BIG DATA APLICADO

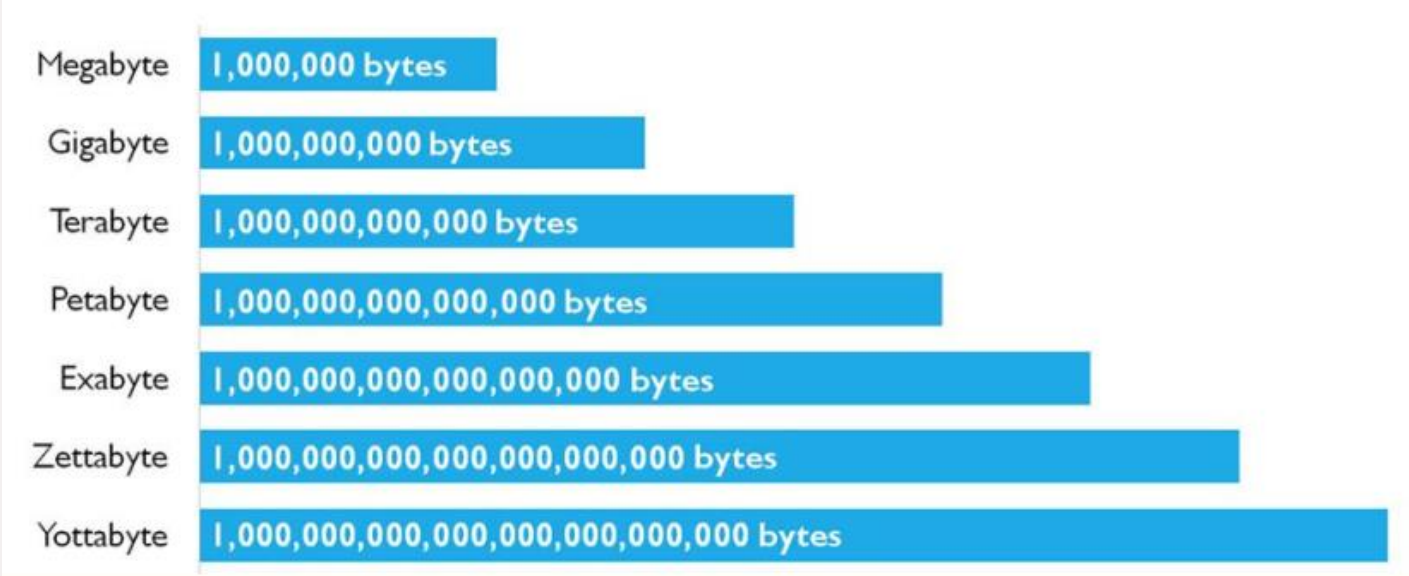
David Granados Zafra

Curso de Big Data Aplicado 2022-2023 v0.2

INTRODUCCIÓN

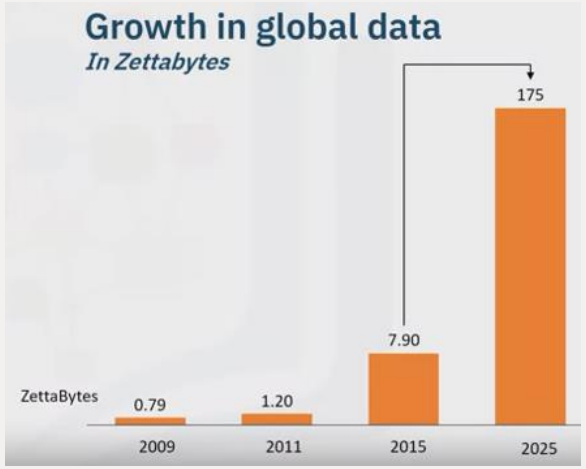
¿QUÉ ES BIG DATA?

- No existe una definición oficialmente aceptada
- El conjunto de herramientas y estretegias que permiten adquirir, almacenar y explotar una cantidad de información tan grande, que no podemos hacerlo con las técnicas o herramientas tradicionales.



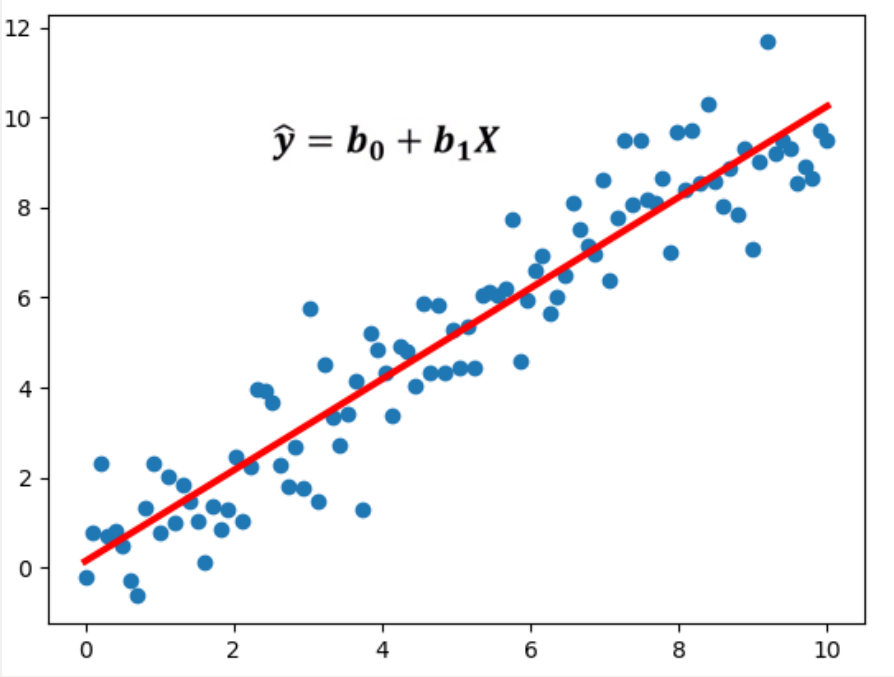
¿QUÉ ES BIG DATA?

- Una película HD en 1080 que ocupase un Zettabyte, duraría 36 millones de años
- El total del video en streaming ya supera los 4 zettabytes
- En los dos ultimos años se ha creado más datos que en toda la humanidad anterior.



¿PORQUÉ NECESITAMOS TANTOS DATOS?

- Desde pequeños hemos aprendido que si tenemos una función podemos generar infinitos valores para la misma, pero qué ocurre cuando no tenemos la función y tenemos que obtener una aproximación a la misma mediante datos?



