

Introducción al Big Data

Conceptos previos y herramientas actuales

Contenido de la clase

¿Qué aprenderemos hoy?

- Definición formal del Big Data
- Herramientas para el Big Data
- Diferencia entre Big Data, Data Science, Machine Learning y AI
- Casos de usos.

¿Qué es el Big Data?

¿Qué es el Big Data ?

Estructura fundamental

- El **Big Data** es un termino que se emplea para una estructura de datos que cumplen con las siguientes condiciones :
 1. Volumen
 2. Velocidad
 3. Variabilidad
 4. Veracidad
 5. Valor
- También se asocia a una estructura de datos que posee dificultades de procesamiento, y que por lo cual requiere procesos no tradicionales para el procesamiento.

¿Qué es el Big Data?

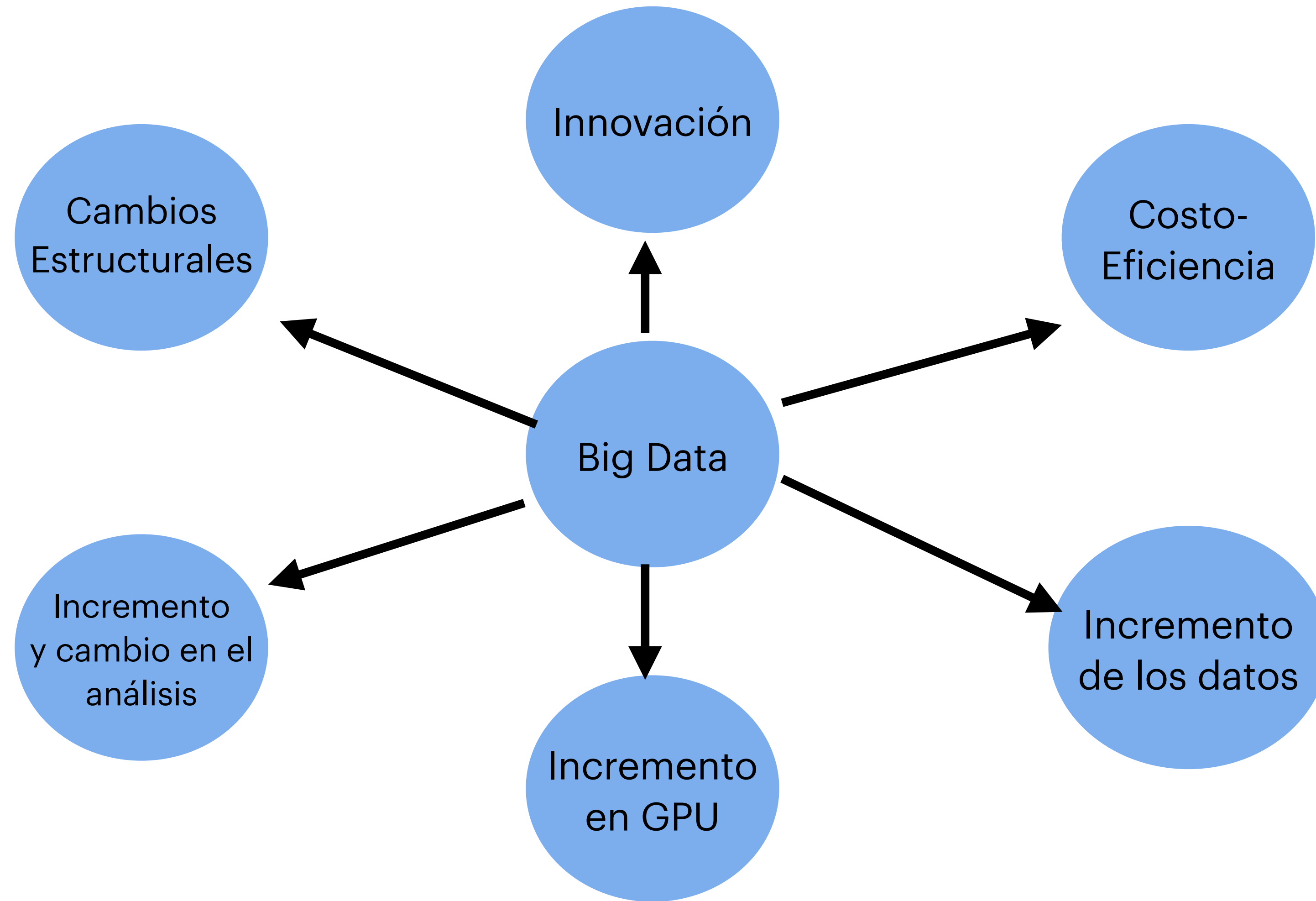
- Una de las complicaciones del Big Data tiene que ver con :
 1. Captura de los datos : Herramientas como Web Scraping son eficientes para estos temas.
 2. Estructurar la data : Convertir los set de datos bajo unos parámetros establecidos para que así se pueda trabajar con ellos.
 3. Confiabilidad de los datos: Conocimiento de las fuentes de datos.
 4. Transferencia de los datos : Que dichos datos se puedan usar en más de un caso.
 5. Análisis de los datos: Los datos deben tener una finalidad.
 6. Visualización de los datos : Formas de generar información.

