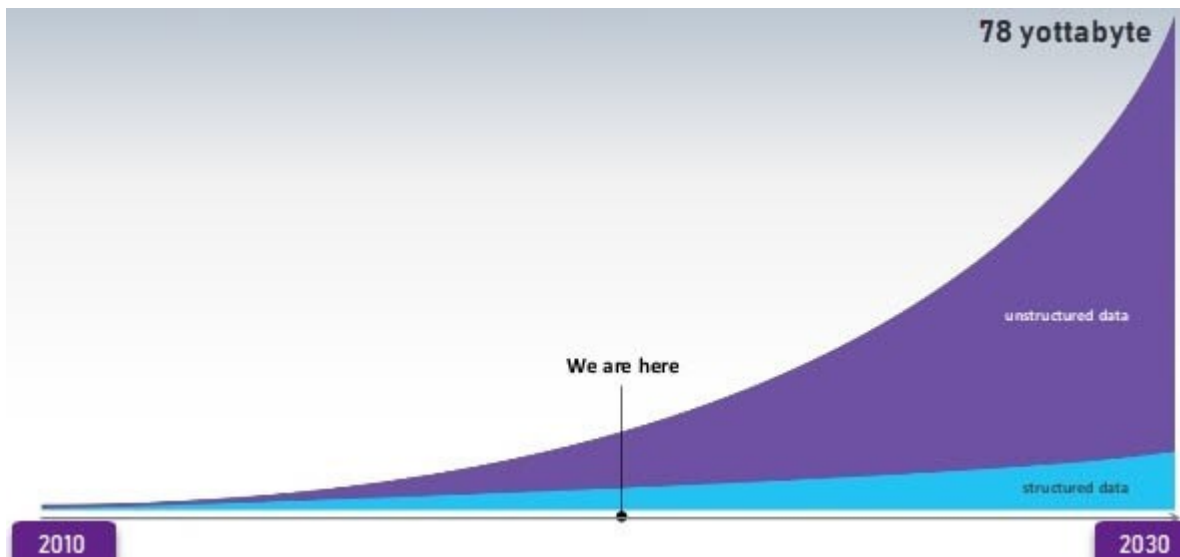
- Mercado financeiro: a New York Stock Exchange (a Bolsa) xera ao redor de 1 TB de novos datos por día.
- Social Media: estatisticamente sábese que Facebook xera 500 TB, cada día, principalmente en forma de fotos e vídeos.
- Aviación: estímase que os sensores dun avión comercial xeran de orde de 10 TB de información cada 30 minutos de voo.



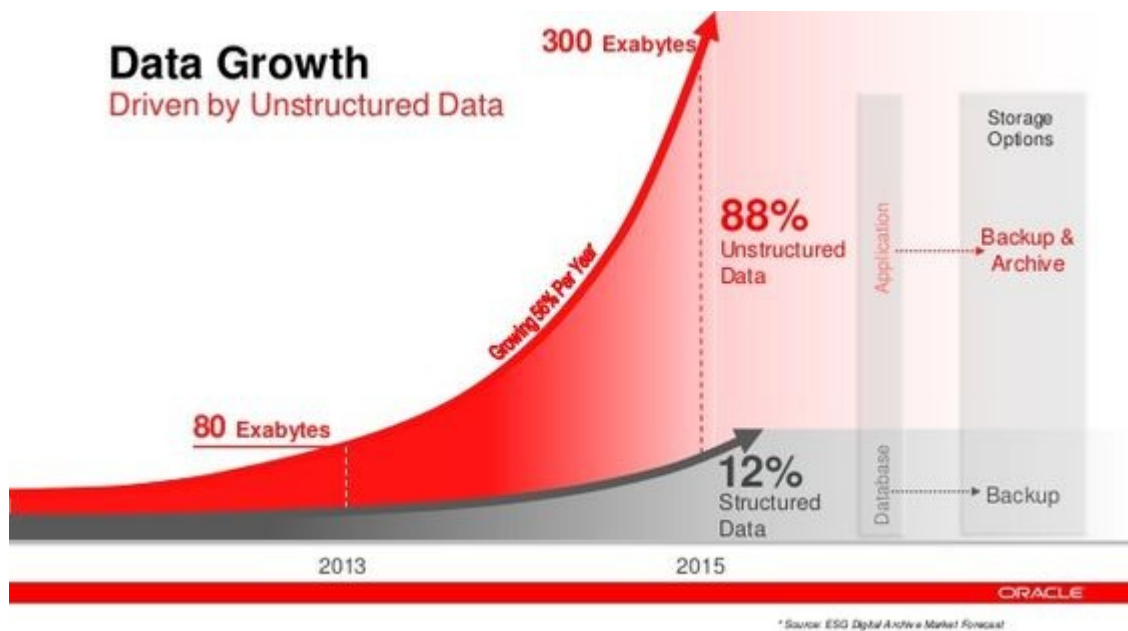
Fonte: https://www.researchgate.net/figure/The-exponential-increase-in-global-open-source-data-in-last-decade-The-dashed-red-line_fig2_309484436



