Big Data, fundamentos y tópicos avanzados

PARTE DE LA FORMACIÓN BIG DATA ENGINEER DE BIG DATA ACADEMY

Concepto

Big Data

Es un marco de trabajo (conceptos + tecnologías) que permite **procesar grandes volúmenes** de datos, de **diferentes estructuras o con carencia** de estas, que **pueden variar** en el tiempo, a **grandes velocidades** que generen **valor al negocio**.

Objetivo fundamental

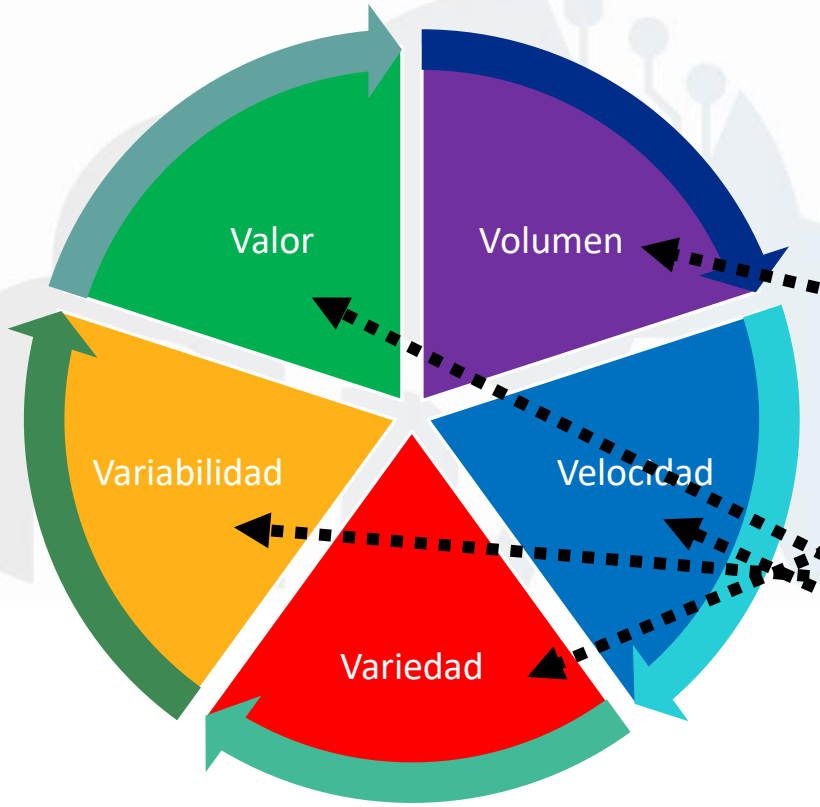
Objetivo fundamental del Big Data

Aumentar el valor del proceso por medio de:

1. **Reducir los tiempos** de procesamiento
2. **Integrar todas las fuentes** de datos disponibles
3. **Reducir los costos** de hardware
4. **Reducir el uso de recursos** computacionales
5. **Crecer fácilmente** en potencia computacional
6. **Aumentar la exactitud** en los cálculos
7. **Potenciar otras tecnologías** y marcos de trabajo

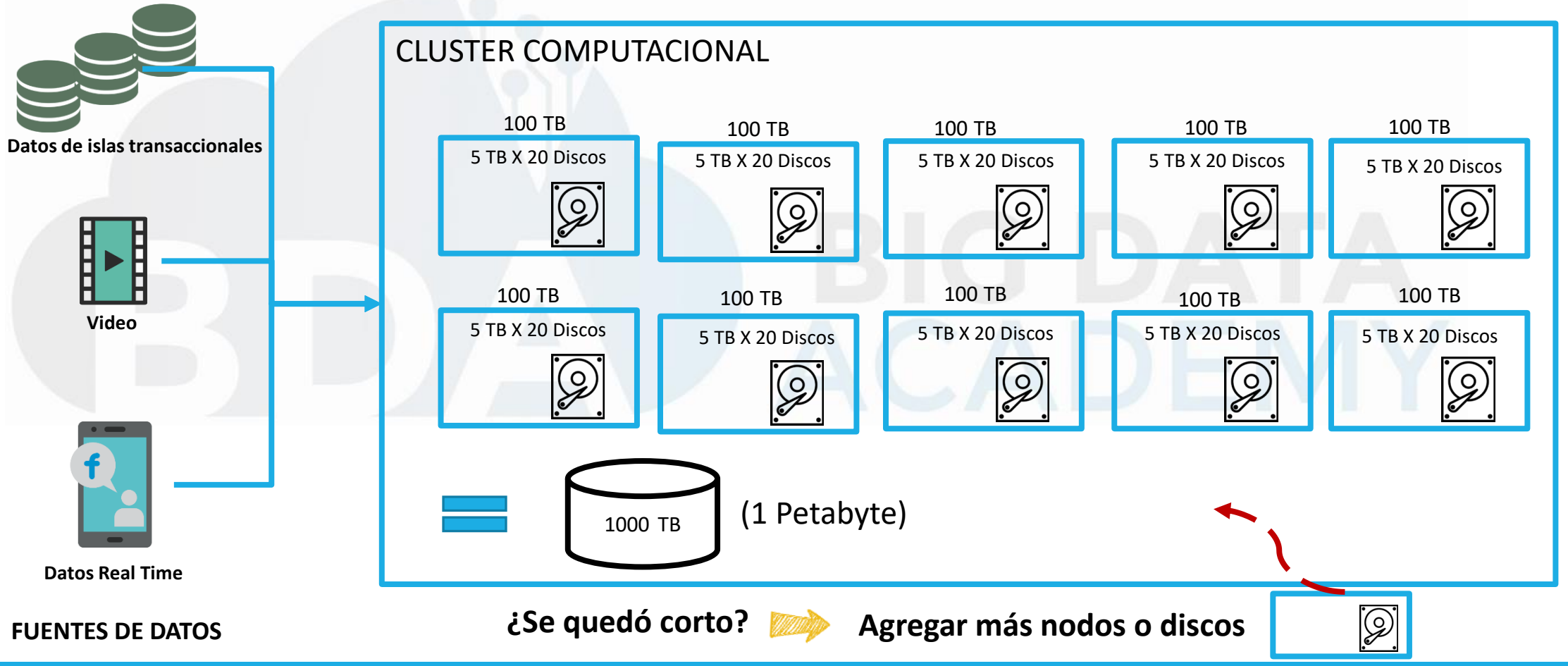
Big Data: Naturaleza

La filosofía de Big Data: Las 5V

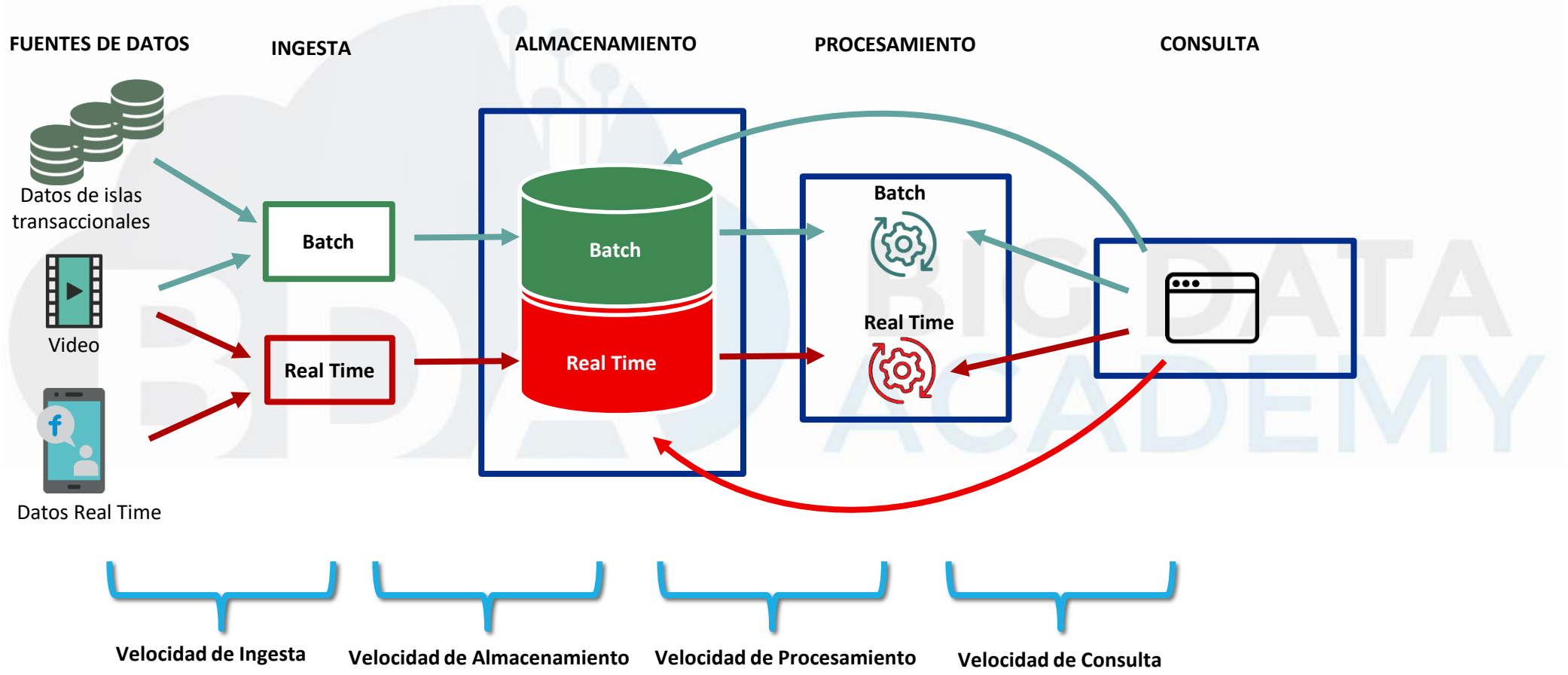


Es un marco de trabajo (conceptos + tecnologías) que permite procesar **grandes volúmenes** de datos, de **diferentes estructuras o con carencia** de estas, que pueden **variar en el tiempo**, a **grandes velocidades** y que generen **valor al negocio**.