¿Qué es el Big Data ?

Contextualizando la definición

- Con base a lo anterior, y usando la definición de Gartner se re-afina la definición como :
“ Big Data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making”.

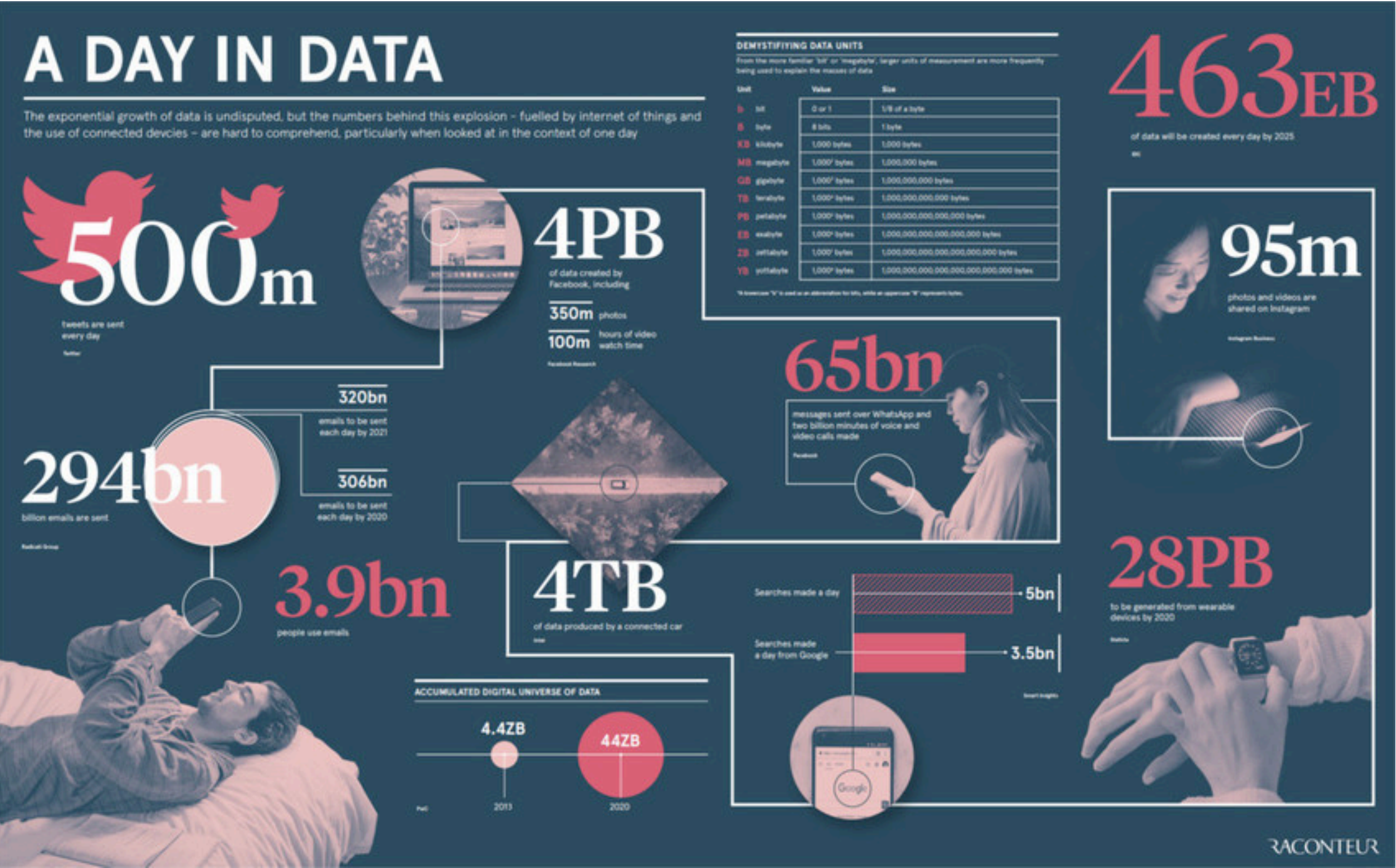
Big Data



Anatomía del Big Data

Velocidad

- Consiste en el cambio y crecimiento de la información
- La creciente de los datos debe basarse en el sentido del crecimiento de la información.
- Puede que dado el crecimiento de la información su velocidad, las bases de datos puedan desbordarse.



Abbreviation	Unit	Value	Size (in bytes)
b	bit	0 or 1	1/8 of a byte
B	bytes	8 bits	1 byte
KB	kilobytes	1,000 bytes	1,000 bytes
MB	megabyte	1,000 ² bytes	1,000,000 bytes
GB	gigabyte	1,000 ³ bytes	1,000,000,000 bytes
TB	terabyte	1,000 ⁴ bytes	1,000,000,000,000 bytes
PB	petabyte	1,000 ⁵ bytes	1,000,000,000,000,000 bytes
EB	exabyte	1,000 ⁶ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	1,000 ⁷ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	1,000 ⁸ bytes	1,000,000,000,000,000,000,000,000 bytes

Tomado de : <https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/>

Volumen

- Tiene que ver con la cantidad de información que se produce a través de dispositivos (IOT) y de forma manual (facturas , etc)
- El volumen es infrecuente a través de sus orígenes, por ello se debe calcular en proyección el tamaño de los set de datos.
- Una recomendación para calcular el crecimiento de los datos es recolectar al menos 30 días de información para calcular la proyección de la tasa de crecimiento de la misma.