EJEMPLO: NETFLIX

- Sus búsquedas
- Dispositivos usados
- Cuántos días a la semana consumen productos en las plataformas
- Qué días de la semana ven más programas
- Cuántas horas invierten en visionado
- Qué fragmentos de una película ven más
- Información de sus perfiles en redes sociales
- Si ven episodios completos o solo fragmentos
- Valoraciones del contenido
- Intereses compartidos

Para promover la exitosa serie "*House of Cards*", crearon más de diez versiones de tráilers. Después, siguiendo las normas del micrortargeting, ofrecían a cada usuario un tráiler adaptado a sus preferencias.

COMPARATIVA

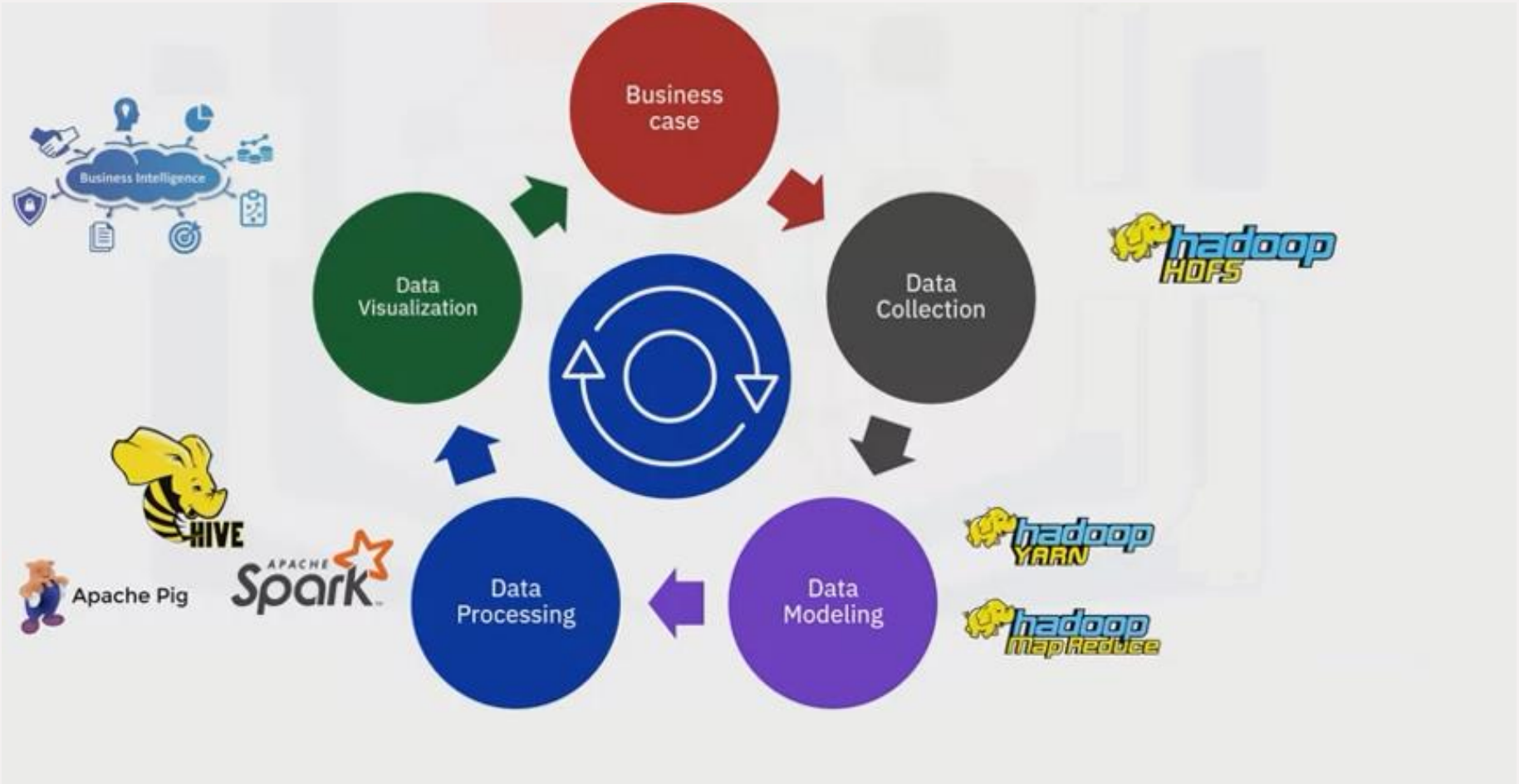
Small Data

- Small enough for human inference
- Accumulated slowly
- Relatively consistent and structured data usually stored in known forms such as JSON and XML
- Mostly located in storage systems within Enterprises or data centers

Big Data

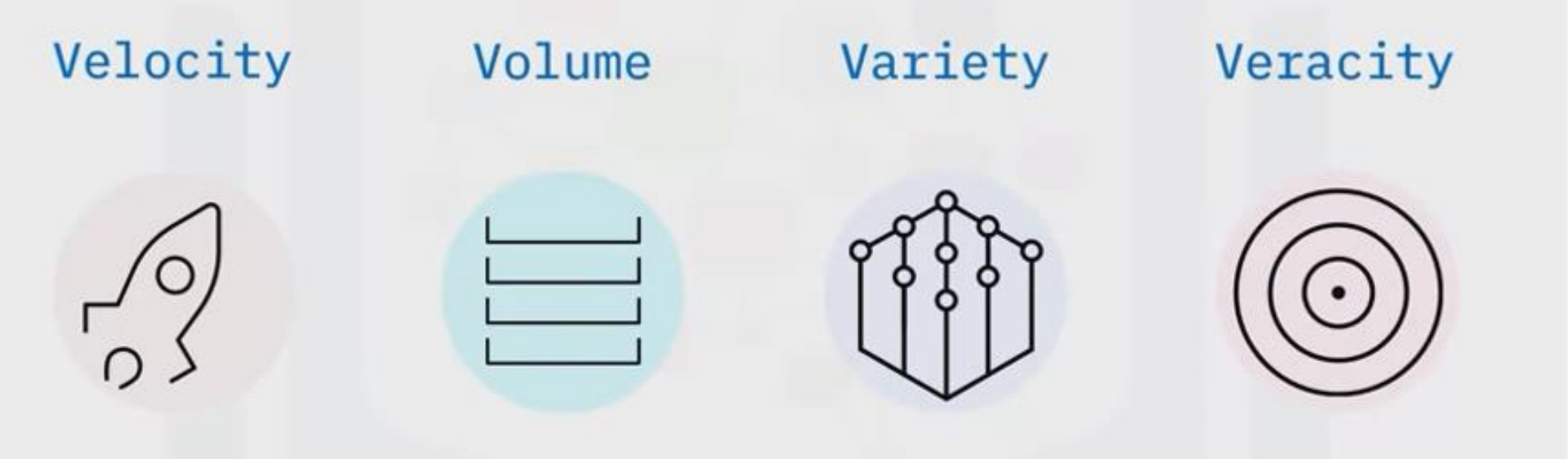
- Data generated in huge volumes and could be structured, semi-structured, or unstructured
- Needs processing to generate insights for human consumption
- Arrives continuously at enormous speed from multiple sources
- Comprises any form of data including video, photos, and more
- Distributed on the cloud and server farms