Fonte: <https://avindh2014.wordpress.com/background/>



Fonte: <https://www.guru99.com/what-is-big-data.html>



Fonte: <http://techwolf4u.blogspot.com/2018/01/big-data.html>



By The Numbers

Projecting the future of digital transformation (2018–2023)



Global

Internet users by 2023



66%

of the population will be using the Internet
up from 51% in 2018

Mobile devices/ connections by 2023



1.6

networked devices and connections per person
up from 1.2 in 2018

Total devices/ connections by 2023



3.6

networked devices and connections per person
up from 2.4 in 2018

Fixed speed by 2023



110 Mbps

average broadband speed
up from 46 Mbps in 2018

Wi-Fi speed by 2023



92 Mbps

average Wi-Fi speed
up from 30 Mbps in 2018

Mobile (cell) speed by 2023



44 Mbps

average mobile speed
up from 13 Mbps in 2018

Cisco Annual Internet Report. Fonte: <https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html>

Western Europe

Internet users by 2023



87%

of the population will be
using the Internet
up from 82% in 2018

Mobile devices/ connections by 2023



2.9

networked devices and
connections per person
up from 1.7 in 2018

Total devices/ connections by 2023



9.4

networked devices and
connections per person
up from 5.6 in 2018

Fixed speed by 2023



123 Mbps

average broadband speed
up from 46 Mbps in 2018

Wi-Fi speed by 2023



97 Mbps

average Wi-Fi speed
up from 31 Mbps in 2018

Mobile (cell) speed by 2023



62 Mbps

average mobile speed
up from 24 Mbps in 2018

[Download the full report](#)

Cisco

Annual Internet Report

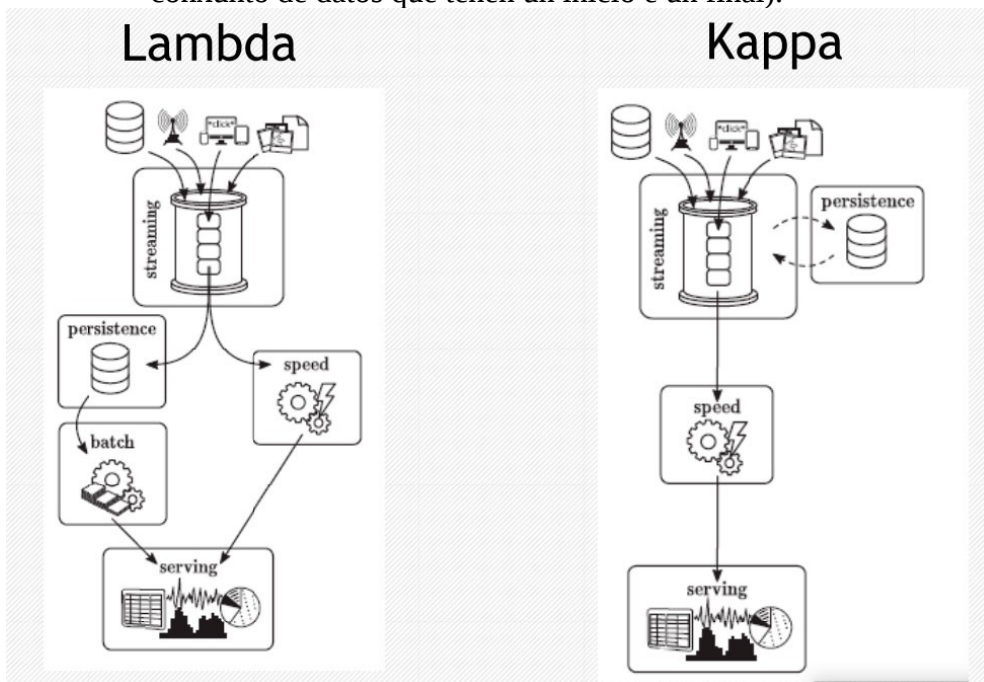


Cisco Annual Internet Report. Fonte: <https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html>