Variedad

- Tiene que ver con la varianza de los datos.
+ Varianza : La cantidad de información que proveen los datos.
- Se refiere a las distintas fines que producen información.
- Se debe verificar la calidad de los datos a través de las fuentes.

Veracidad

- ¿Qué tan creíble es la fuente de información?
- ¿Tiene cómo compararse dicha información?
- ¿Puede diferenciarse de una fake ?

Herramientas del Big Data

Herramientas del Big Data

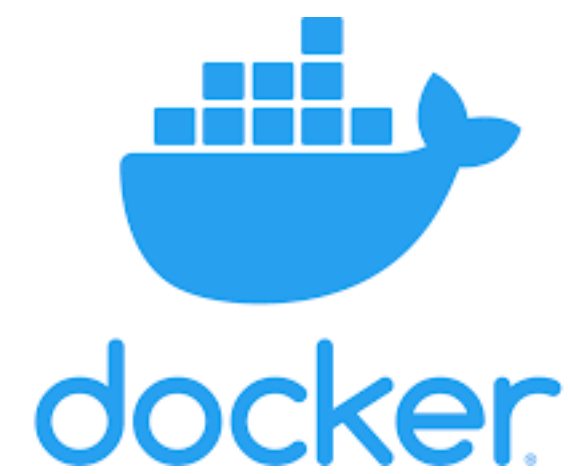
- El Big Data no puede tratarse de manera local, por las siguientes razones :
 1. Capacidad de Procesamiento de las GPUS, CPUS, Procesador o Nodos.
 2. Capacidad de Memoria
 3. Lugar de Almacenamiento
 4. Redes y estructura de la información.
- Otro factor importante es la capacidad de escalar dicha información.

Herramientas para el Big Data

Estructura

- Programación en Paralelo : Enfoque de Alto rendimiento a la hora de procesamiento de datos.
- Storage: Almacenamiento
- Distribución de sistemas: Aprovechamiento al máximo de las máquinas y capacidad de reproducir los procesos
- Alta velocidad de redes: Arquitectura de la información a alta velocidad
- Analítica : Capacidad de interpretar a los datos
- Machine Learning : Capacidad de producir resultados a un problema que se estudia con los datos.
- Visualización: Crear Valor en los datos.

Herramientas del Big Data




Hadoop

Herramientas del Big Data

- Estructura open source para el almacenamiento de datos.
- Se basa en el trabajo de clusters
- Proporciona almacenamiento masivo para cualquier tipo de datos
- Su capacidad o funcionalidad es : Procesar y analizar datos.

<https://hadoop.apache.org/>

 Apache Hadoop

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

[Learn more »](#) [Download »](#) [Getting started »](#)

Latest news

Release 3.2.2 available2021 Jan 9

This is the second stable release of Apache Hadoop 3.2 line. It contains 516 bug fixes, improvements and enhancements since 3.2.1.

Users are encouraged to read the [overview of major changes](#) since 3.2.1. For details of 516 bug fixes, improvements, and other enhancements since the previous 3.2.1 release, please check [release notes](#) and [changelog](#) detail the changes since 3.2.1.

Modules

The project includes these modules:

- **Hadoop Common:** The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™):** A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN:** A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce:** A YARN-based system for parallel processing of large data sets.
- **Hadoop Ozone:** An object store for Hadoop.

Who Uses Hadoop?

Related projects

Other Hadoop-related projects at Apache include:

- **Ambari™:** A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.
- **Avro™:** A data serialization system.
- **Cassandra™:** A scalable multi-master database