CICLO DE VIDA



LAS V'S DEL BIG DATA

V'S DE BIG DATA



VELOCIDAD

Description

- Data that is generated fast
- Process that never stops

Attributes

- Batch
- Close to real time
- Streaming

Drivers

- Improved connectivity and hardware
- Rapid response times
- Precalculated analysis

VOLUMEN

Description

- Scale of data
- Increased amount of stored data

Attributes

- Petabytes
- Exa
- Zetta

Drivers

- Increase in data sources
- Higher resolution sensors
- Scalable infrastructure

VARIEDAD

Description

- Data that comes from machines, people, and processes
- Structured, semi-structured, and unstructured data

Attributes

- Structure, complexity, and origin

Drivers

- Mobile technologies
- Scalable infrastructure
- Resilience
- Fault recovery
- Efficient storage and retrieval

RESILENCIA

La **resiliencia informática** es la capacidad de un sistema para recuperarse de un fallo y conservar la confiabilidad del servicio cuando este las presenta, su objetivo es **asegurar que todas las operaciones comerciales estén protegidas, para que así una amenaza o incumplimiento no afecte todo el negocio**.

VERACIDAD

Description

- Quality, origin, and conformity of facts
- Accuracy of data
- Data that comes from people and processes

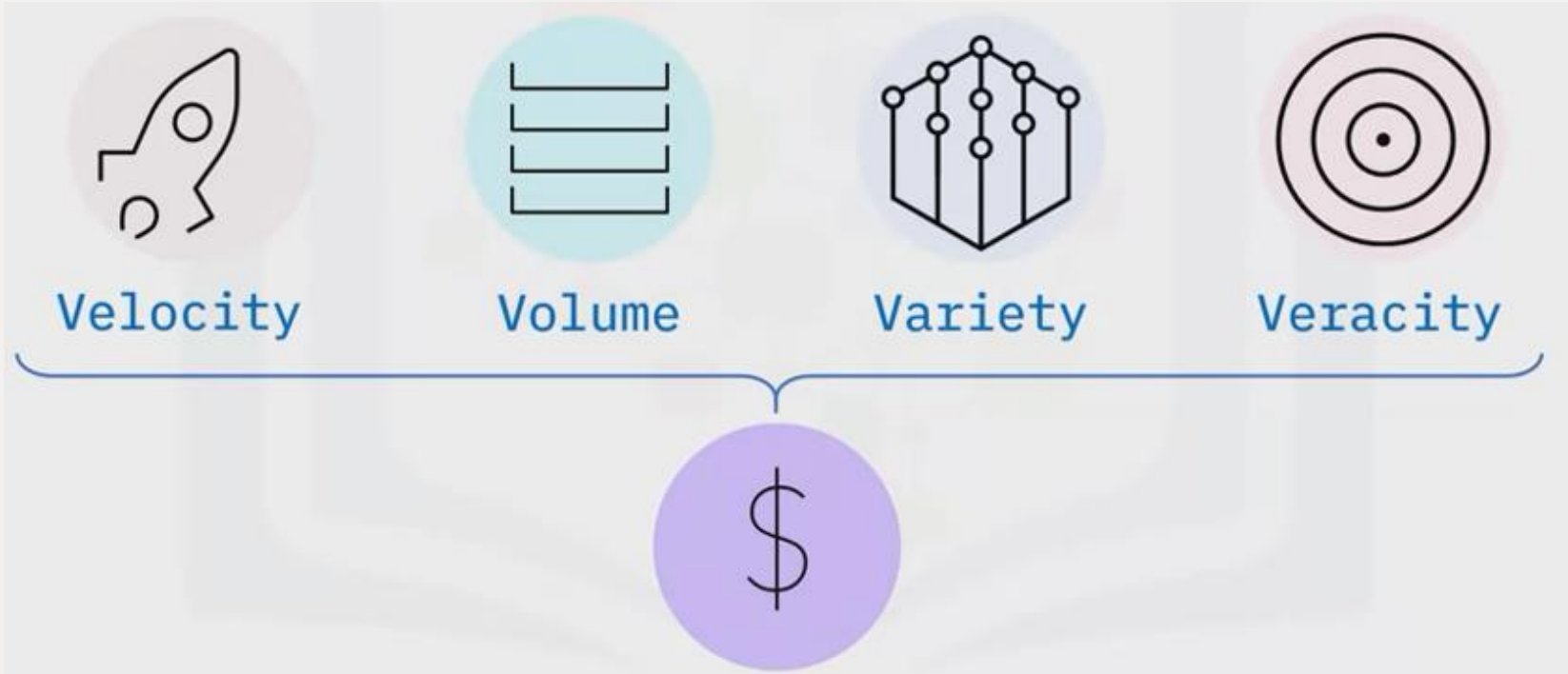
Attributes

- Consistency and completeness
- Integrity
- Ambiguity

Drivers

- Cost and traceability
- Robust ingestion
- ETL mechanisms

LA QUINTA V: VALOR



PARADIGMAS DE COMPUTACIÓN

BATCH PROCESSING

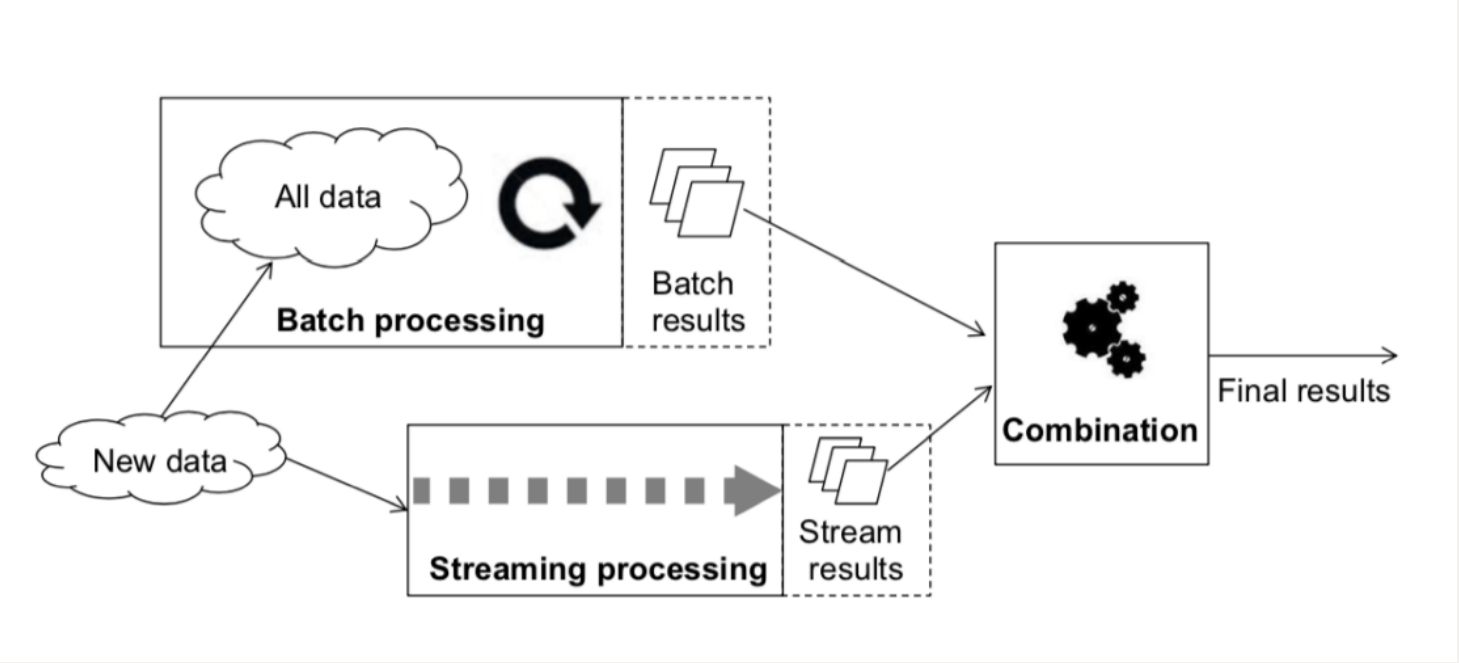
- Scalable
- Distributed
- Parallel
- Fault Tolerant
- Large volumes of static data
- High Latence

STREAMING PROCESSING

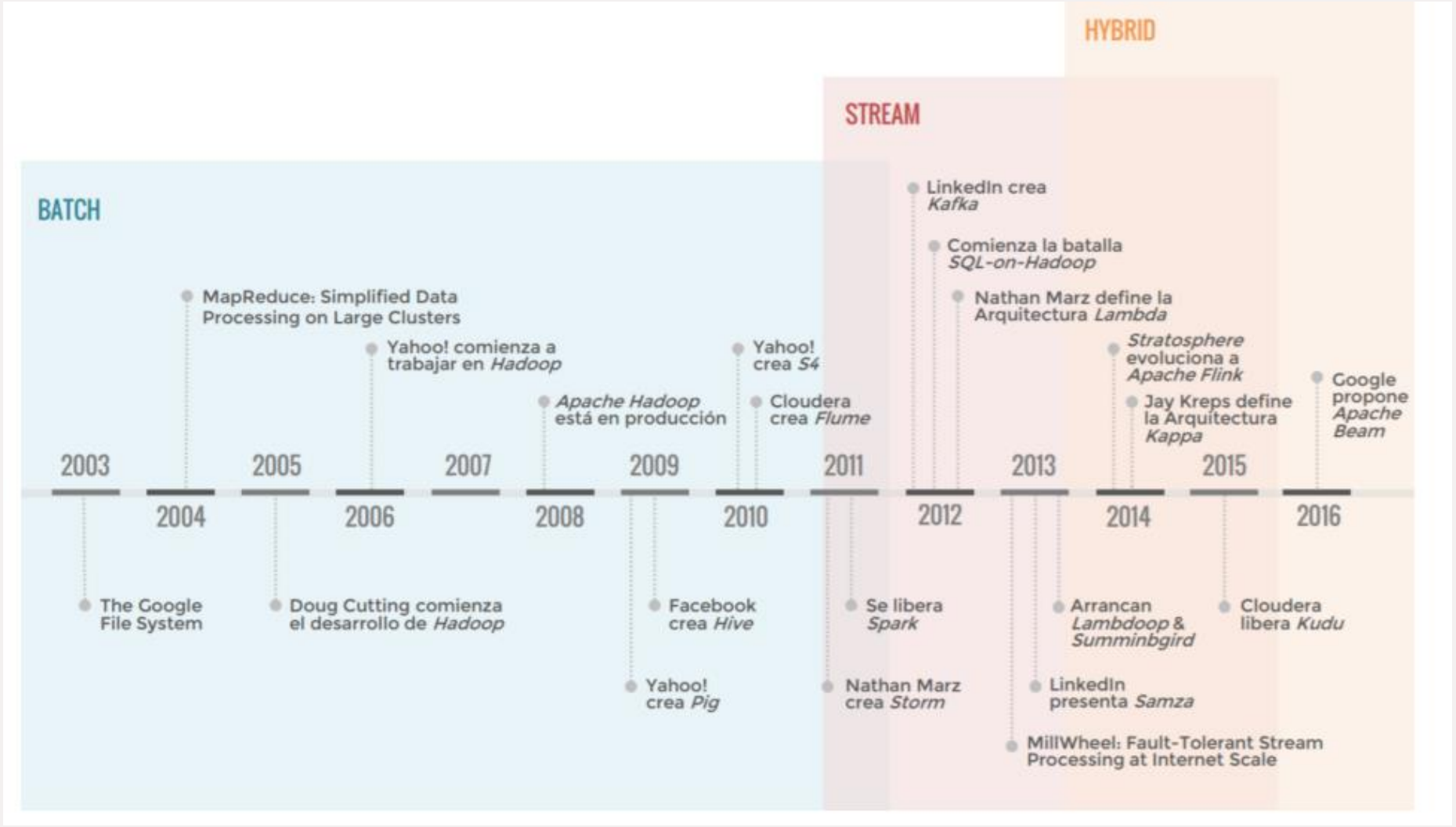
- Low latency
- Distributed
- Parallel
- Fault Tolerant
- Information generated countinously

HYBRID PROCESSING

- Low latency
- Scalable
- Batch and Streaming results combined



HYBRID PROCESSING



EJEMPLOS EN LA VIDA REAL.

