Arquitecturas de Big Data

BIG DATA ACADEMY

Alonso Melgarejo [alonsoraulmgs@gmail.com]

Escalabilidad del proceso

Necesidad de negocio

PERSONA DE NEGOCIO

“Tengo una necesidad de negocio”:

- Construye un reporte
- Haz un proceso de limpieza de datos
- Construye una red neuronal

DESARROLLADOR

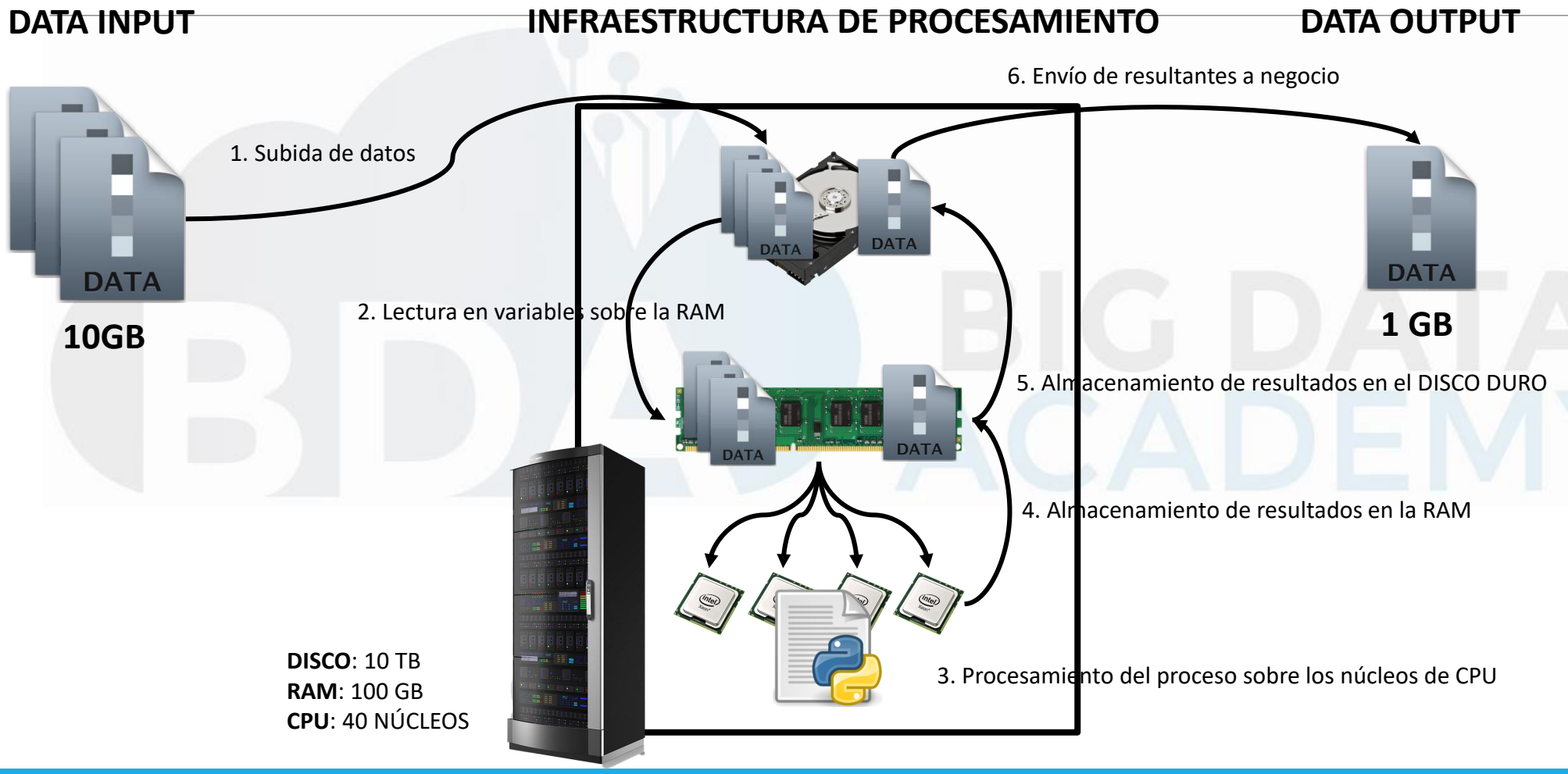
Implementa la necesidad con algún lenguaje:

- Python
- Scala
- R
- SQL

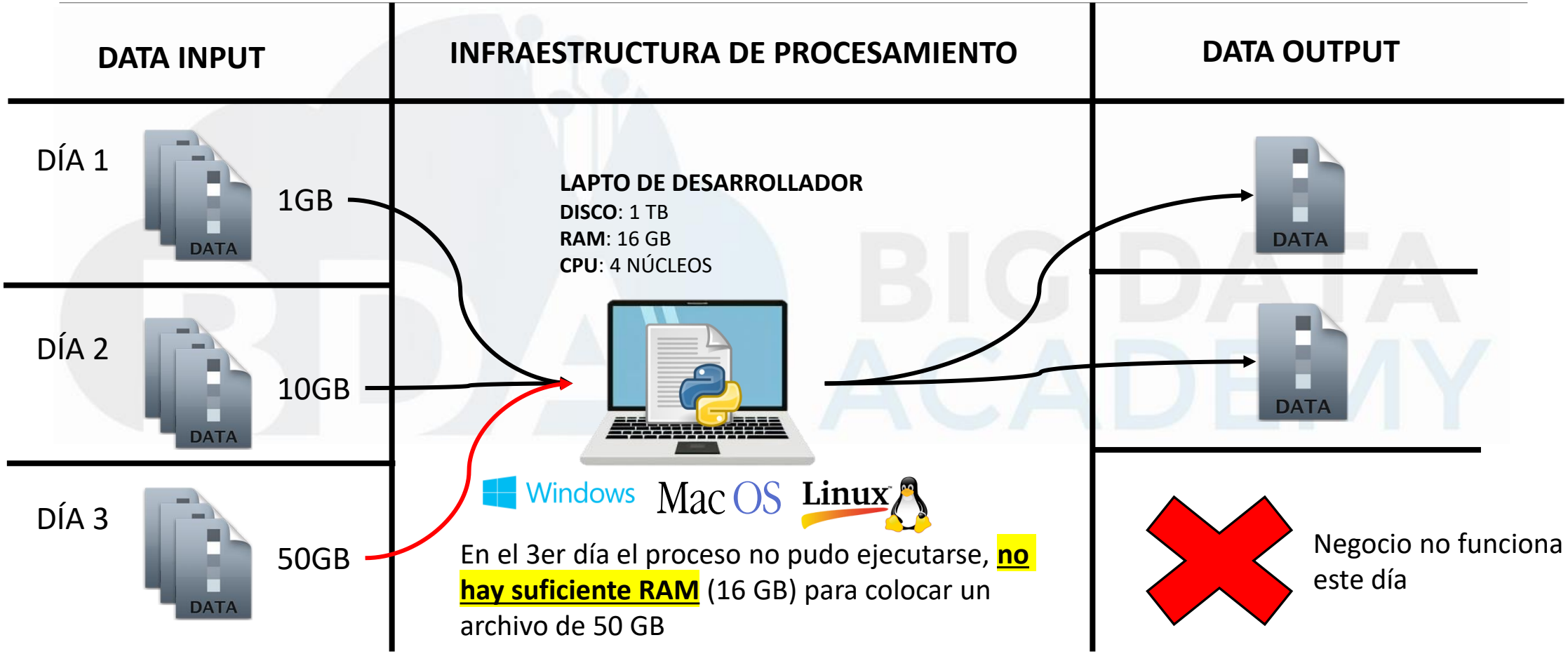
PROCESO DE NEGOCIO

El proceso se ejecutará sobre algún tipo de infraestructura

Ejecución del proceso de negocio



Procesamiento sobre computadoras de escritorio



Solución: Escalamos a un servidor empresarial

PERSONA DE NEGOCIO

“Compraremos un servidor empresarial”

ADQUISICIÓN DE SERVIDOR

5 SEMANAS DESPUÉS

ALGUNA ACTIVIDADES

- Preparación de licitación [1 semana]
- Contacto a proveedores [1 semana]
- Elección de proveedor [1 semana]
- Compra y despliegue [1 semana]
- Instalación de software [1 semana]

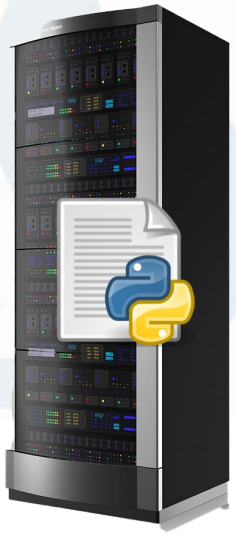
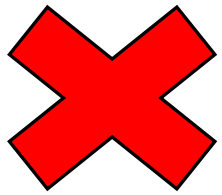
Linux

SERVIDOR EMPRESARIAL

DISCO: 10 TB
RAM: 100 GB
CPU: 40 NÚCLEOS

Durante 5 semanas, negocio no funcionó

Después de 5 semanas: Procesamiento sobre servidor empresarial

DATA INPUT	INFRAESTRUCTURA DE PROCESAMIENTO	DATA OUTPUT
DÍA 1  50GB	 **SERVIDOR EMPRESARIAL** **DISCO**: 10 TB **RAM**: 100 GB **CPU**: 40 NÚCLEOS Linux	
DÍA 2  90GB		
DÍA 3  300GB	En el 3er día el proceso no pudo ejecutarse, **no hay suficiente RAM** (100 GB) para colocar un archivo de 300 GB	 Negocio no funciona este día

Solución: Escalamos a un clúster de servidores

PERSONA DE NEGOCIO

“Compraremos un clúster de servidores”

ADQUISICIÓN DE CLÚSTER

5 SEMANAS DESPUÉS

ALGUNA ACTIVIDADES

- Preparación de licitación [1 semana]
- Contacto a proveedores [1 semana]
- Elección de proveedor [1 semana]
- Compra y despliegue [1 semana]
- Instalación de software [1 semana]

10 SERVIDORES

...