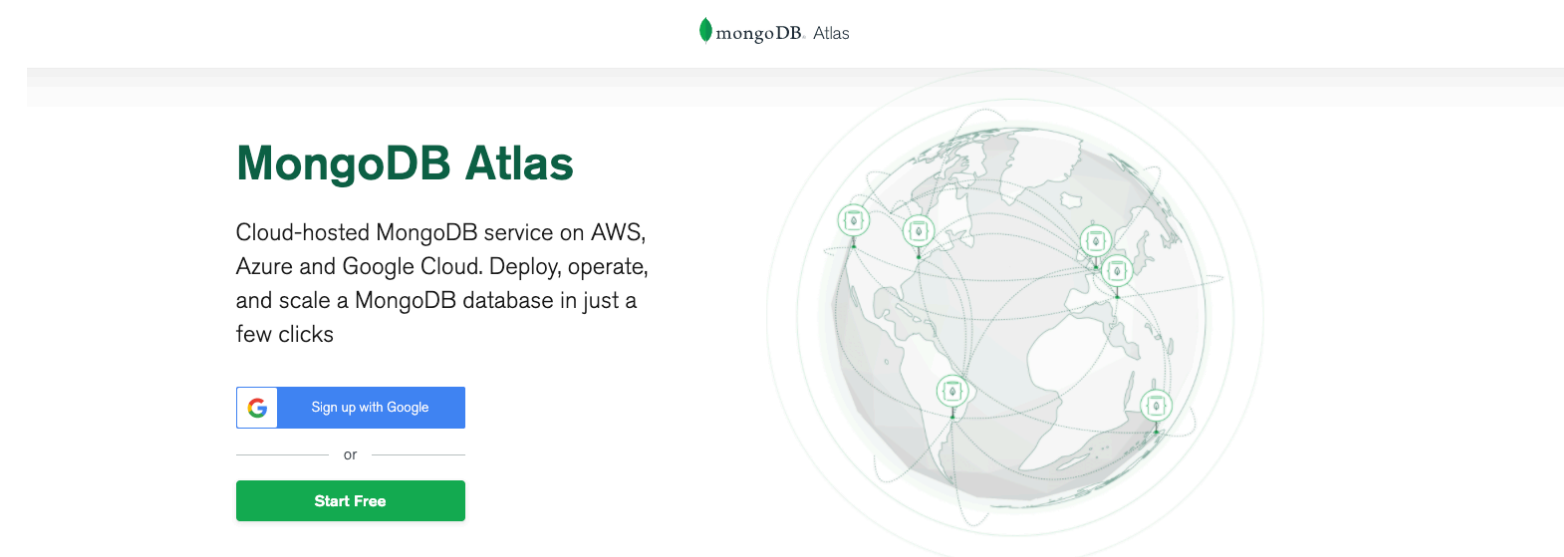
MongoDB

Herramientas del Big Data

- Motor de bases de datos NoSQL
- Gestor de bases de datos orienta a objetos y documentos JSON

[https://www.mongodb.com/cloud/atlas/lp/try2?](https://www.mongodb.com/cloud/atlas/lp/try2?utm_source=google&utm_campaign=gs_americas_colombia_search_core_brand_atlas_desktop&utm_term=mongodb&utm_medium=cpc_paid_search&utm_ad=e&utm_campaign_id=12212624317&gclid=CjwKCAiAm-2BBhANEiwAe7eyFCroMNzfro3nXgfE5az6whBYNodnyFQTpoleyOXk3viDh6GJ9VU1MBoCdhQQAvD_BwE)

[utm_source=google&utm_campaign=gs_americas_colombia_search_core_brand_atlas_desktop&utm_term=mongodb&utm_medium=cpc_paid_search&utm_ad=e&utm_campaign_id=12212624317&gclid=CjwKCAiAm-2BBhANEiwAe7eyFCroMNzfro3nXgfE5az6whBYNodnyFQTpoleyOXk3viDh6GJ9VU1MBoCdhQQAvD_BwE](https://www.mongodb.com/cloud/atlas/lp/try2?utm_source=google&utm_campaign=gs_americas_colombia_search_core_brand_atlas_desktop&utm_term=mongodb&utm_medium=cpc_paid_search&utm_ad=e&utm_campaign_id=12212624317&gclid=CjwKCAiAm-2BBhANEiwAe7eyFCroMNzfro3nXgfE5az6whBYNodnyFQTpoleyOXk3viDh6GJ9VU1MBoCdhQQAvD_BwE)



Apache Spark

Herramientas del Big Data

- Es un motor de procesamiento de datos
- Es el primer motor para hacer programación distribuida en clusters
- Permite hasta 100 veces mayor velocidad en el procesamiento.

<https://spark.apache.org/>

The screenshot shows the Apache Spark website homepage. At the top is the Apache Spark logo with the tagline "Lightning-fast unified analytics engine". Below the logo is a navigation bar with links: Download, Libraries, Documentation, Examples, Community, Developers, and Apache Software Foundation. The main content area features a headline: "Apache Spark™ is a unified analytics engine for large-scale data processing." Below this, there are three sections: "Speed", "Ease of Use", and "Built-in Libraries". The "Speed" section includes a bar chart comparing Hadoop and Spark running times for logistic regression. The "Ease of Use" section shows a code snippet for reading JSON files using the Spark Python DataFrame API. The "Built-in Libraries" section lists various libraries available with Spark. On the right side, there is a "Latest News" section with updates on Spark releases and an "APACHE EVENTS" banner with a "LEARN MORE" button. A green "Download Spark" button is located at the bottom right.

Speed
Run workloads 100x faster.

Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.

Running time (s)	Hadoop	Spark
110		
0.9		

Logistic regression in Hadoop and Spark

Ease of Use
Write applications quickly in Java, Scala, Python, R, and SQL.

```
df = spark.read.json("logs.json")
df.where("age > 21")
  .select("name.first").show()
```

Spark's Python DataFrame API
Read JSON files with automatic schema inference

Built-in Libraries:
[SQL and DataFrames](#)
[Spark Streaming](#)
[MLlib \(machine learning\)](#)
[GraphX \(graph\)](#)
[Third-Party Projects](#)

Latest News
Spark 3.0.2 released (Feb 19, 2021)
Next official release: Spark 3.1.1 (Jan 07, 2021)
Spark 2.4.7 released (Sep 12, 2020)
Spark 3.0.1 released (Sep 08, 2020)
[Archive](#)

APACHE EVENTS [LEARN MORE](#)