EJEMPLOS. SISTEMA DE RECOMENDACIONES

- Búsquedas de productos
- Pedidos pasados
- Items que están en la cesta de la compra.
- Opiniones y puntuaciones
- Lo que han comprado o mirado otros clientes

ASISTENTES VIRTUALES

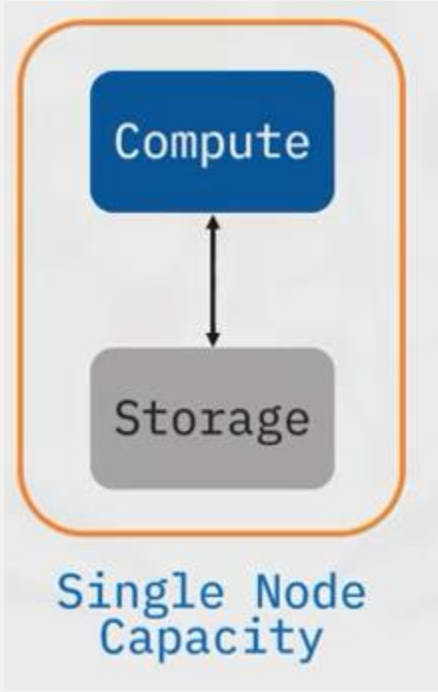
- Elaborar Respuestas
  - Usan redes neuronales
- Traducir de un lenguaje a otro
- Detectan dónde estás o qué haces e intentan predecir el futuro. Por ejemplo si la gente que sale de trabajar, te puede decir porqué ruta vas a llegar antes a casa.

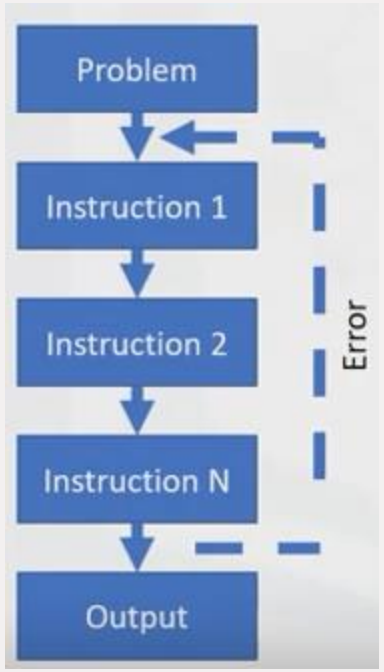
IOT

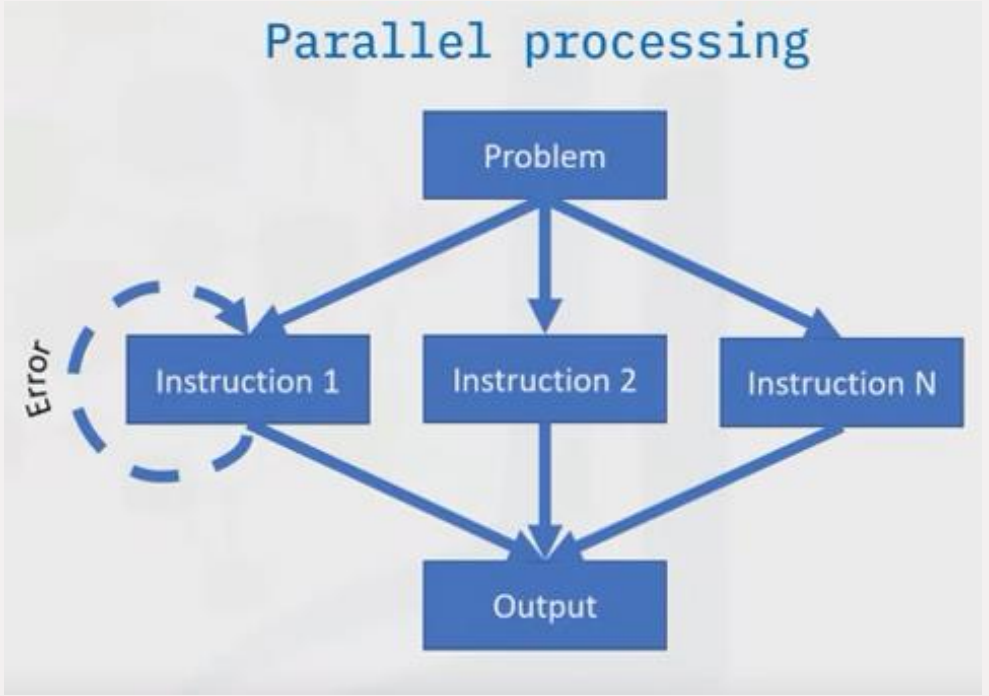

Thing
or device
Wireless
network
Data
Gateway
Data
Cloud
Data
Analytics
Data
User interface