La filosofía de Big Data: Volumen



La filosofía de Big Data: Velocidad



La filosofía de Big Data: Variedad

La variedad hace referencia al tipo de estructura del dato

FUENTES DE DATOS

Datos de islas transaccionales

DATA ESTRUCTURADA
Misma estructura para todos los registros

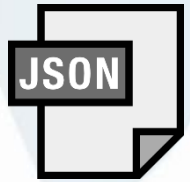
Tablas

CSV

XML

Datos Real Time

DATA SEMI-ESTRUCTURADA
Cada registro tiene su propia estructura



JSON

XML

Video

DATA NO ESTRUCTURADA
No tiene estructura ni registro

Videos

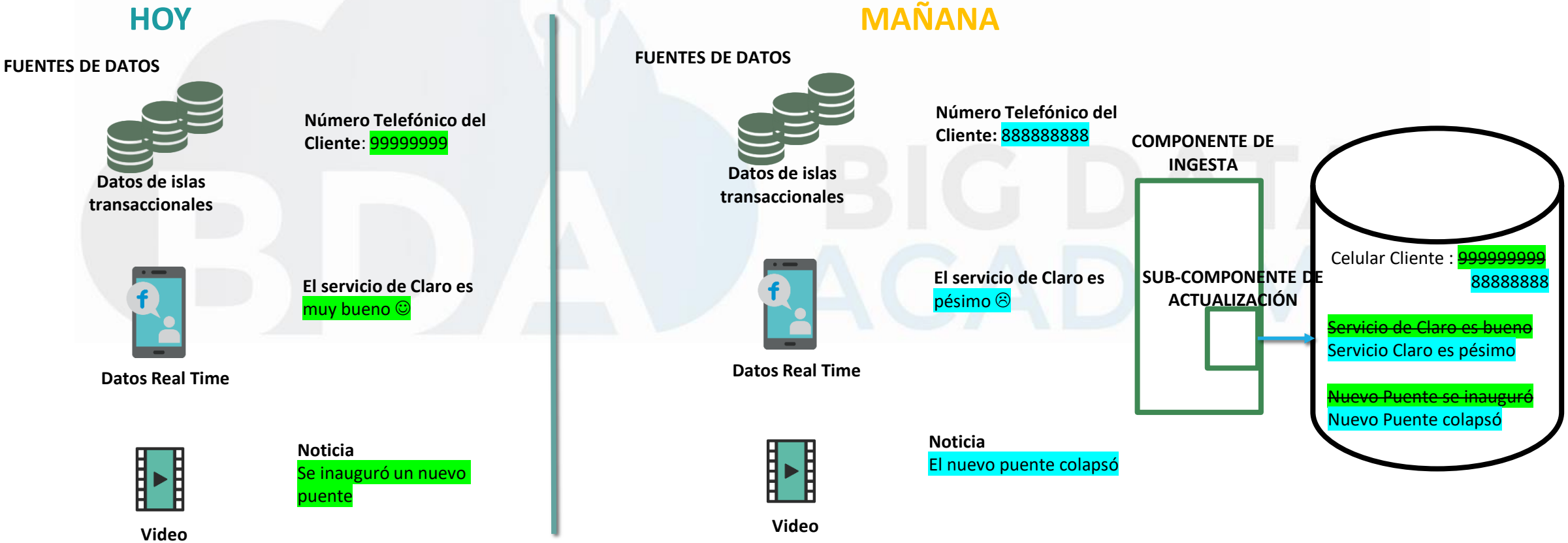
E-mail

Documentos

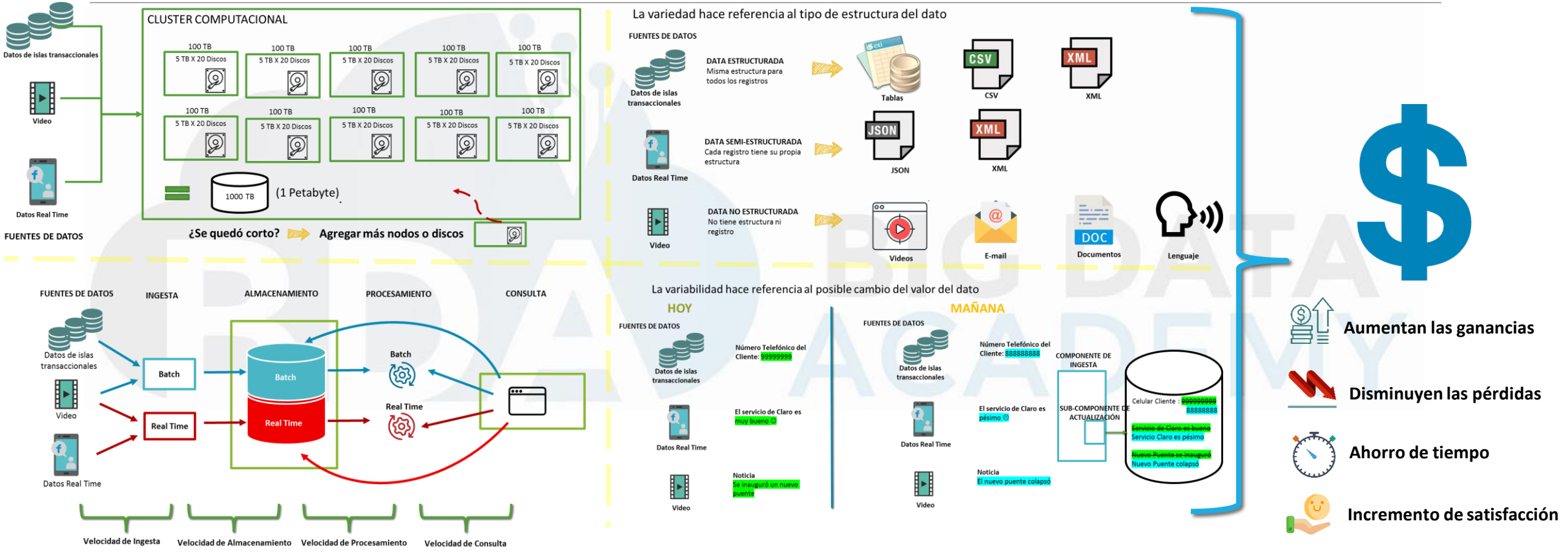
Lenguaje

La filosofía de Big Data: Variabilidad

La variabilidad hace referencia al posible cambio del valor del dato

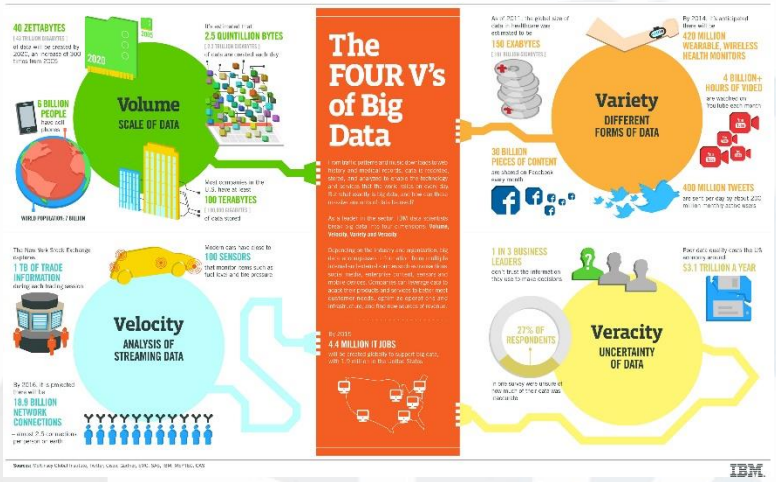


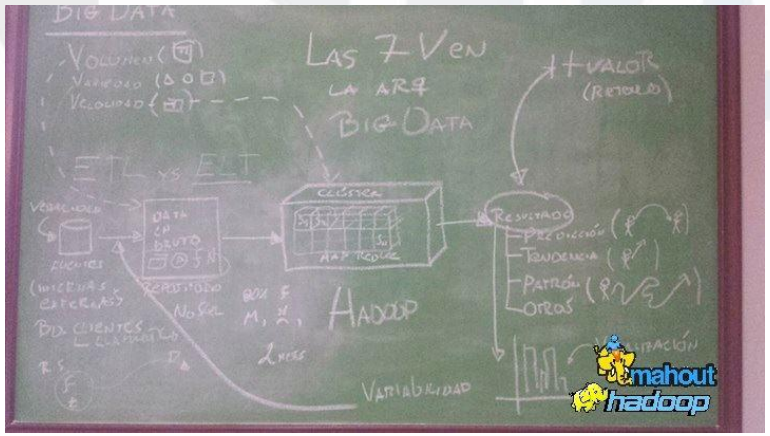
La filosofía de Big Data: Valor

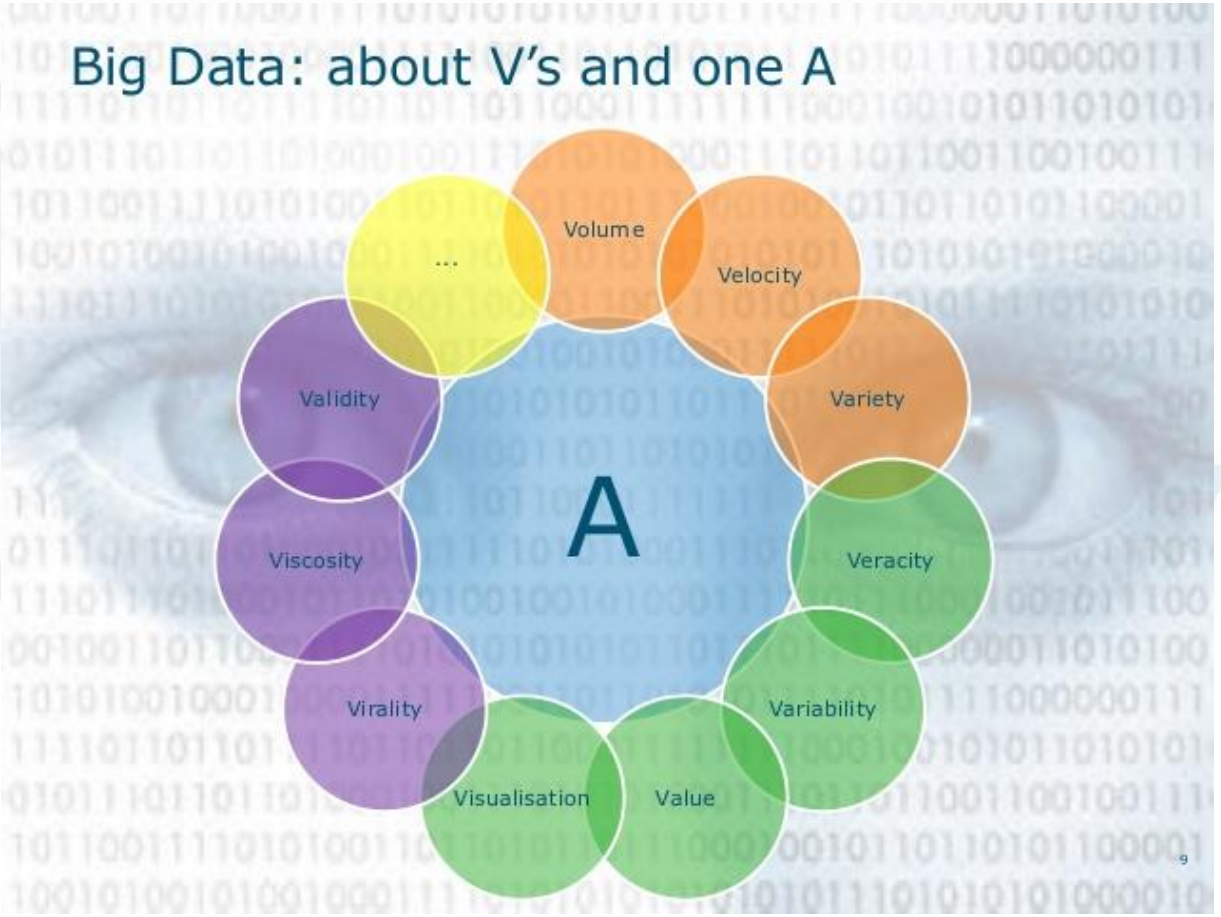


Todos nuestros esfuerzos deben de traducirse en una ganancia para la organización

Otras V


The FOUR V's of Big Data
Volume
SCALE OF DATA
Variety
DIFFERENT FORMS OF DATA
Velocity
ANALYSIS OF STREAMING DATA
Veracity
UNCERTAINTY OF DATA
IBM


BIG DATA
LAS 7 VEN
LA ARQ
BIG DATA
ETL vs ELT
HADOOP
VARIABILIDAD
mahout
hadoop


Big Data: about V's and one A
A
Volume
Velocity
Variety
Veracity
Variability
Value
Visualisation
Virality
Viscosity
Validity
...

¿Por qué Big Data es un marco de trabajo?