Características do Big Data

- Entre os V do Big Data, o **Volume e a Velocidade son fundamentais**, polo que precisase dunha **infraestrutura coa potencia e robustez** necesarias para o tratamento dos datos.
- Unha arquitectura de big data deséñase para **manexar**
 - a inxestión,
 - o procesamento
 - e a análise dos datos
- que son **demasiado grandes ou complexos para un sistema tradicional de base de datos**.
- As **características** da tecnoloxía e infraestrutura necesaria para o Big Data son as seguintes:
 - **Escalabilidade**: permite aumentar facilmente as capacidades de procesamento e almacenamento de datos.
 - **Tolerancia a fallos**: garante a dispoñibilidade do sistema, aínda que se produzan fallos nunhas das máquinas, evitando a perda de datos.
 - **Datos distribuídos**: os datos deben estar almacenados entre diferentes máquinas evitando así o problema de almacenar grandes volumes de datos nun único nodo central (SPOF. Single Point Of Failure).
 - **Procesamento distribuído**: o tratamento dos datos realízase entre diferentes máquinas para mellora-los tempos de execución e dotar ó sistema de escalabilidade.
 - **Localidade do dato**: os datos para traballar e os procesos que os tratan deben estar preto, para evita-las transmisións por rede que engaden latencias e aumentan os tempos de execución.
- Antes de falar das arquitecturas de Big Data, convén facer un inciso para defini-los dous **tipos de procesamento dos datos**:
 - Procesamento **Batch**.
 - Fai referencia a un proceso no que interveñen un conxunto de datos e que **ten un inicio e un fin no tempo**. Tamén se lle coñece como procesamento **por lotes** e execútase sen control directo do/a usuario/a.
 - Por exemplo, se temos un conxunto de datos moi grande con múltiples relacións, pode levarnos da orde de horas executar as consultas que necesita o cliente, e polo tanto, non se poden executar en tempo real e necesitan de algoritmos paralelos (por exemplo, Map Reduce). Nestes casos, os resultados almacénanse nun lugar diferente ó de orixe para posteriores consultas.
 - Outro exemplo, se temos unha aplicación que amosa o total de casos COVID que hai en cada cidade, no canto de realiza-lo cálculo sobre o conxunto completo dos datos, podemos realizar unha serie de operacións que fagan eses cálculos e os almacenen en táboas temporais (por exemplo, mediante INSERT ... SELECT), de maneira que se queremos volver realizar a consulta sobre tódolos datos, accederíamos ós datos xa calculados da táboa temporal. O problema é que este cálculo necesita actualizarse, por exemplo, de maneira diaria, e por iso é polo que habería que refacer tódalas táboas temporais.
 - É o procesamento que se realizou desde os inicios do traballo con datos, tanto a nivel de bases de datos como con *Data Warehouses*.
 - Da man do procesamento batch implantouse o ecosistema Hadoop con todas as ferramentas que abarcan un proceso ETL (extracción, transformación e carga dos datos).
 - Procesamento en **Streaming**.¶

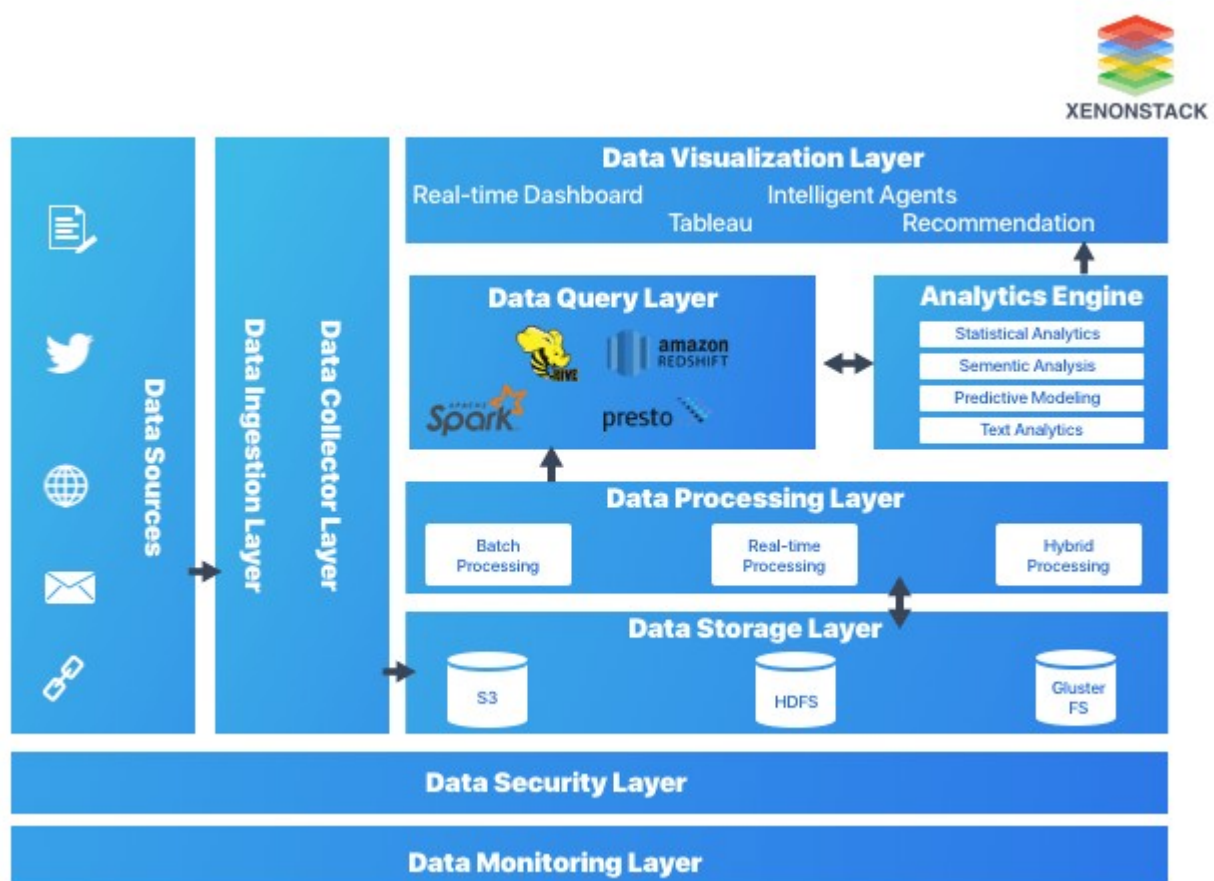
- Un procesamento é de tipo streaming cando **está continuamente recibindo e tratando nova información segundo vai chegando sen ter un fin** no tempo.
 - Este procesamento relaciónase coa análise en tempo real. Para iso, utilízanse diferentes sistemas baseados no uso de colas de mensaxes.
 - Ollo, non todo é tempo real; **non se debe confundir tempo real con immediatez**. En informática, un sistema de tempo real é aquel que responde nun período de tempo finito, normalmente moi pequeno, pero non ten por que ser instantáneo. Polo tanto o *Streaming* refírese a un *fluxo continuo de datos en tempo real*.
- En canto ós **tipos de arquitecturas**, as máis comúns para aborda-los problemas en Big Data son principalmente dúas, diferenciadas nos fluxos de tratamento de datos que interveñen:
 - **Kappa**, se o sistema necesita realizar un tratamento dos datos en **streaming** (proceso que trata fluxos de datos que se están recibindo continuamente),
 - **e Lambda**, se pola contra, pódese realizar un procesamento **batch** (proceso que trata un conxunto de datos que teñen un inicio e un final).