PROCESAMIENTO PARALELO

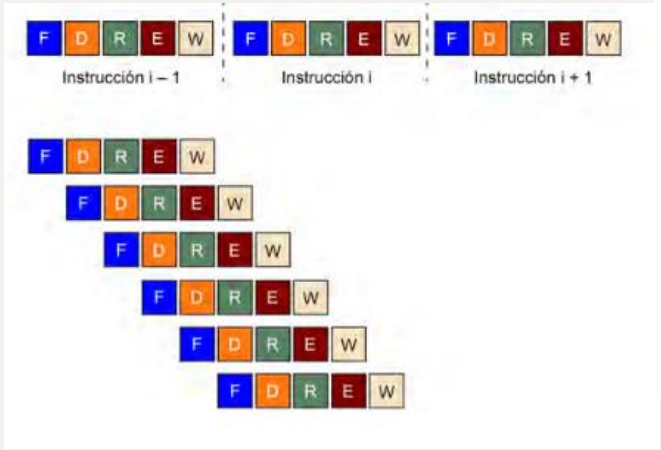
PROCESAMIENTO LINEAL VS PARALELO

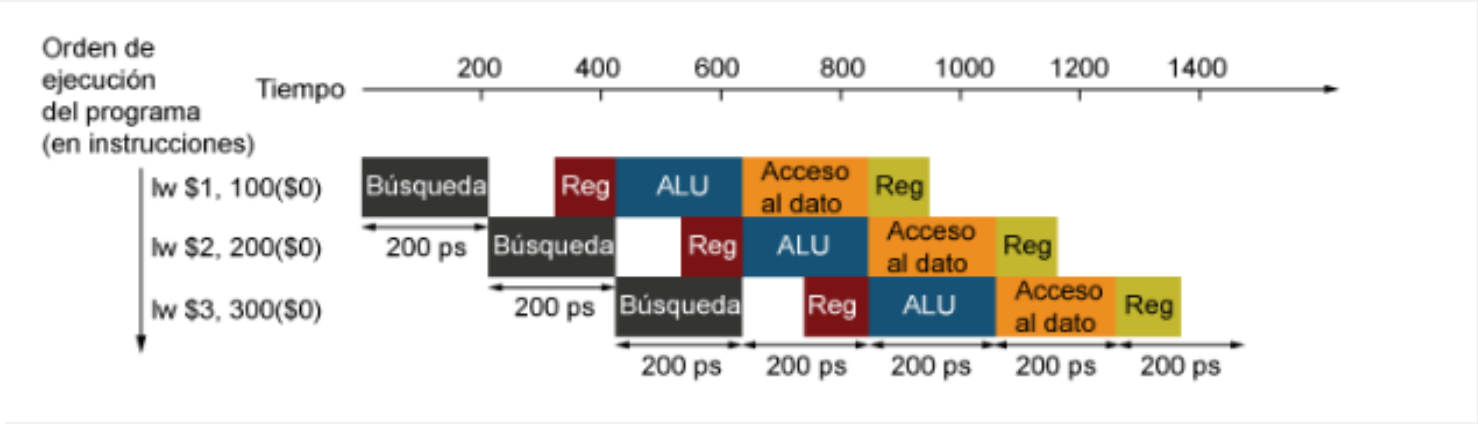




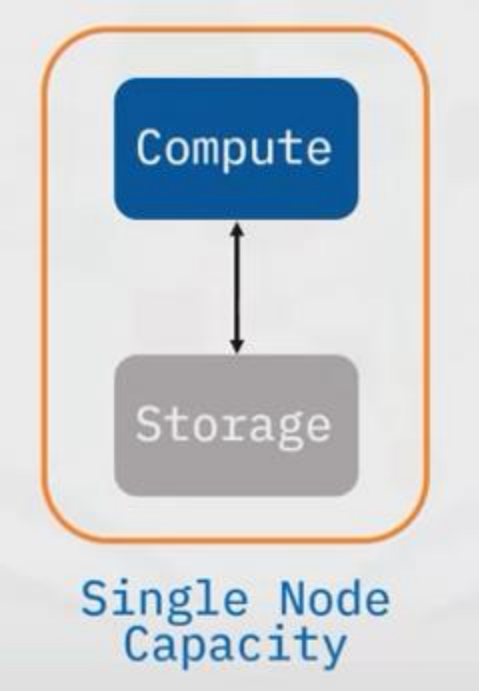


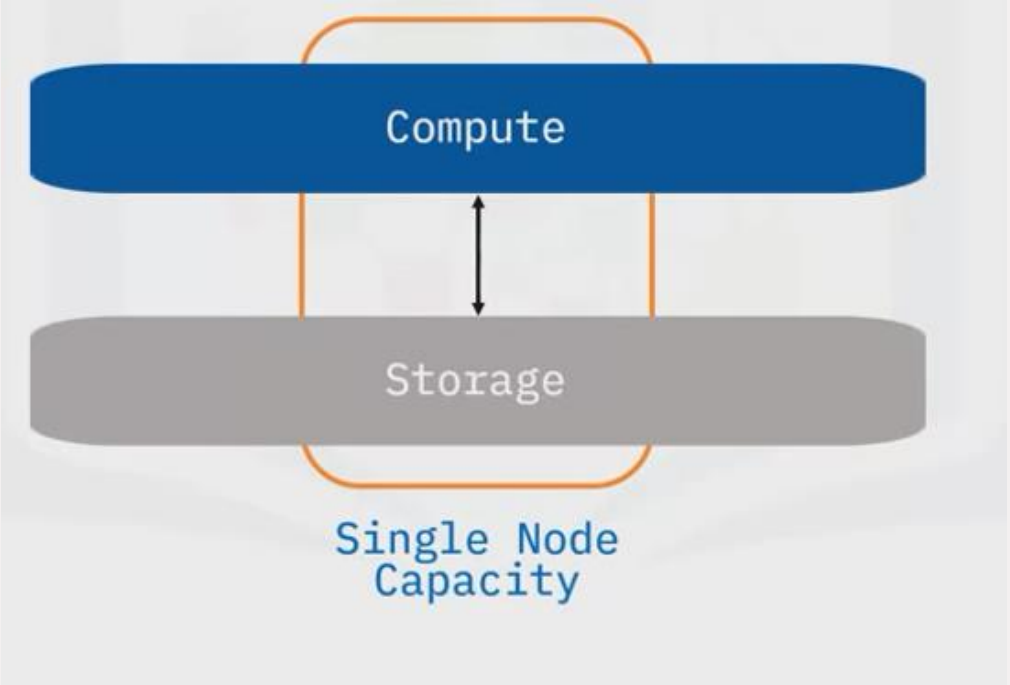
PROCESAMIENTO LINEAL VS PARALELO



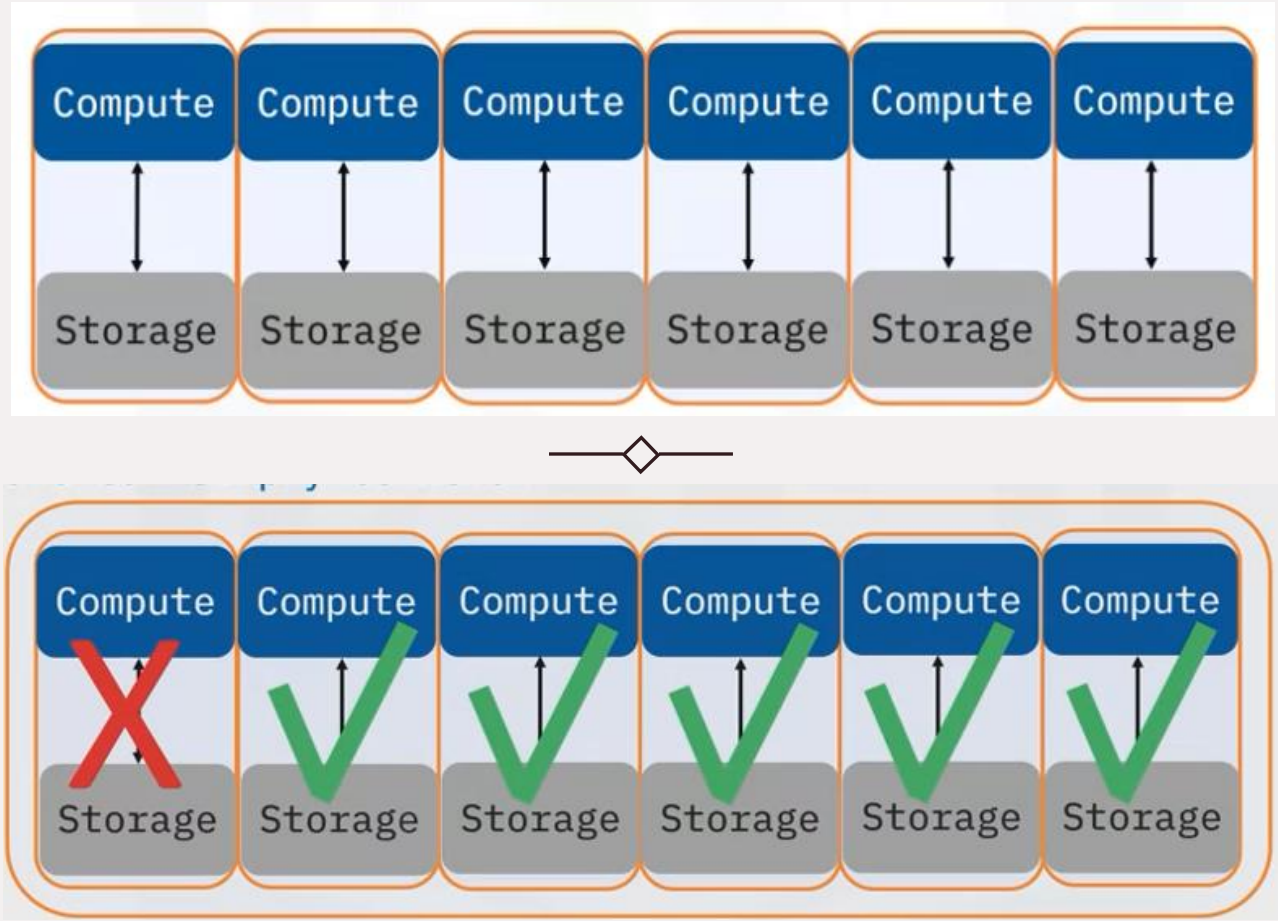


ESCALADO DE DATOS

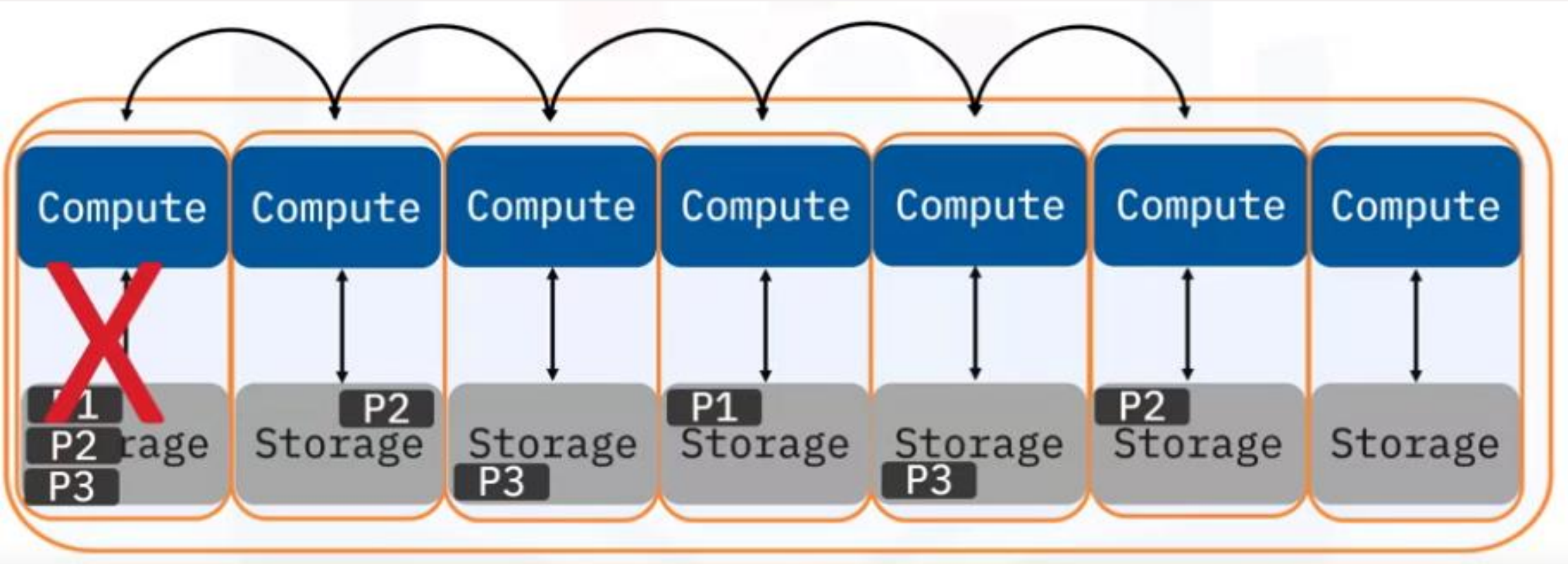




ESCALADO HORIZONTAL



TOLERANCIA A FALLO. MODELO HADOOP



ECOSISTEMA BIG DATA


Data Technologies
Programming Tools
Analytics & Visualization
Big Data Ecosystem
NoSQL Databases