- Las 5V
- Clúster computacional
- Paralelización
- Distribución de carga de trabajo
- Escalabilidad
- Alta disponibilidad
- Seguridad
- Gobierno
- Patrones de diseño

- Hadoop
- Hive
- HBase
- Spark
- Kafka
- Cassandra
- Lenguajes de programación

BIG DATA

Big Data: Conceptos

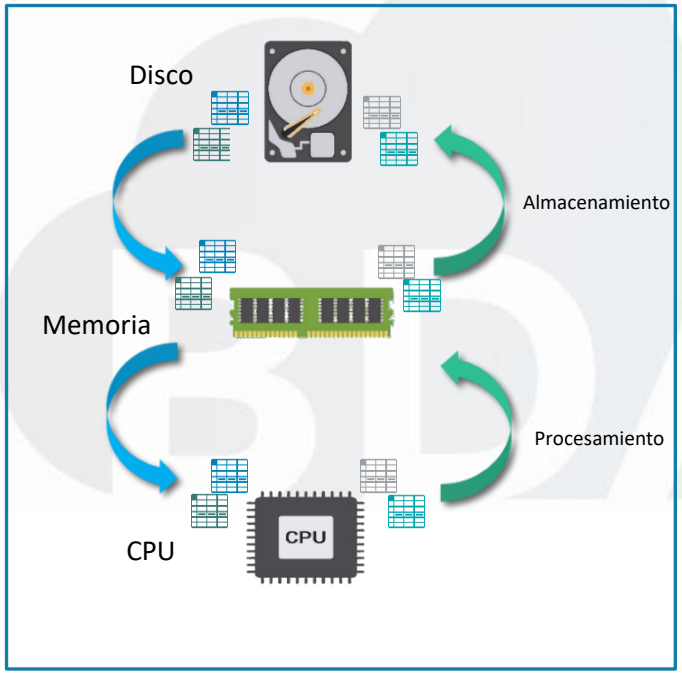
Clúster computacional

¿Cómo trabaja una computadora?

100 TB DISCO

256 GB RAM

40 NÚCLEOS DE CPU



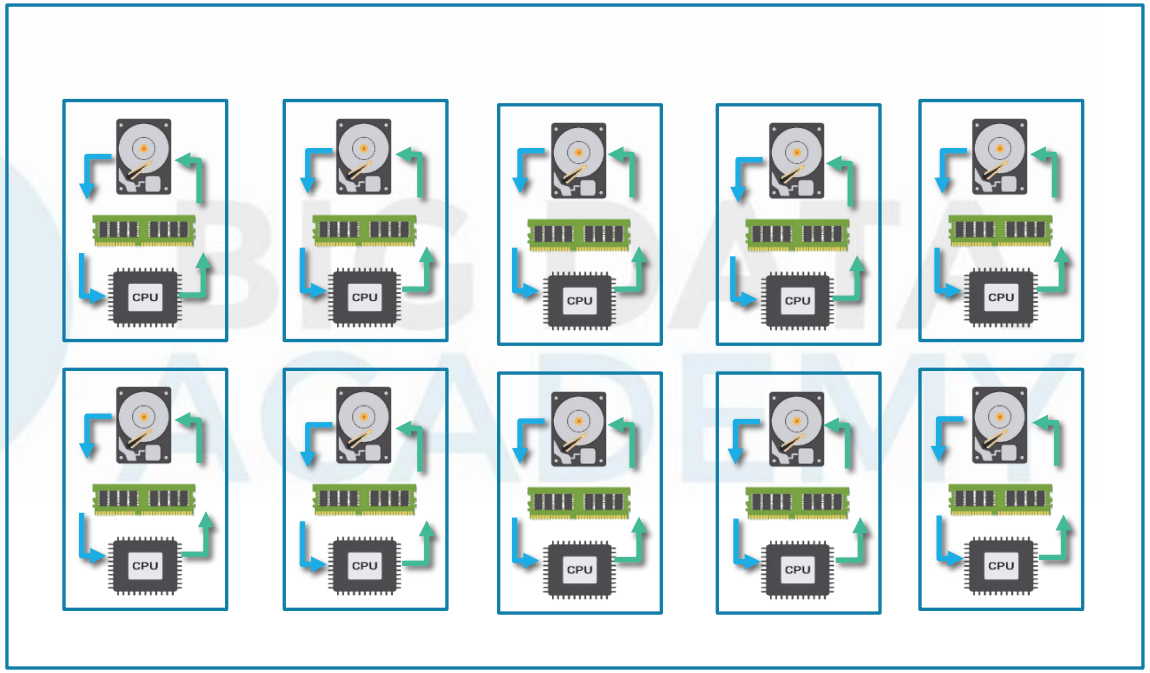
Un servidor común

¿Cómo trabaja un clúster?

1000 TB DISCO

2560 GB RAM

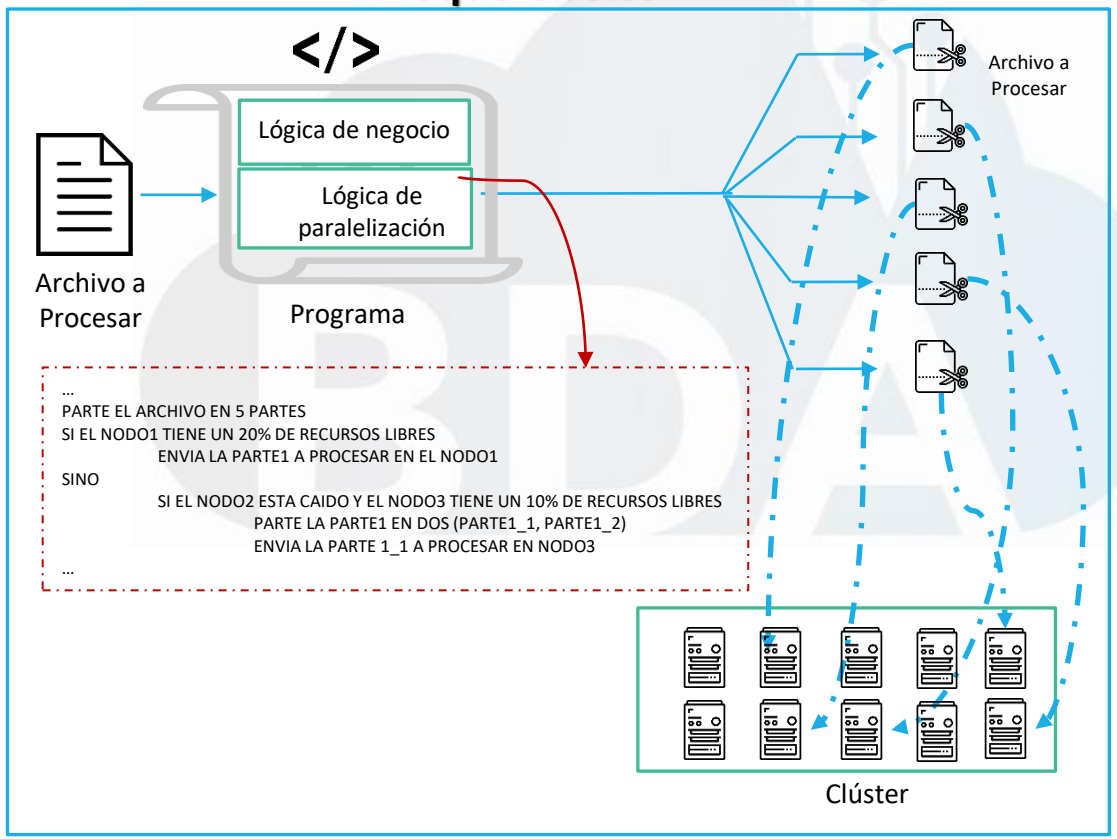
400 NÚCLEOS DE CPU



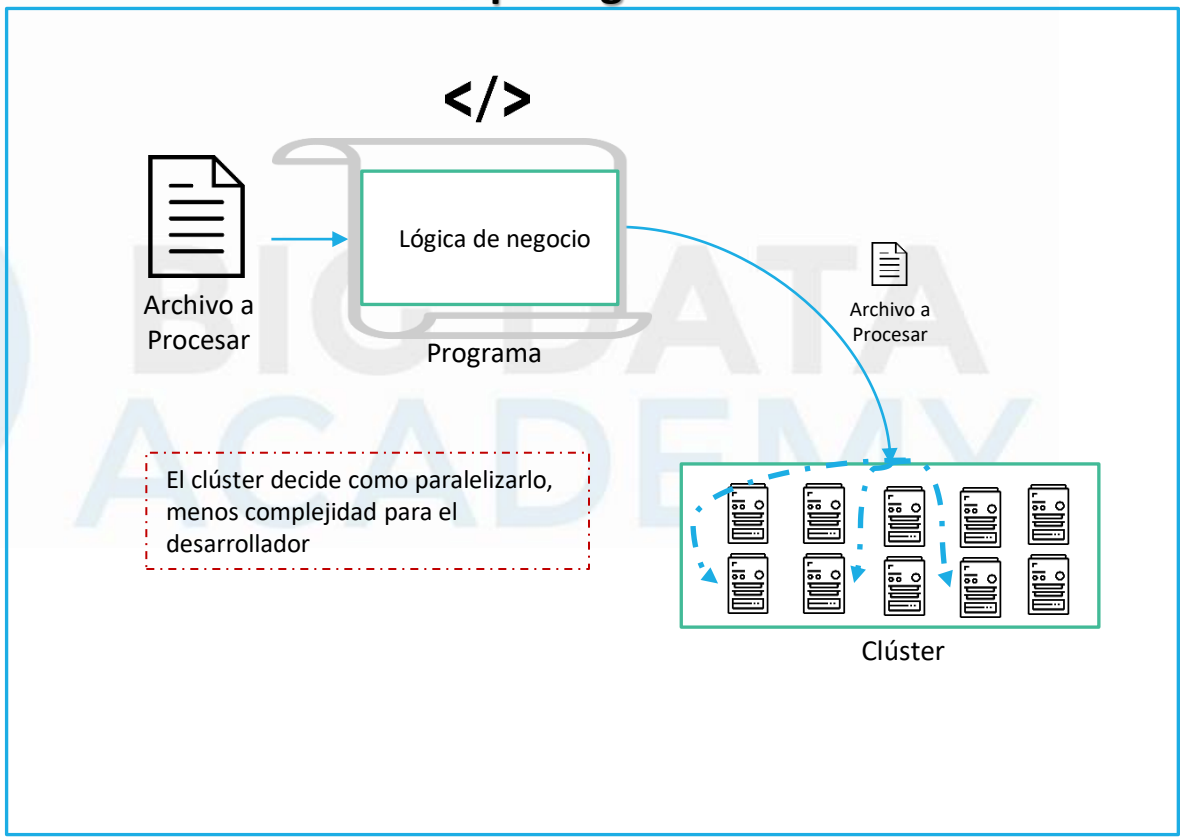
Un clúster es la suma de los recursos computacionales de los servidores que lo conforman, es como si tuviésemos una “super-computadora”