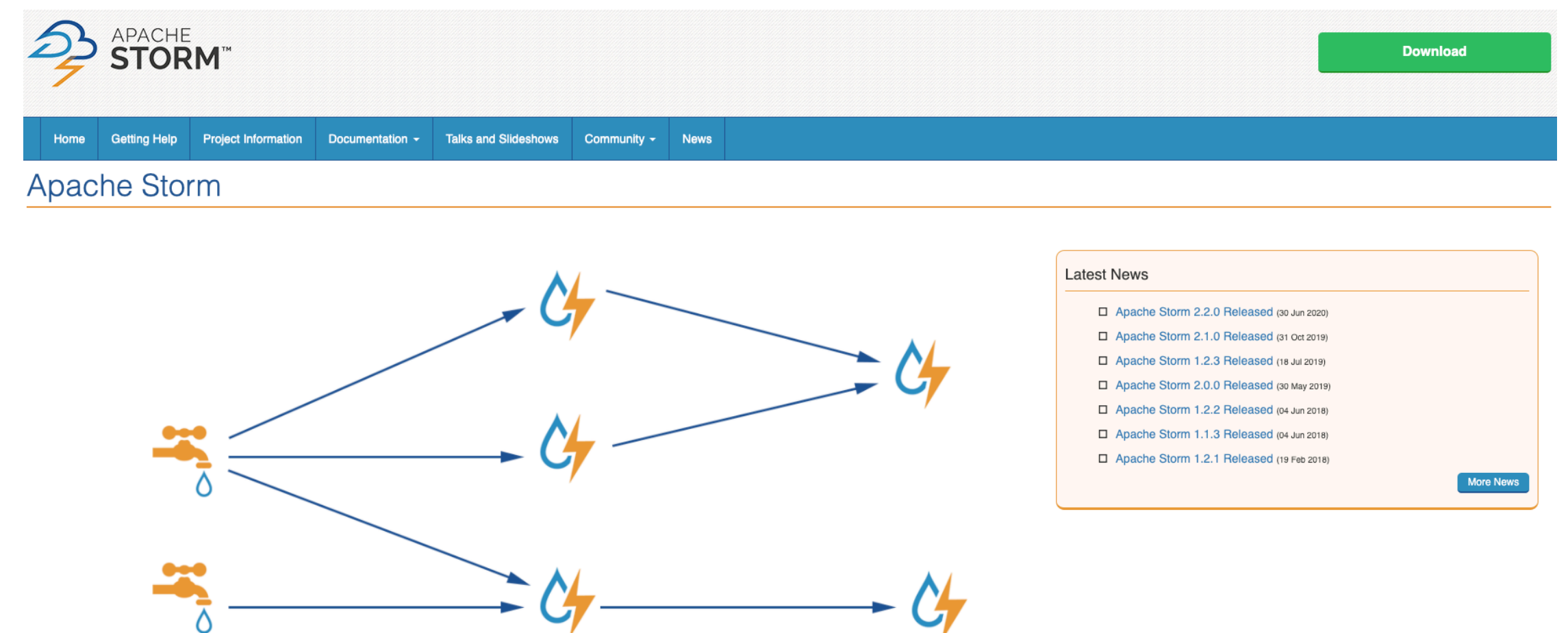
[Download Spark](#)

Apache Storm

Herramientas del Big Data

- Es un framework para el trabajo de procesamiento distribuido.
- Se basa para procesamiento real-time
- Se recomienda para el trabajo con Redes Sociales, sensores, cámaras de seguridad entre otros.

<https://storm.apache.org/>



R

Herramientas de Big Data

- Es un open source basado en la filosofía del trabajo de Analytics y ML con foco estadístico.
- Se basa en C#
- Esta optimizado para el trabajo de visualización de datos y ML.

<https://www.r-project.org/>



[\[Home\]](#)

Download

[CRAN](#)

R Project

[About R](#)

[Logo](#)

[Contributors](#)

[What's New?](#)

[Reporting Bugs](#)

[Conferences](#)

[Search](#)

[Get Involved: Mailing Lists](#)

[Developer Pages](#)

[R Blog](#)