Business Intelligence
Cloud Providers

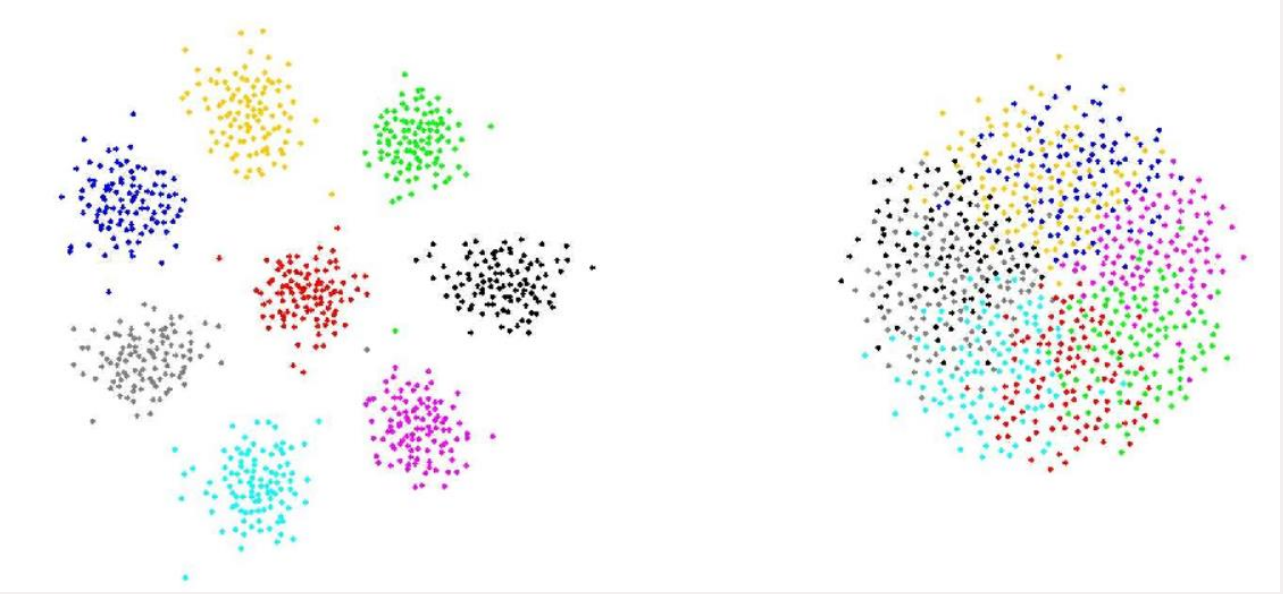
DATA TECHNOLOGIES

- Permite a las empresas:
  - Capturar, procesar y compartir datos a escala en cualquier formato
  - Trabajar con datos estructurados y desestructurados
  - Aprovechar high-performance y procesamiento paralelo
- Herramientas
  - Hadoop
  - HDFS
  - SPARK

ANÁLISIS Y VISUALIZACIÓN

- La análitica examina grandes cantidades de datos:
  - Detectar patrones, correlaciones y tendencias

- Visualización, por otro lado nos permite representar esos datos de forma gráfica.