La paralelización

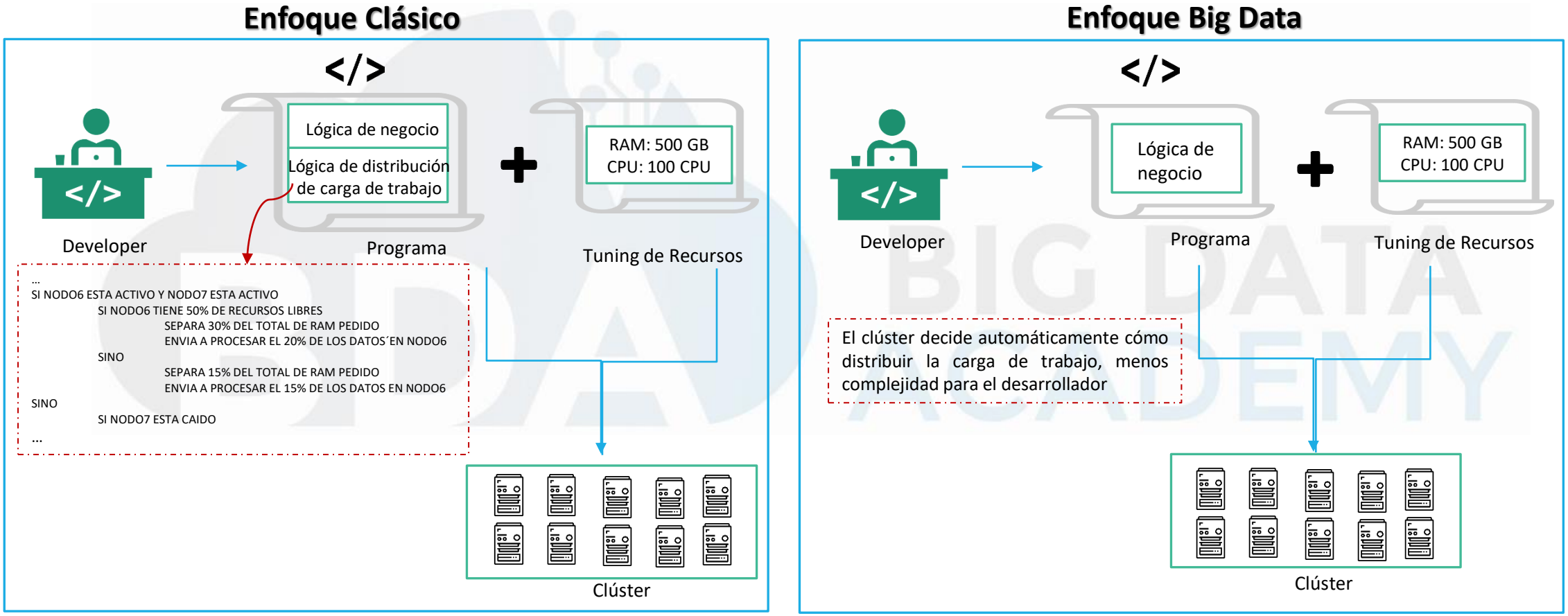
Enfoque Clásico



Enfoque Big Data



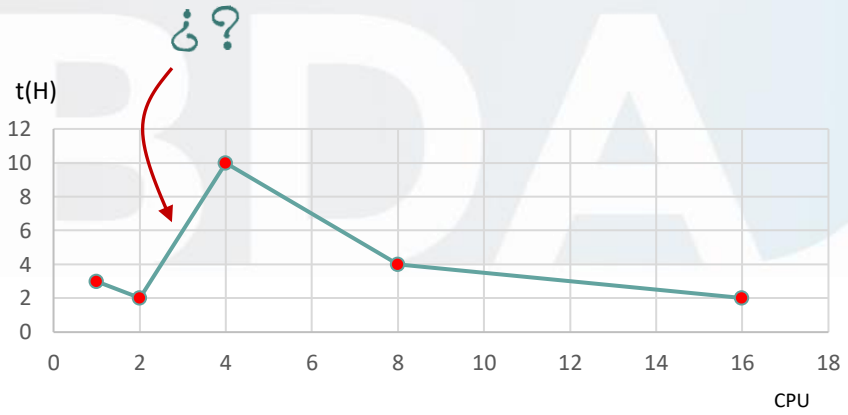
La distribución de carga de trabajo



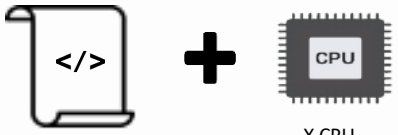
Escalabilidad (de proceso)

Enfoque Clásico

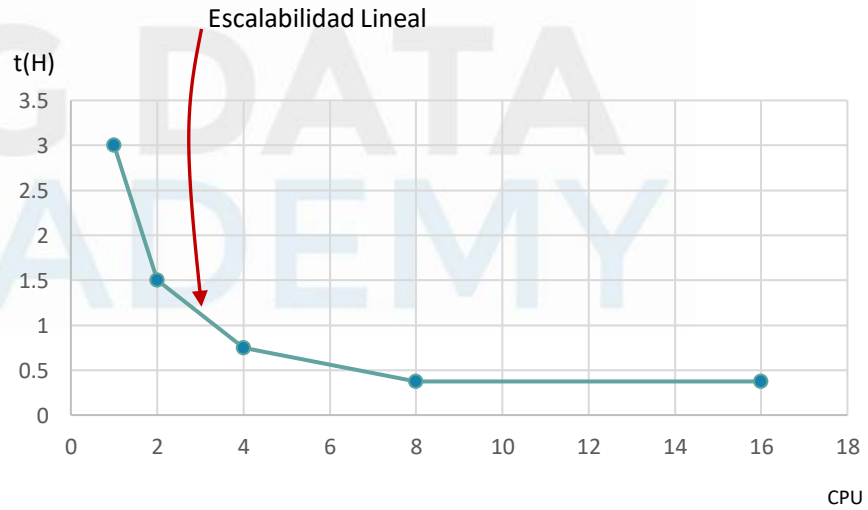
CPU	t
1	3H
2	2H
4	10H
8	4H
16	2H



Enfoque Big Data



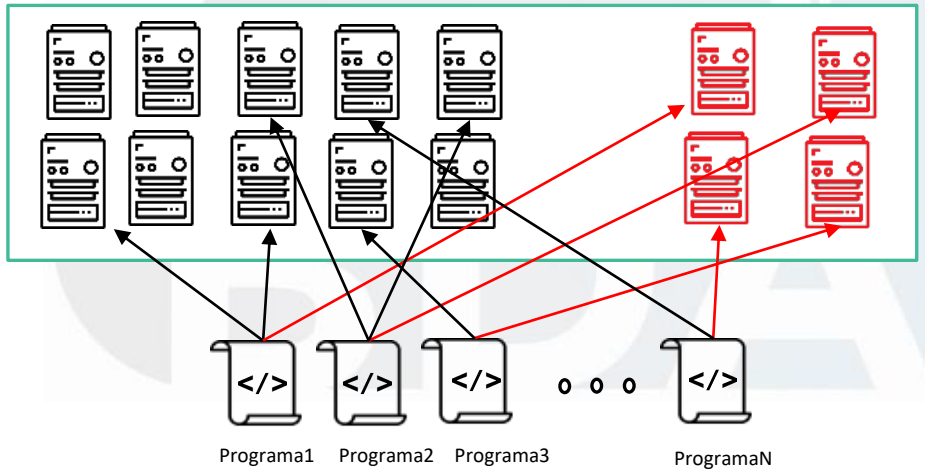
CPU	t
1	3H
2	1.5H
4	0.75H
8	0.375 H
16	0.375H



Escalabilidad (de hardware)

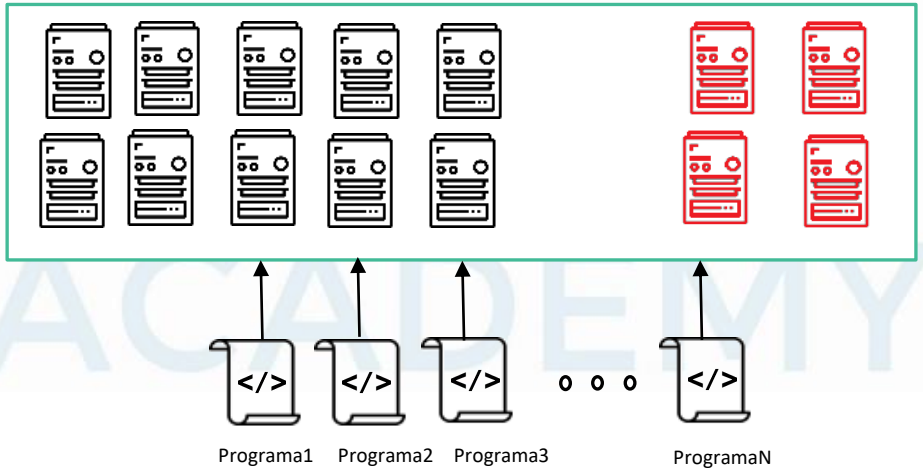
Enfoque Clásico

Agreguemos más nodos al clúster



Modificar los programas para que hagan uso de los nuevos nodos

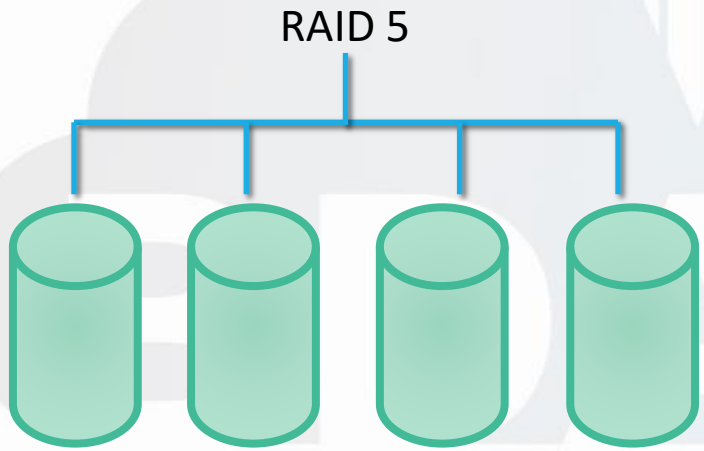
Enfoque Big Data



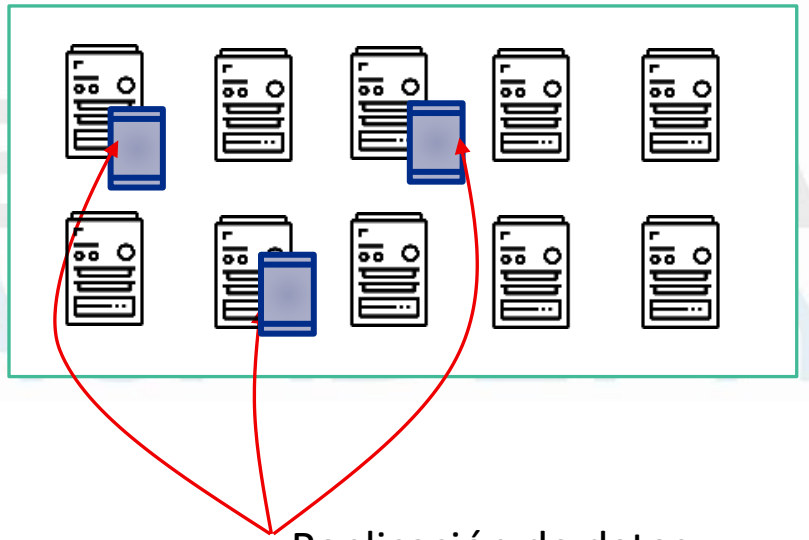
No es necesario modificar los programas, el clúster es visto como un todo

Alta disponibilidad (de datos)