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

News

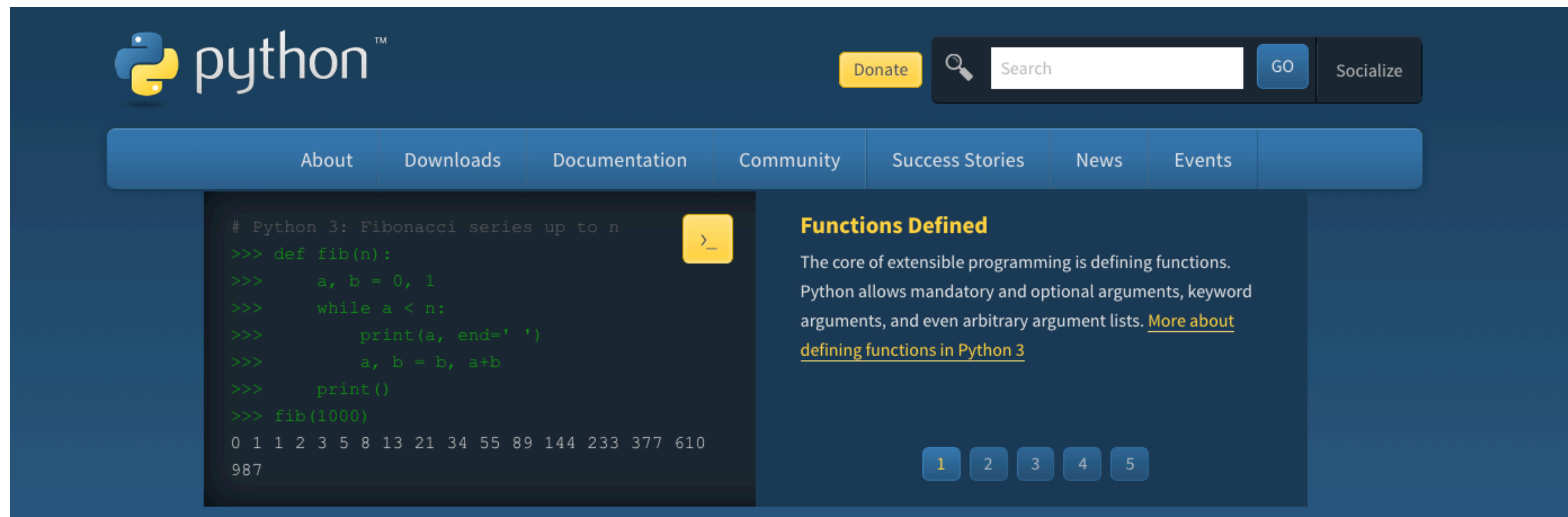
- [R version 4.0.4 \(Lost Library Book\)](#) has been released on 2021-02-15.
- Thanks to the organisers of useR! 2020 for a successful online conference. Recorded tutorials and talks from the conference are available on the [R Consortium YouTube channel](#).
- [R version 3.6.3 \(Holding the Windsock\)](#) was released on 2020-02-29.
- You can support the R Foundation with a renewable subscription as a [supporting member](#)

Python

Herramientas del Big Data

- Es un lenguaje de programación basado en OOP
- Es uno de los lenguajes más usados del mundo
- Es sencillo de integrar y fácil de implementar a nivel de industria

<https://www.python.org/>



GCP

Herramientas del Big Data

- Es la tecnología de la nube de Google.
- Se utiliza para implementar diferentes tipos de soluciones tecnológicas
- En el almacenamiento de datos provee Map reduce
- Incluye aplicaciones como AutoML.

Recomendación: <https://www.coursera.org/professional-certificates/cloud-engineering-gcp>

AWS

Herramientas del Big Data

- Es la nube de Amazon
- Es la más popular en la actualidad gracias a los servicios de tracking sobre los modelos y bases de datos.
- Es fácil de integrar a procesos de gran escala
- Incluye la función lambda.

<https://aws.amazon.com/es/free/?>

[trk=ps_a134p000003yhOBAAY&trkCampaign=acq_paid_search_brand&sc_channel=ps&sc_campaign=acquisition_LATAMO&sc_publisher=google&sc_category=core-main&sc_country=LATAMO&sc_geo=LATAM&sc_outcome=Acquisition&sc_detail=aws&sc_content=Brand_Core_aws_e&sc_matchtype=e&sc_segment=453309389434&sc_medium=ACQ-P|PS-GO|Brand|Desktop|SU|Core-Main|Core|LATAMO|EN|Text&s_kwid=AL!4422!3!453309389434!e!!g!!aws&ef_id=CjwKCAiAm-2BBhANEiwAe7eyFLGzcsIqVp3OHKP7F4g8fkrxUXVJHr1OtcIom_kRTxViTExM8DOLxoCCOwQAvD_BwE:G:s&s_kwid=AL!4422!3!453309389434!e!!g!!aws&all-free-tier.sort-by=item.additionalFields.SortRank&all-free-tier.sort-order=asc](https://aws.amazon.com/es/free/?trk=ps_a134p000003yhOBAAY&trkCampaign=acq_paid_search_brand&sc_channel=ps&sc_campaign=acquisition_LATAMO&sc_publisher=google&sc_category=core-main&sc_country=LATAMO&sc_geo=LATAM&sc_outcome=Acquisition&sc_detail=aws&sc_content=Brand_Core_aws_e&sc_matchtype=e&sc_segment=453309389434&sc_medium=ACQ-P|PS-GO|Brand|Desktop|SU|Core-Main|Core|LATAMO|EN|Text&s_kwid=AL!4422!3!453309389434!e!!g!!aws&ef_id=CjwKCAiAm-2BBhANEiwAe7eyFLGzcsIqVp3OHKP7F4g8fkrxUXVJHr1OtcIom_kRTxViTExM8DOLxoCCOwQAvD_BwE:G:s&s_kwid=AL!4422!3!453309389434!e!!g!!aws&all-free-tier.sort-by=item.additionalFields.SortRank&all-free-tier.sort-order=asc)



Tipos de ofertas

Explore los más de 85 productos y comience a crear en AWS mediante la capa gratuita. Hay tres tipos diferentes de ofertas gratuitas disponibles según el producto utilizado. Observe a continuación los detalles de cada producto.