- Tableau
- Pentaho
- SAS
- Teradata

BUSINESS INTELLIGENCE

- Herramientas que de forma "sencilla" permiten obtener determinados "indicadores"
- Estos resultados ayudan a los analistas a tomar decisiones estratégicas.
- Herramientas:
  - PowerBi
  - Cognos
  - Oracle

CLOUD PROVIDERS

- Ofrecen la infraestructura y los servicios necesarios para poder realizar cualquier fase del ciclo de big data.

CLOUD PROVIDERS

Basis Of	IAAS	PAAS	SAAS
Stands for	Infrastructure as a service.	Platform as a services.	Software as a services.
Uses	IAAS is used by network architects.	PAAS is used by developers.	SAAS is used by end user.
Access	IAAS gives access to the resources like virtual machines and virtual storage.	PAAS give access to run time environment to deployment and development tools for application.	SAAS give access to the end user.
Model	It is a service model that provides visualized computing resources over the internet.	It is a cloud computing model that delivers tools that are used for development of application.	It is a service model in cloud computing that host software make available for clients.
Technical understanding.	It requires technical knowledge.	In this some knowledge is required for the basic setup.	There is no requirement about technicalities company handles everything.
Popularity.	It is popular between developer and researchers.	It popular between developer who focus on the development of apps and scripts.	It is popular between consumer and company, such as file sharing, email and networking.
Cloud services.	Amazon Web Services, sun, vCloud Express.	Facebook, and Google search engine.	MS Office web, Facebook and Google Apps.
Enterprise services.	AWS virtual private cloud.	Microsoft azure.	IBM cloud analysis.
Outsourced cloud services.	Salesforce	Force.com, Gigaspaces.	AWS, Terremark
User Controls	Operating System, Runtime, Middleware, and Application data	Data of the application	Nothing

NOSQL

- Rompen con lo que representan las bases de datos transaccionales.
  - Por ejemplo:
    - No tienen en cuenta la consistencia de datos
    - Datos redundantes
    - Entidades con distintos atributos
- Herramientas
  - MongoDB
  - HBase
  - Redis

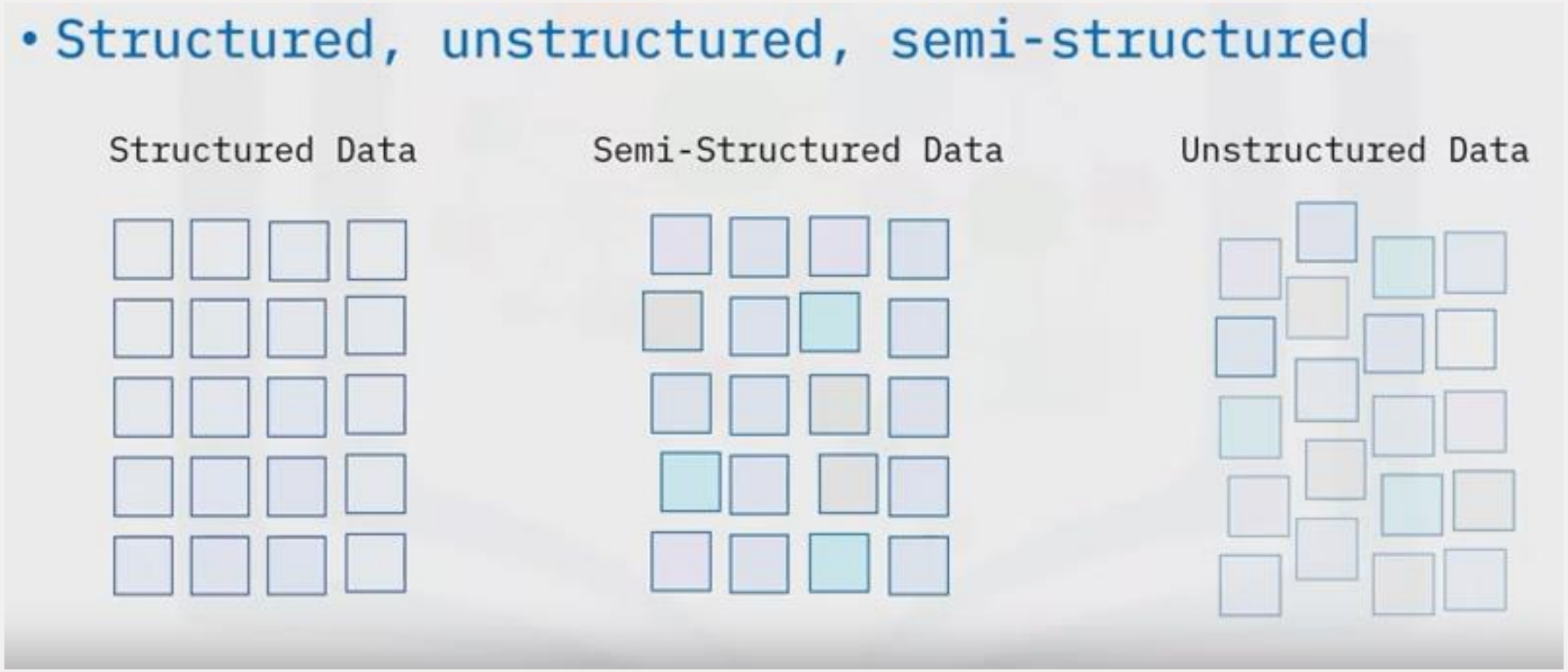
HERRAMIENTAS DE PROGRAMACIÓN

- Preparadas para trabajar con cualquier fase del ciclo de big data.
- Herramientas
  - R
  - Python
  - Scala

TIPOS DE BIG DATA

TIPOS

- Structured, unstructured, semi-structured



FUENTES DE DATOS

- Estructurado
  - Bases de datos estructurados
  - Excel
- Semiestructurado
  - Xml
  - JSON
- Desestructurado
  - La que se genera por Internet
  - Las empresas que producen contenido multimedia.
  - Las redes sociales.

DESESTRUCTURADO

- PRODUCCIÓN DE VIDEO
- SOCIAL MEDIA
- INTERNET

FIN