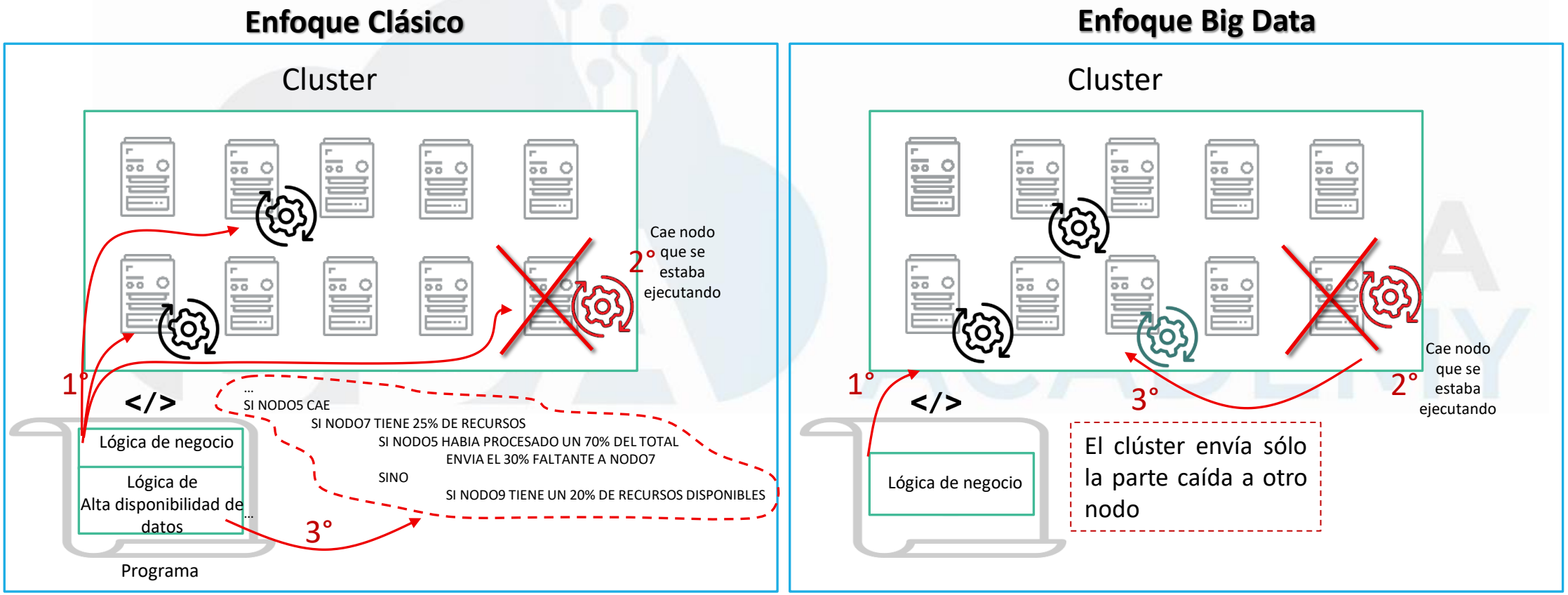
Enfoque Clásico



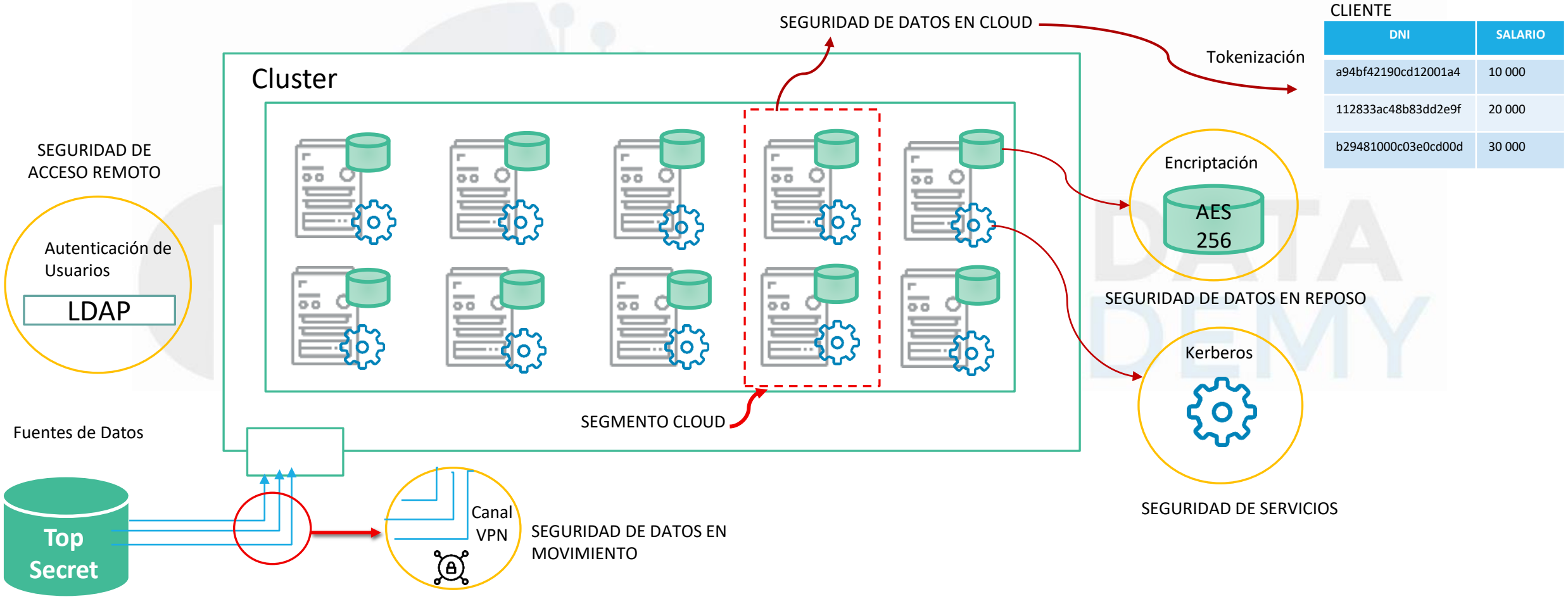
Enfoque Big Data



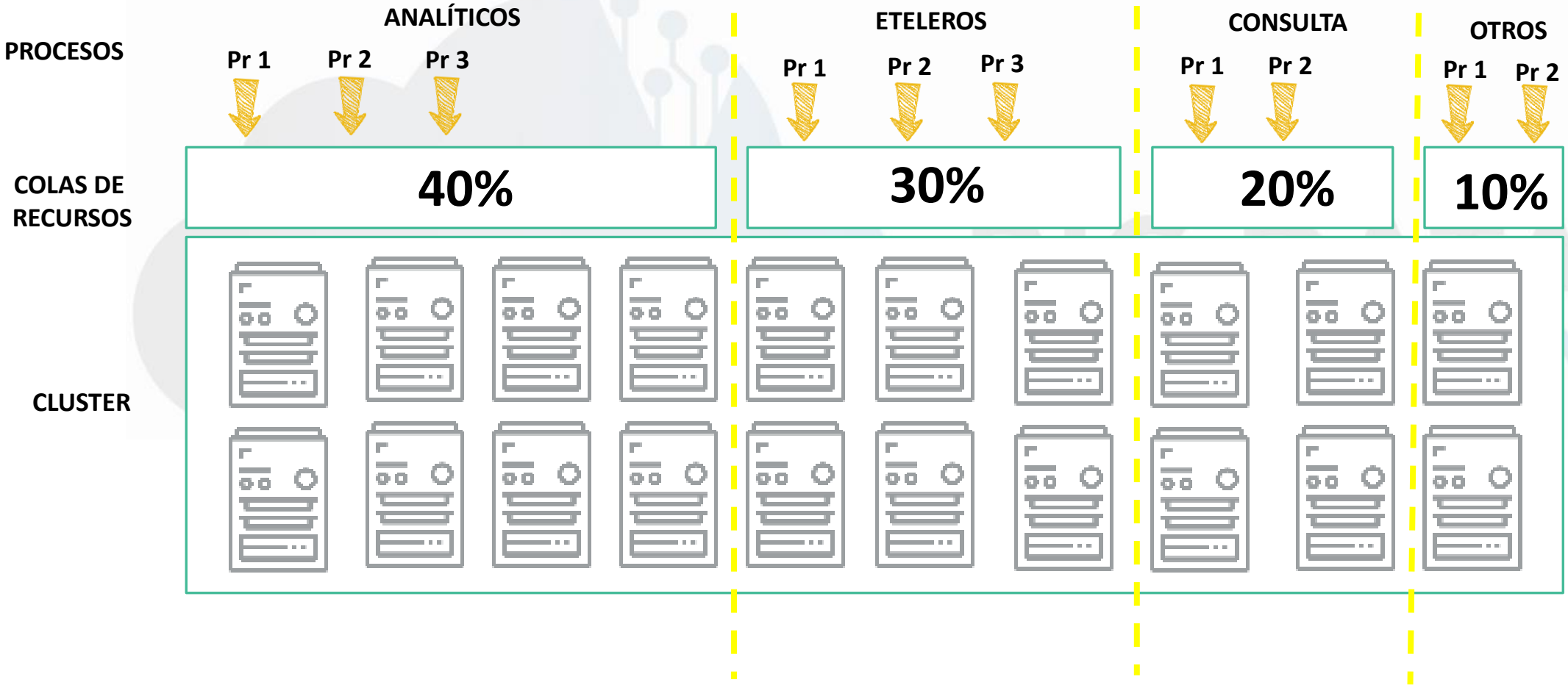
Alta disponibilidad (de proceso)