Gratis para siempre

Estas ofertas de la capa gratuita no caducan y están disponibles para todos los clientes de AWS



12 meses de uso gratuito

Disfrute de estas ofertas durante 12 meses después de su fecha de registro inicial en AWS



Pruebas

Las ofertas de prueba gratuita a corto plazo se inician a partir de la fecha en la que se activa un servicio en particular

Diferencia entre Big Data, Data Science, Machine Learning y AI



David Robinson

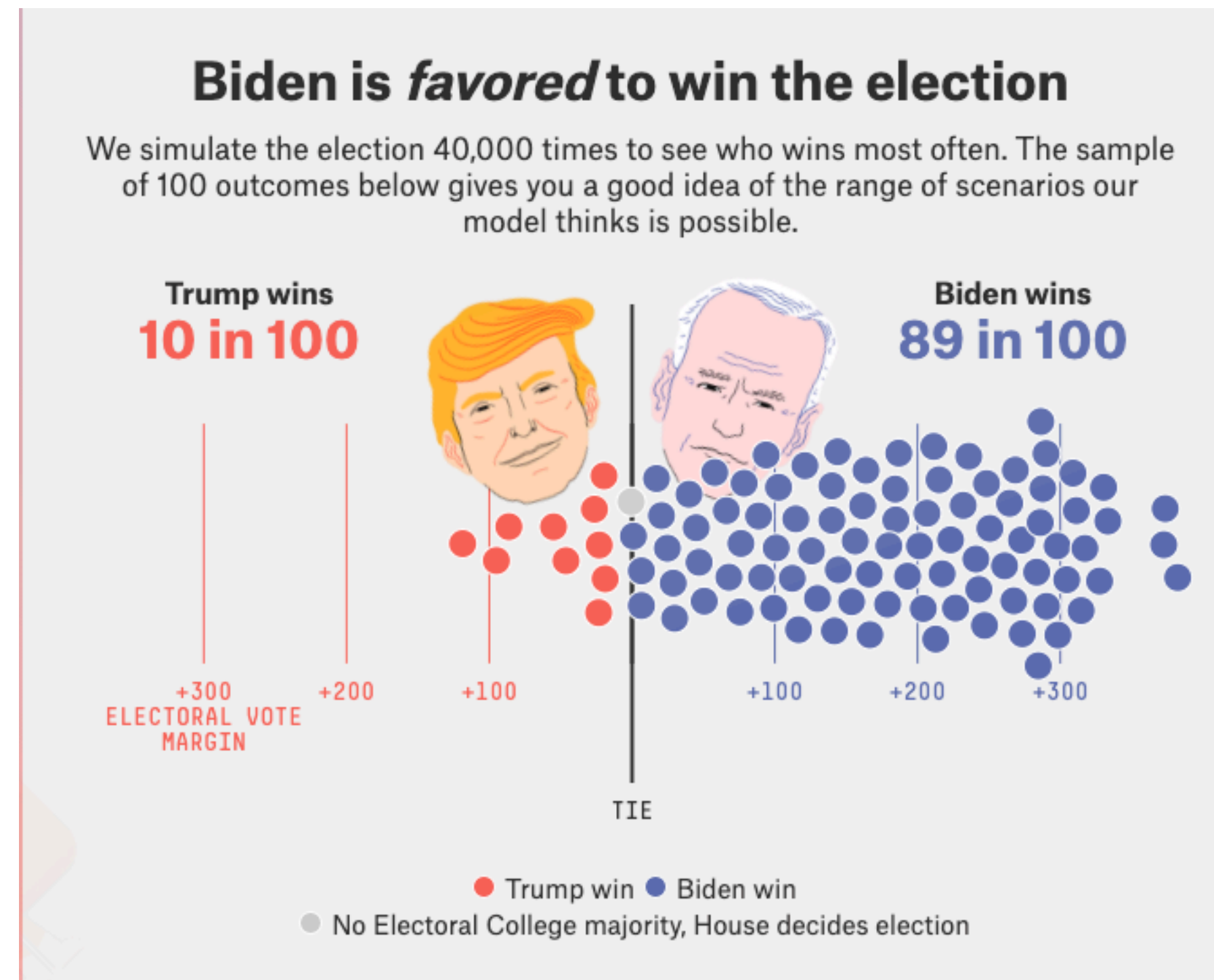
*Principal Data Scientist at
Heap, works in R and
Python.*

- **Data science** produces **insights**
- **Machine learning** produces **predictions**
- **Artificial intelligence** produces **actions**

<https://es.mailjet.com/blog/news/big-data/>

Casos de usos


Predicción de las elecciones



Detección de Fake News

MACHINE BOX

machinebox.io



Fakebox

Analyze news content and detect fake news

● ready

localhost:8003

fakebox v1da3e61c

Community support

Open an issue

@machineboxio

machinebox.io

hello@machinebox.io

Do you ❤️ this box?

We rely on developers like you spreading the word about us, so please [tell your friends and followers](#)

Fakebox uses language models from [spaCy.io](#) licensed under [Creative Commons Attribution 3.0 License](#)

© 2017 [Machine Box Ltd](#)
All rights reserved

Fakebox lets you analyze news content and detect fake news.