Arquitecturas Lambda e Kappa. Fonte: <https://www.trentia.net/procesamiento-de-grandes-volumenes-de-datos-el-caso-de-las-entidades-financieras/>

- Un exemplo real dunha arquitectura Kappa sería un sistema de xeolocalización de usuarios/as pola proximidade a unha antena de telefonía móbil. Cada vez que se aproximase a unha antena que lle dese cobertura xerárase un evento. Este evento procesárase na capa de streaming e serviría para pintar sobre un mapa o seu desprazamento respecto a a súa posición anterior.
- Un caso de uso real para unha arquitectura Lambda podería ser un sistema que recomende películas en función dos gustos dos usuarios. Por unha banda, tería unha capa batch encargada de adestrar o modelo e ir mellorando as predicións; e por outro, unha capa streaming capaz de encargarse das valoracións en tempo real.
- Máis: <https://www.ericsson.com/en/blog/2015/11/data-processing-architectures--lambda-and-kappa-examples>

- Ademais destas solucións, outra forma de deseña-las capas dunha arquitectura Big Data consiste en separa-las diferentes fases do dato en capa diferenciadas, no que se coñece como **arquitectura por capas**.
 - A arquitectura por capas **dá soporte tanto ó procesamento batch como por streaming**.
 - Consiste en 6 capas que aseguran un fluxo seguro dos datos:
 - **Capa de inxestión:** é a primeira capa que recolle os datos que proveñen de fontes diversas. Os datos se categorizan e priorizan, facilitando o fluxo destes en posteriores capas.
 - **Capa de recolección:** centrada no transporte dos datos dende a inxesta ó resto do pipeline de datos. Nesta capa os datos desfáanse para facilita-la analítica posterior.
 - **Capa de procesamento:** esta é a capa principal. Procésanse os datos recollidos nas capas anteriores (xa sexa mediante procesos batch, streaming ou modelos híbridos), e clasifícanse para decidir cara a que capa se dirixen.
 - **Capa de almacenamento:** céntrase en decidir onde almacenar de forma eficiente a enorme cantidade de datos. Normalmente nun almacén de arquivos distribuído, que dá pé ó concepto de *data lake*.
 - **Capa de consulta:** capa onde se realiza o procesado analítico, centrándose en obter valor a partir dos datos.
 - **Capa de visualización:** tamén coñecida como capa de presentación, é coa que interactúan os/as usuarios/as.



Arquitectura por capas. Fonte: <https://www.xenonstack.com/blog/big-data-ingestion>