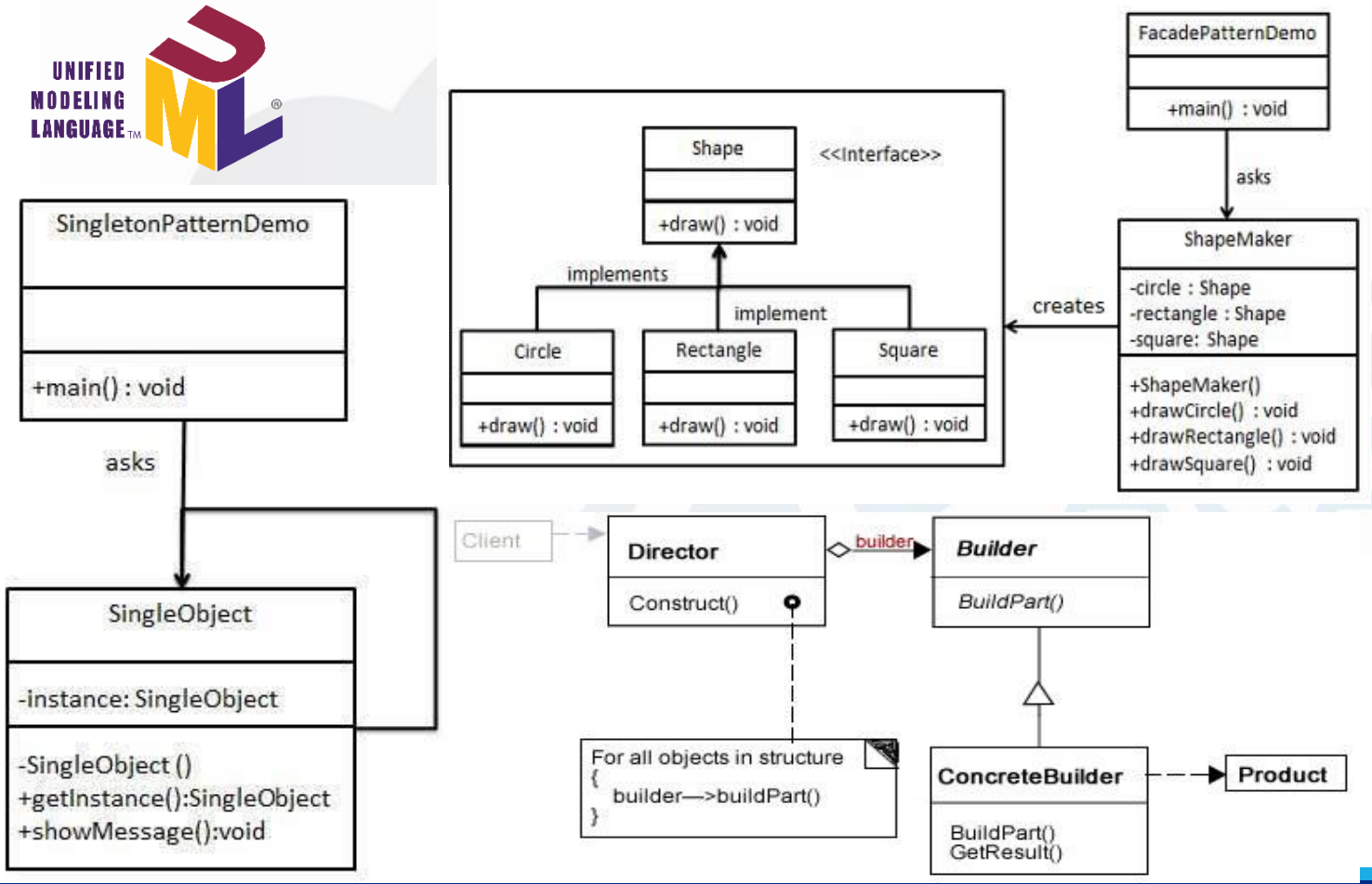
Seguridad

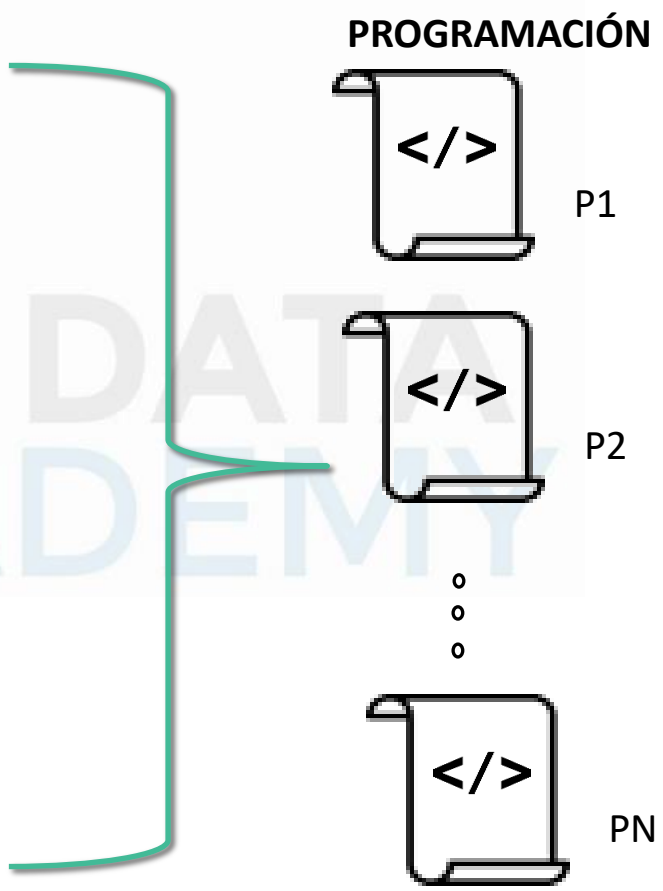

SEGURIDAD DE DATOS EN CLOUD
Tokenización
CLIENTE
DNI
SALARIO
a94bf42190cd12001a4
10 000
112833ac48b83dd2e9f
20 000
b29481000c03e0cd00d
30 000
Cluster
SEGURIDAD DE ACCESO REMOTO
Autenticación de Usuarios
LDAP
Encriptación
AES 256
SEGURIDAD DE DATOS EN REPOSO
Kerberos
SEGMENTO CLOUD
Fuentes de Datos
Top Secret
Canal VPN
SEGURIDAD DE DATOS EN MOVIMIENTO
SEGURIDAD DE SERVICIOS

Gobierno


PROCESOS
COLAS DE RECURSOS
CLUSTER
ANALÍTICOS
Pr 1
Pr 2
Pr 3
40%
ETELEROS
Pr 1
Pr 2
Pr 3
30%
CONSULTA
Pr 1
Pr 2
20%
OTROS
Pr 1
Pr 2
10%

Patrones de diseño





Big Data: Tecnologías

Tipos de tecnologías

INGESTA

- Sqoop
- kafka
- IBM DataStage
- Flume

ALMACENAMIENTO

- hadoop
- Apache HBase
- Ignite

PROCESAMIENTO

- Hive
- Apache Spark
- Apache Phoenix
- cloudera Impala
- H2O.ai
- Storm
- Solr
- Apache Kudu
- mahout

EXPLOTACIÓN

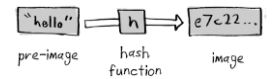
- Hue
- rapidminer
- Qlik
- jupyter
- tableau

GOBIERNO

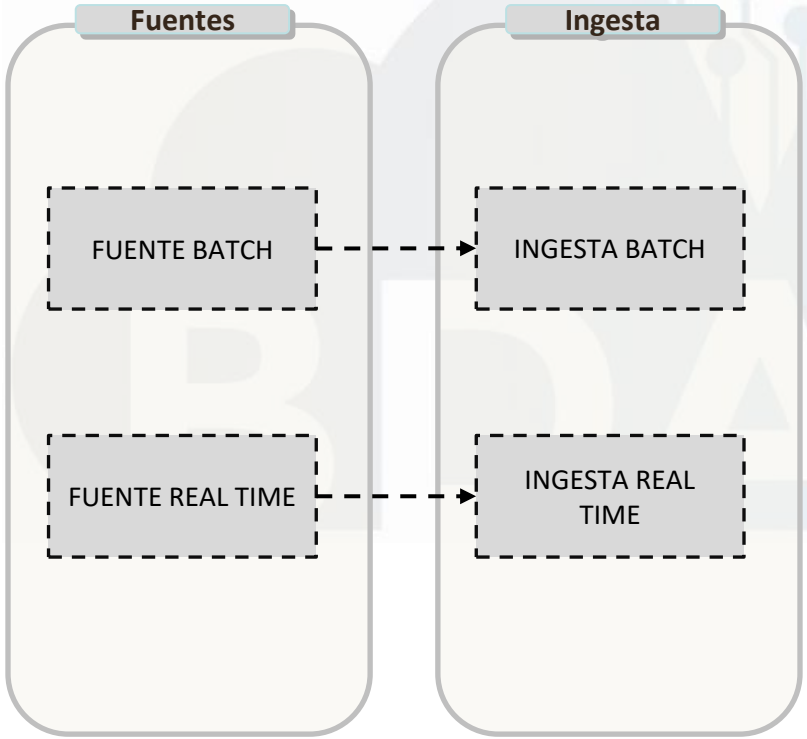
- cloudera
- Hortonworks (Powering the future of data)
- Bamboo
- Jenkins