/fakebox/check endpoint

Use the `/fakebox/check` endpoint to analyze a news article.

Fakebox takes the title, content and URL of a news article and analyzes each of them. The web domain is checked against a database of known websites, and the title and content are judged to be either impartial or biased. If you know the domain but not the entire URL, it is enough to send the URL value as the homepage.

Fakebox will pull out any recognized [entities](#) (people, dates, locations, etc.) mentioned in the title and content, as well as significant keywords from the body of the article.

Article analyzer

On this page

Analyzing news articles

Entities

Domain categories

Environment variables

JSON

HTTP POST form

To analyze a news article, make the following `POST` request with a JSON body:

```
POST http://localhost:8003/fakebox/check
{
  "url": "http://www.bbc.co.uk/news/world-asia-40909468",
  "title": "North Korea: China urges Trump not to worsen situation",
  "content": "China's President Xi Jinping has urged Donald Trump..."
}
```

Pass in at least one of the required fields:

- `title` - (string) The title of the article
- `content` - (string) The entire content of the article
- `url` - (string) The URL of where the article is hosted

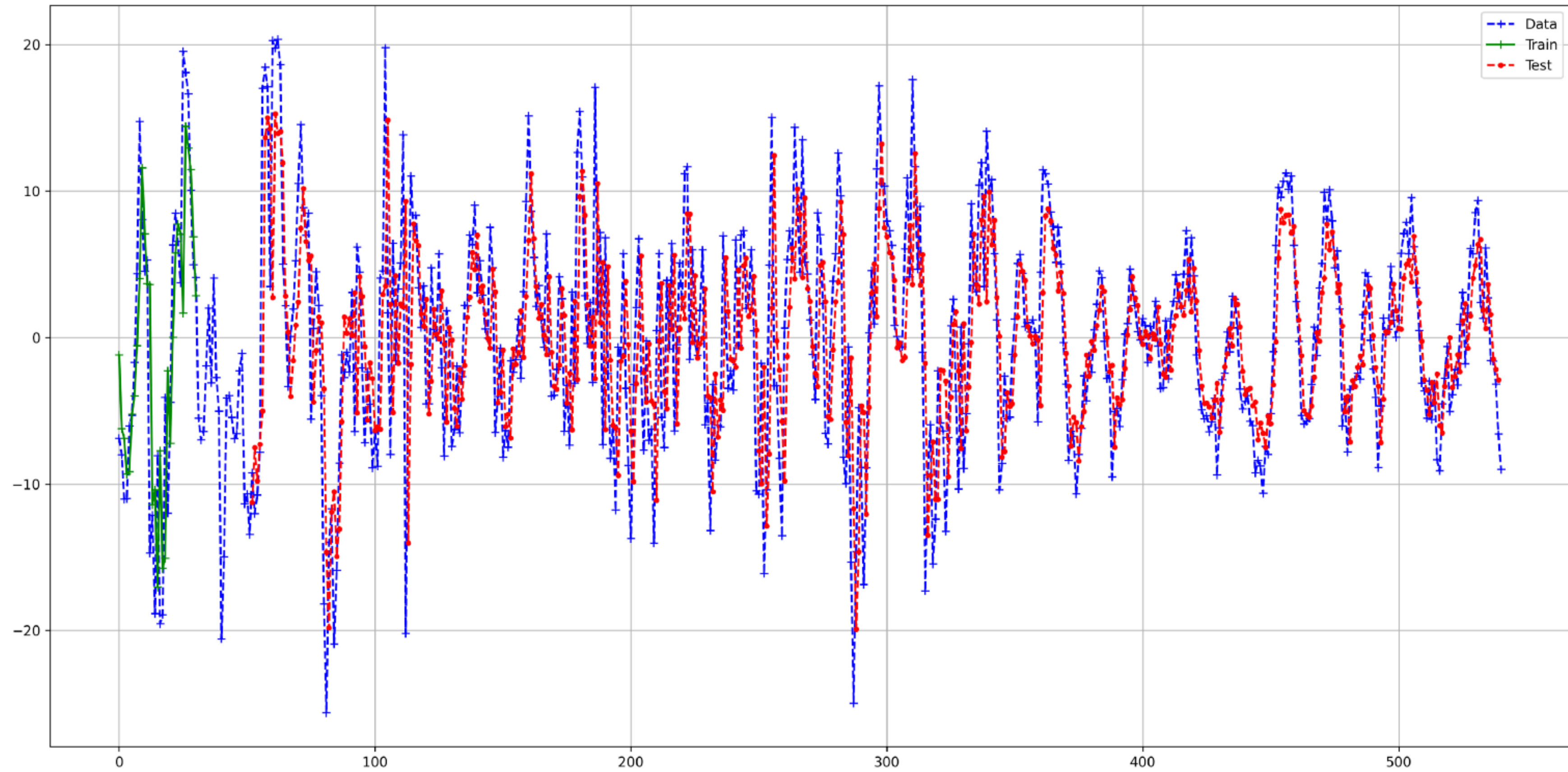
Remember also to:

- Set both the `Accept` and `Content-Type` headers to `"application/json; charset=utf-8"`

Try it now

cURL

Forecasting



Reflexión Final