SEGURIDAD

- data encryption
- VPN
- ACL
- LDAP
- PCI Compliant
- Kerberos



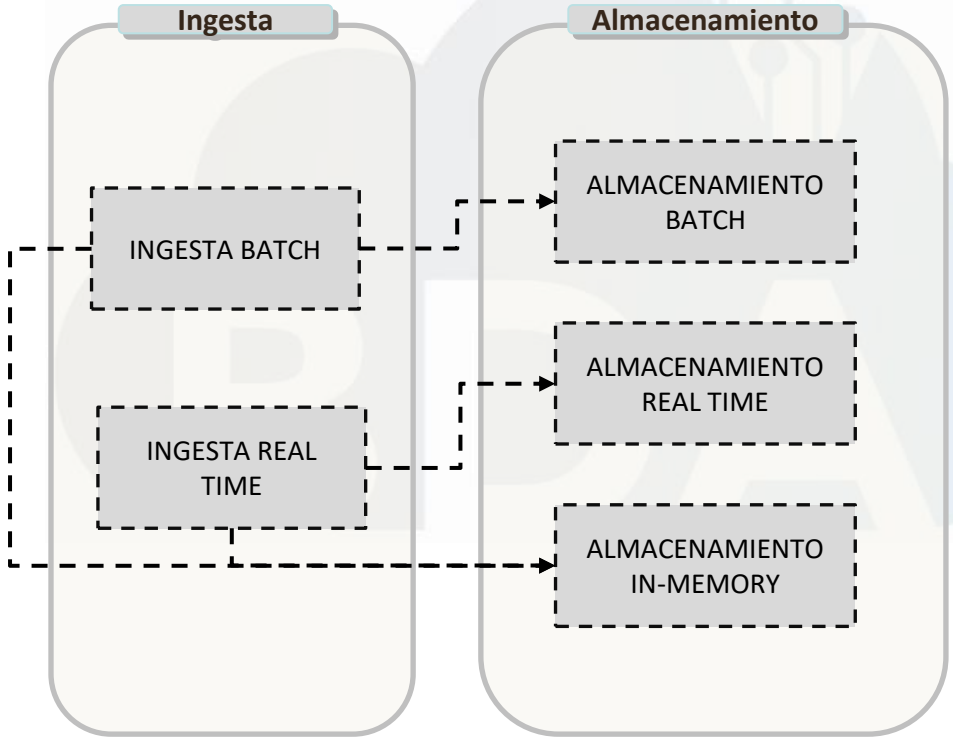
Tecnologías de ingesta



Quiero ingestar <u>de forma batchera</u>

Quiero ingestar <u>de forma real time</u>

Tecnologías de almacenamiento

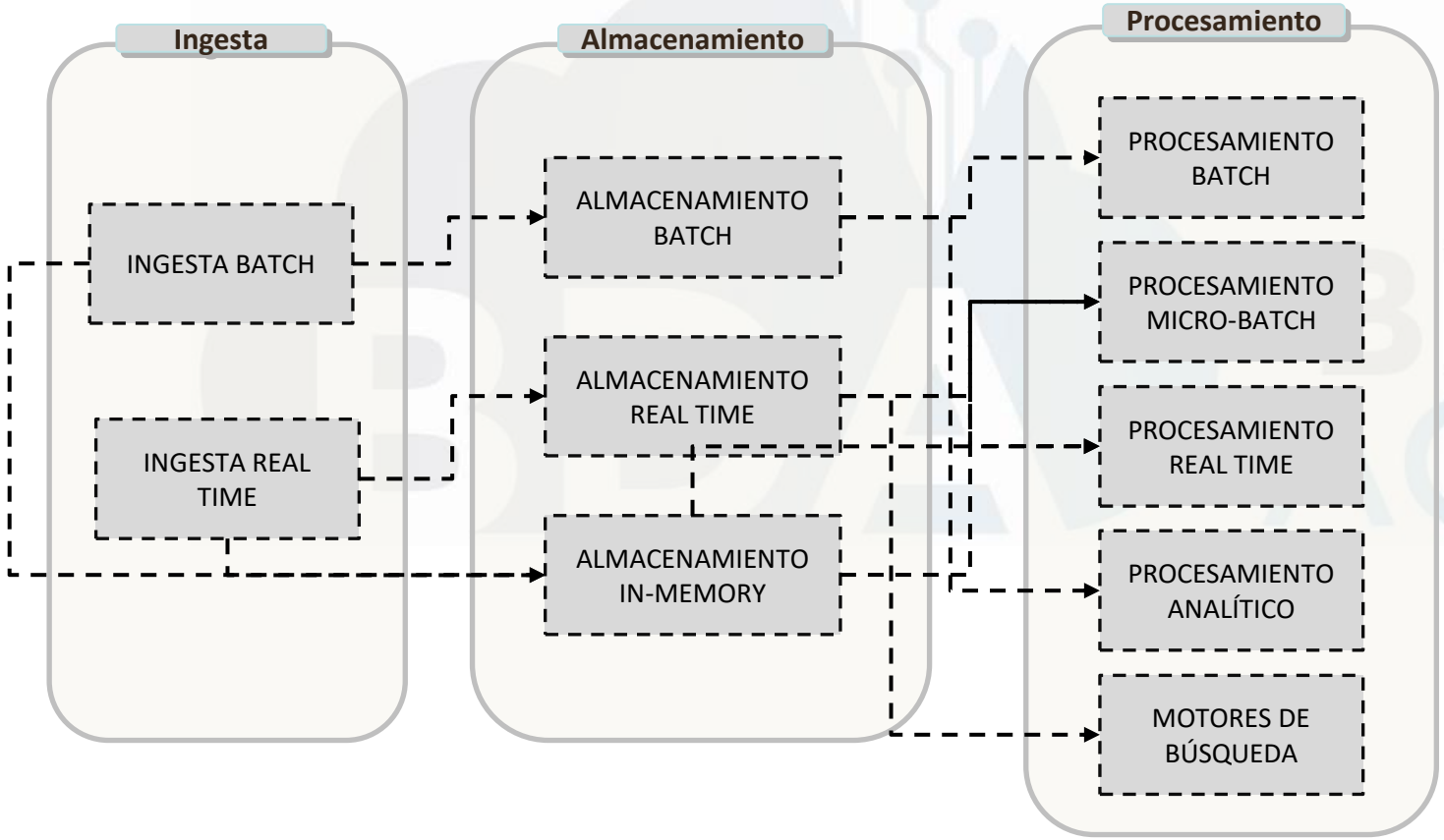


Quiero almacenar <u>de forma batchera</u>

Quiero almacenar <u>de forma real time</u>

Quiero almacenar <u>in-memory</u>

Tecnologías de procesamiento



Quiero procesar <u>de forma batchera</u>

HIVE · APACHE Spark

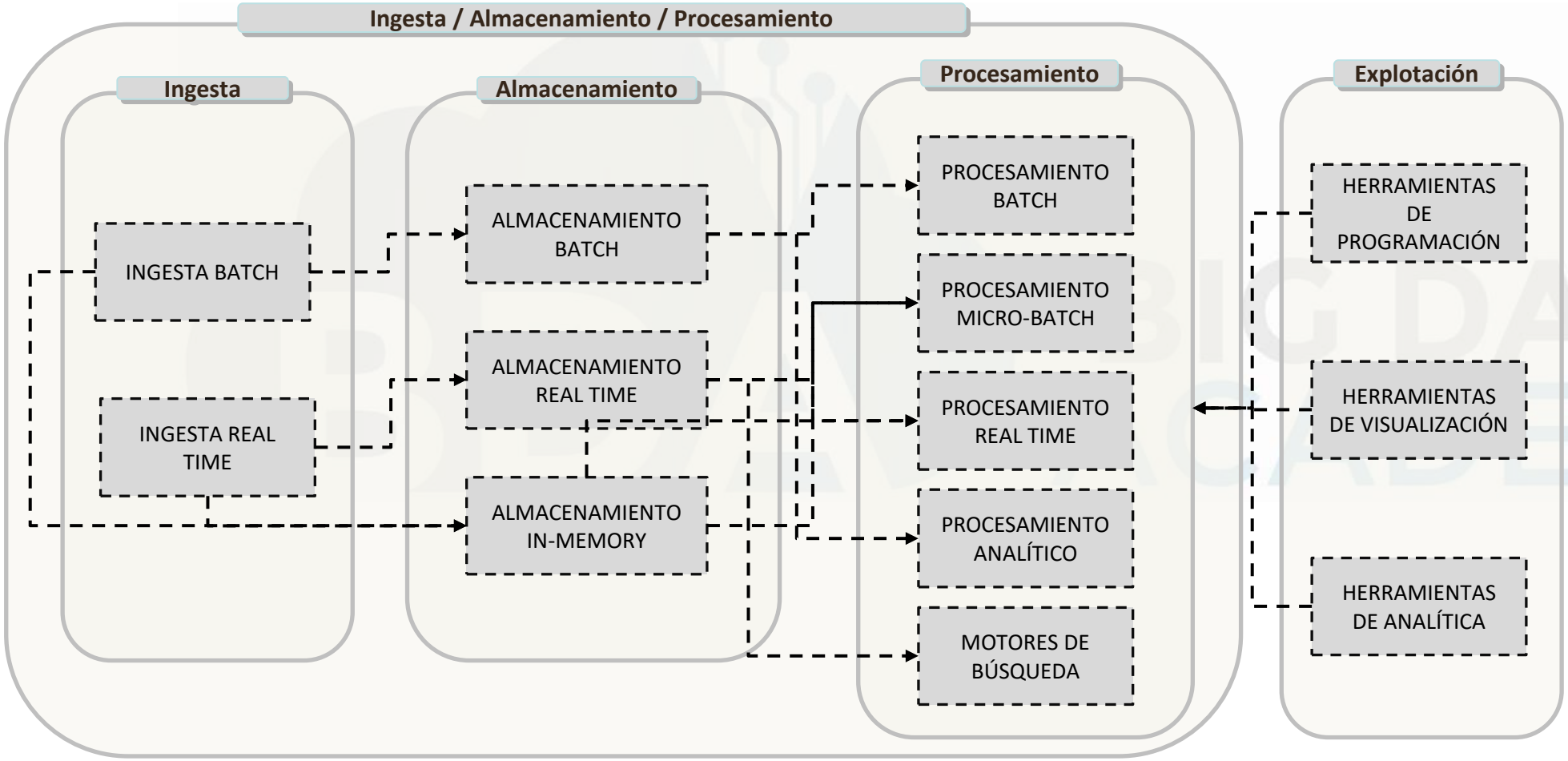
Quiero procesar de forma <u>real time</u>

APACHE Spark · APACHE PHOENIX · STORM

Quiero procesar de forma <u>analítica</u>

mahout · APACHE Spark · H_2O.ai · cloudera IMPALA · APACHE KUDU · Solr

Tecnologías de explotación



Quiero programar

jupyter HUE

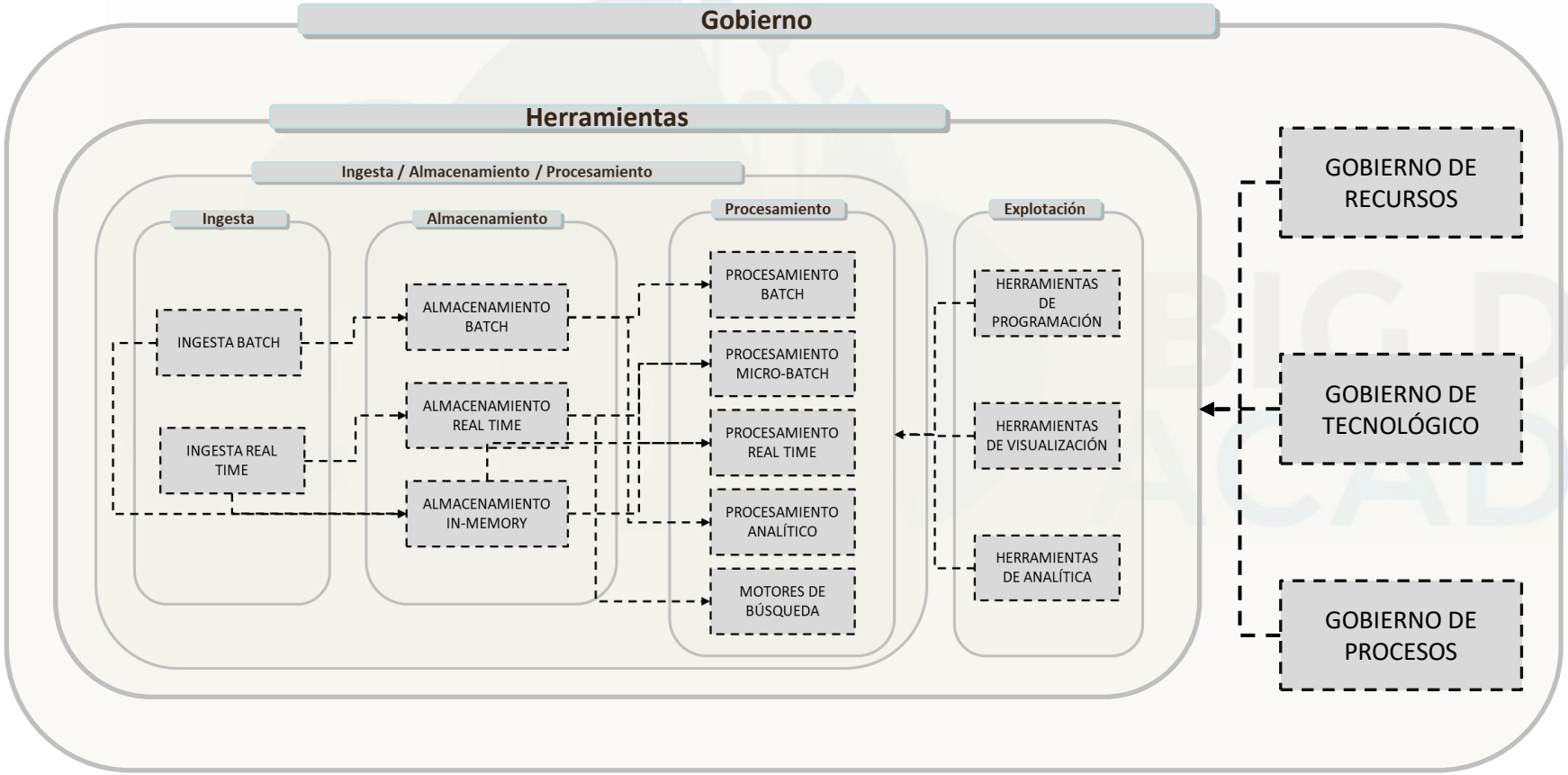
Quiero pintar gráficos y reportes

Qlik +ableau

Quiero analizar datos

rapidminer

Tecnologías de gobierno



Quiero <u>administrar el acceso a RAM, CPU, prioridades</u>

cloudera HORTONWORKS POWERING THE FUTURE OF DATA™

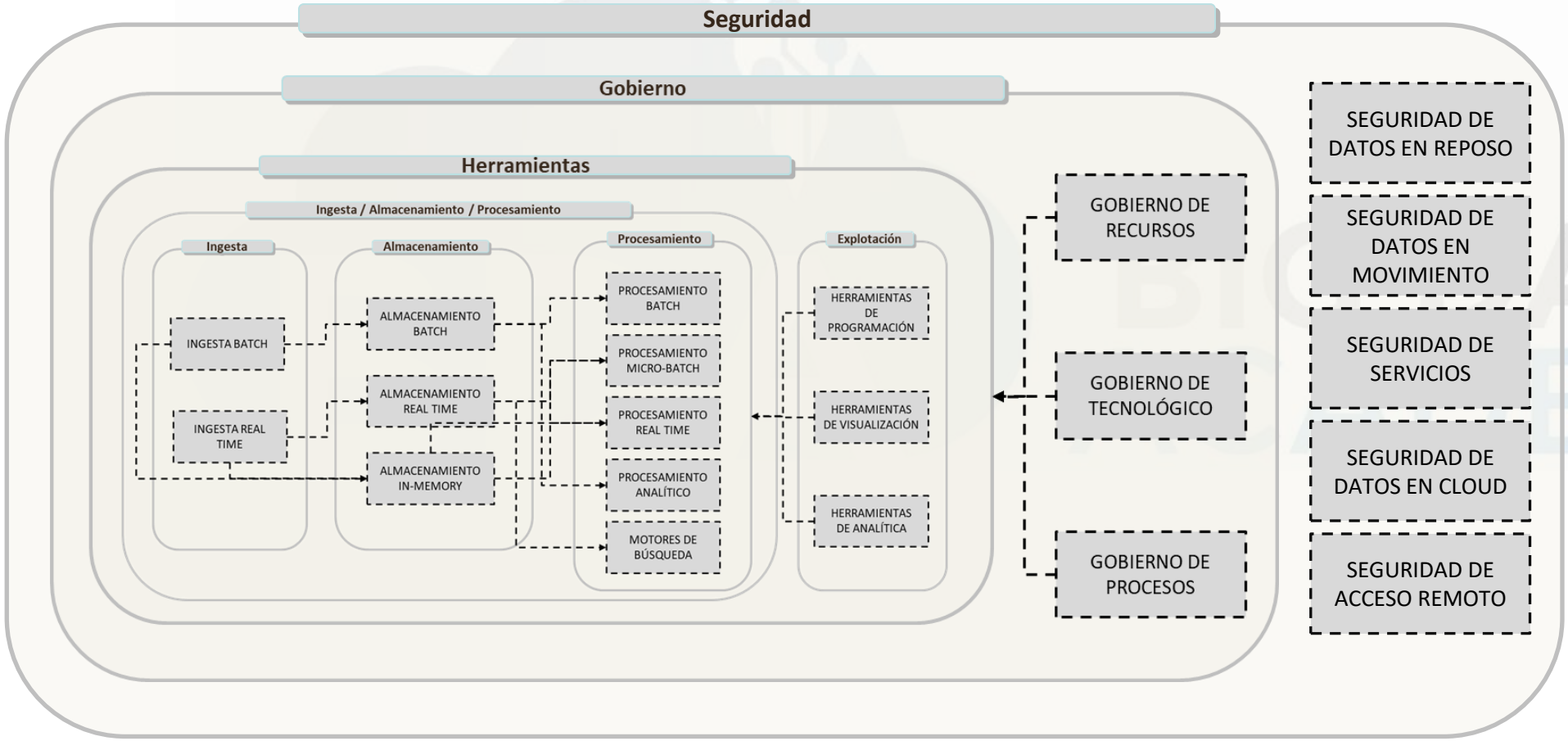
Quiero <u>administrar las herramientas instaladas</u>

cloudera HORTONWORKS POWERING THE FUTURE OF DATA™

Quiero <u>administrar el ciclo de desarrollo y puesto en producción</u>

Jenkins Bamboo IBM

Tecnologías de seguridad



Quiero asegurar datos en mis discos duros

Quiero asegurar los datos que viajan por la red

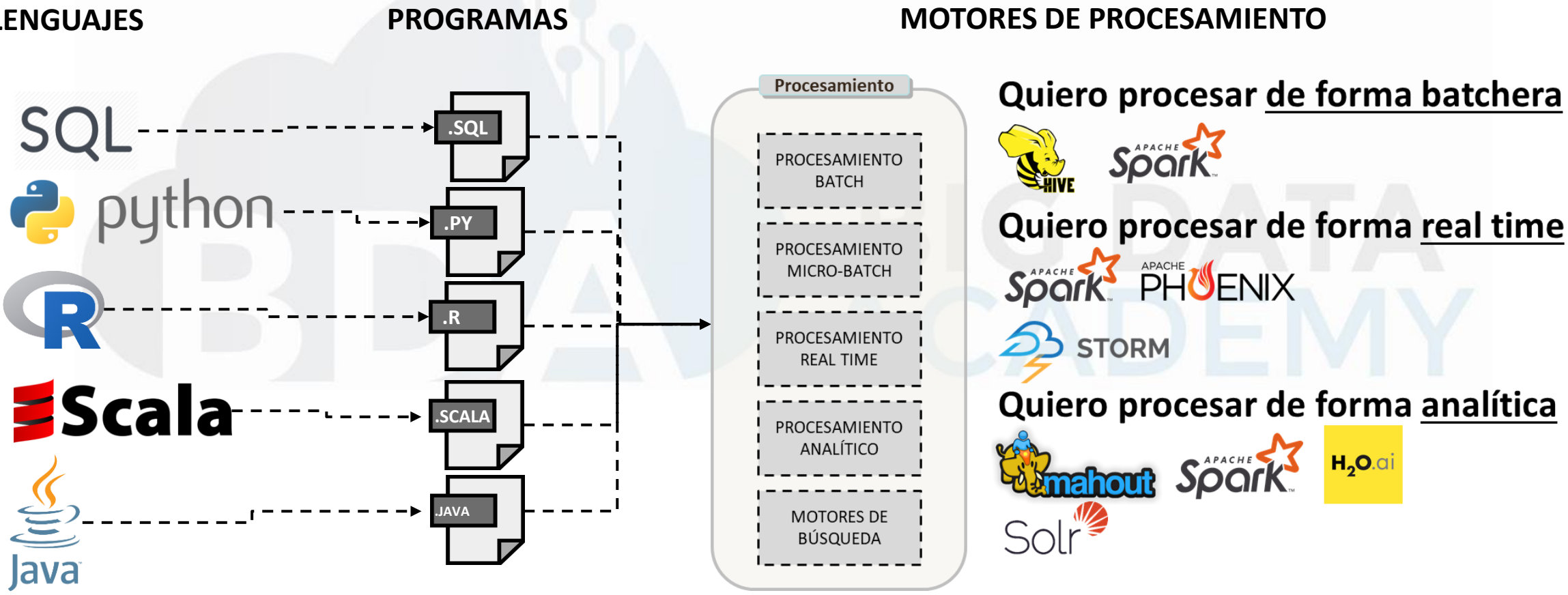
Quiero asegurar el acceso a servicios y procesos

Quiero asegurar mis datos en la nube

Quiero asegurar el acceso al clúster

Lenguajes de programación

Lenguajes de programación



Características

- Orientado a **datos estructurados**
- Aprendizaje fácil
- **Estándar para consultas a bases de datos**

- Orientado a la **programación funcional**
- Interpretado, no tipado y muy flexible
- **Librerías casi para todo**

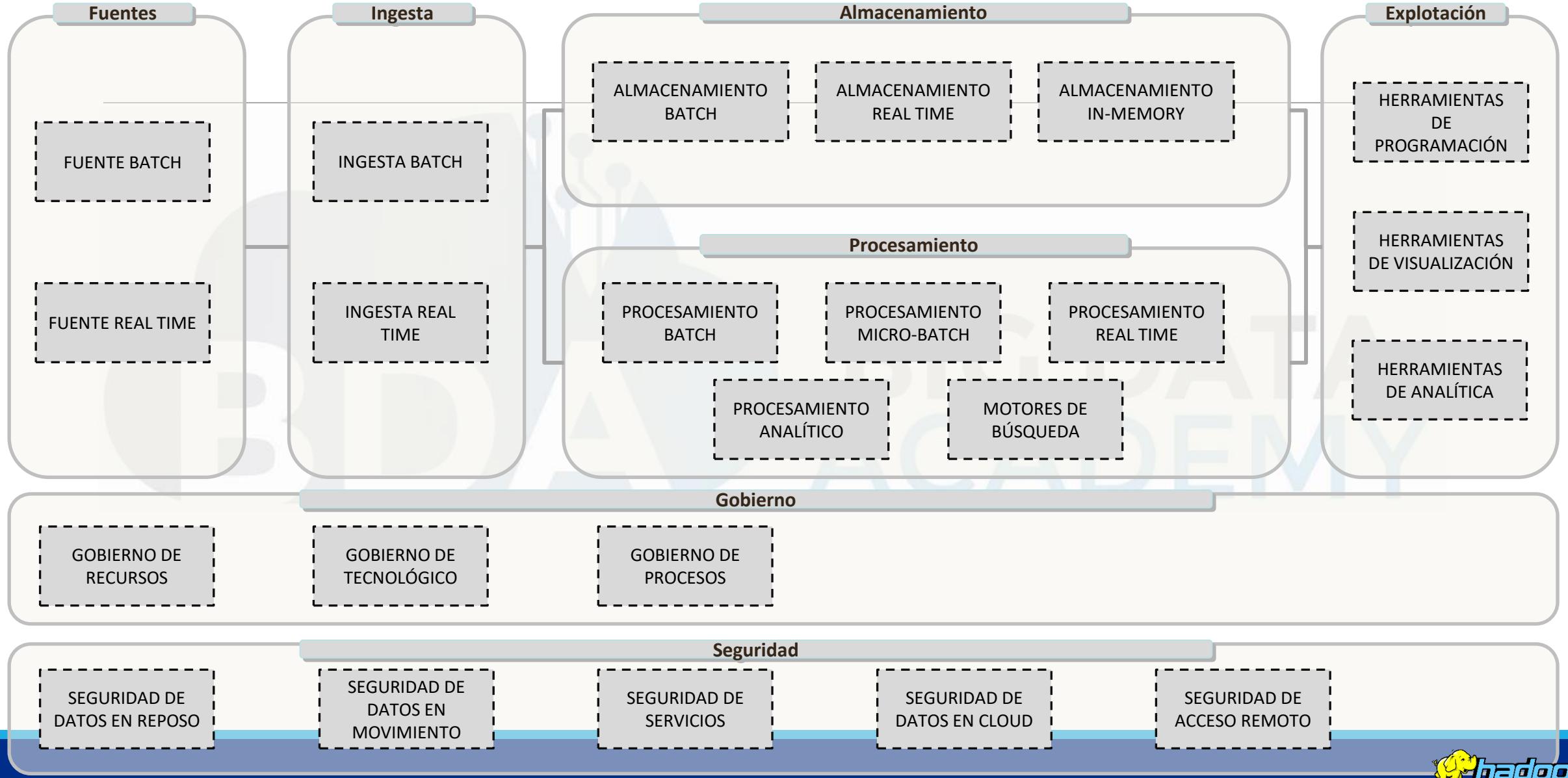
- Orientado a la **estadística**
- Interpretado
- **Gráficos avanzados simples de realizar**

- Orientado a la **programación funcional y de objetos**
- Compilado y no tipado
- **Fork de Java**

- Orientado a la **programación de objetos**
- Compilado y tipado
- **Aprovecha muy bien los servidores con grandes recursos computacionales**

Arquitectura general de Big Data

Arquitectura conceptual



Tipos de tecnologías

INGESTA

- Sqoop
- kafka
- IBM DataStage
- Flume

ALMACENAMIENTO

- hadoop
- Apache HBase
- Ignite

PROCESAMIENTO

- Hive
- Apache Spark
- Apache Phoenix
- cloudera Impala
- H2O.ai
- Storm
- Solr
- Apache Kudu
- mahout

EXPLOTACIÓN

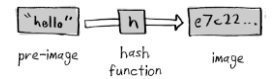
- Hue
- rapidminer
- Qlik
- jupyter
- tableau

GOBIERNO

- cloudera
- Hortonworks (Powering the future of data)
- Bamboo
- Jenkins

SEGURIDAD

- data encryption
- VPN
- ACL
- LDAP
- PCI Compliant
- Kerberos



Arquitectura tecnológica

BIG DATA ACADEMY

Fuentes

FUENTE BATCH

FUENTE REAL TIME

Ingesta

INGESTA BATCH

INGESTA REAL TIME

Almacenamiento

ALMACENAMIENTO BATCH

ALMACENAMIENTO REAL TIME

ALMACENAMIENTO IN-MEMORY

Procesamiento

PROCESAMIENTO BATCH

PROCESAMIENTO MICRO-BATCH

PROCESAMIENTO REAL TIME

PROCESAMIENTO ANALÍTICO

MOTORES DE BÚSQUEDA

Explotación

HERRAMIENTAS DE PROGRAMACIÓN

HERRAMIENTAS DE VISUALIZACIÓN

HERRAMIENTAS DE ANALÍTICA

Gobierno

GOBIERNO DE RECURSOS

GOBIERNO DE TECNOLÓGICO

GOBIERNO DE PROCESOS

Seguridad

SEGURIDAD DE DATOS EN REPOSO

SEGURIDAD DE DATOS EN MOVIMIENTO

SEGURIDAD DE SERVICIOS

SEGURIDAD DATOS EN CLOUD

SEGURIDAD DE ACCESO REMOTO

Ejercicios teóricos

Ejercicios teóricos

1. ¿Qué es Big Data?
2. ¿Cuál es el objetivo del Big Data?
3. ¿Cuáles son las 5V?
4. ¿Cuál es la diferencia entre variedad y variabilidad?
5. ¿Qué es un clúster computacional?
6. ¿Qué es la paralelización?
7. ¿Qué es la escalabilidad?
8. ¿Cuáles son los puntos claves en la seguridad de un clúster?
9. ¿Qué tipos de tecnologías existen en el Big Data?
10. ¿Qué motores de procesamiento conoce?
11. ¿Qué lenguaje de programación es mejor para programar sobre Big Data?
12. ¿Por qué se prefiere un proceso micro-batch sobre uno real time?
13. Dibuje la arquitectura conceptual de Big Data

Resumen

Hablemos…