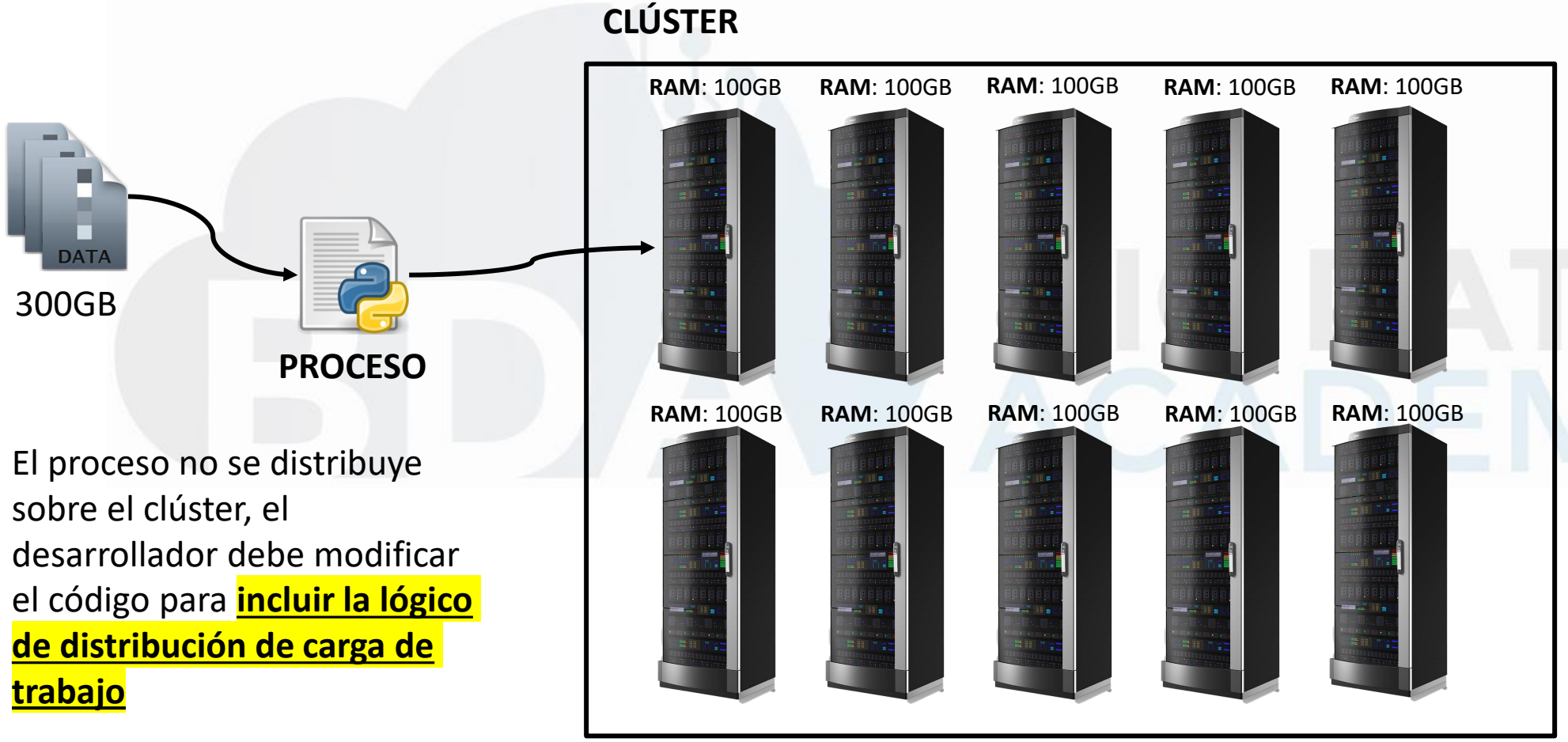
Linux

CADA SERVIDOR:
DISCO: 10 TB
RAM: 100 GB
CPU: 40 NÚCLEOS

POTENCIA DEL CLÚSTER
DISCO: 100 TB
RAM: 1000 GB
CPU: 400 NÚCLEOS

Durante 5 semanas, negocio no funcionó

Problema: El proceso no distribuye la carga de trabajo sobre el clúster



El proceso no se distribuye sobre el clúster, el desarrollador debe modificar el código para **incluir la lógico de distribución de carga de trabajo**

Solución: Adaptar el proceso para distribuir la carga de trabajo

PROCESO ORIGINAL

Sólo contiene la necesidad de negocio (P.E.: 1000 líneas de código)

ADAPTACIÓN DEL PROCESO

5 SEMANAS DESPUÉS

Algunas actividades:

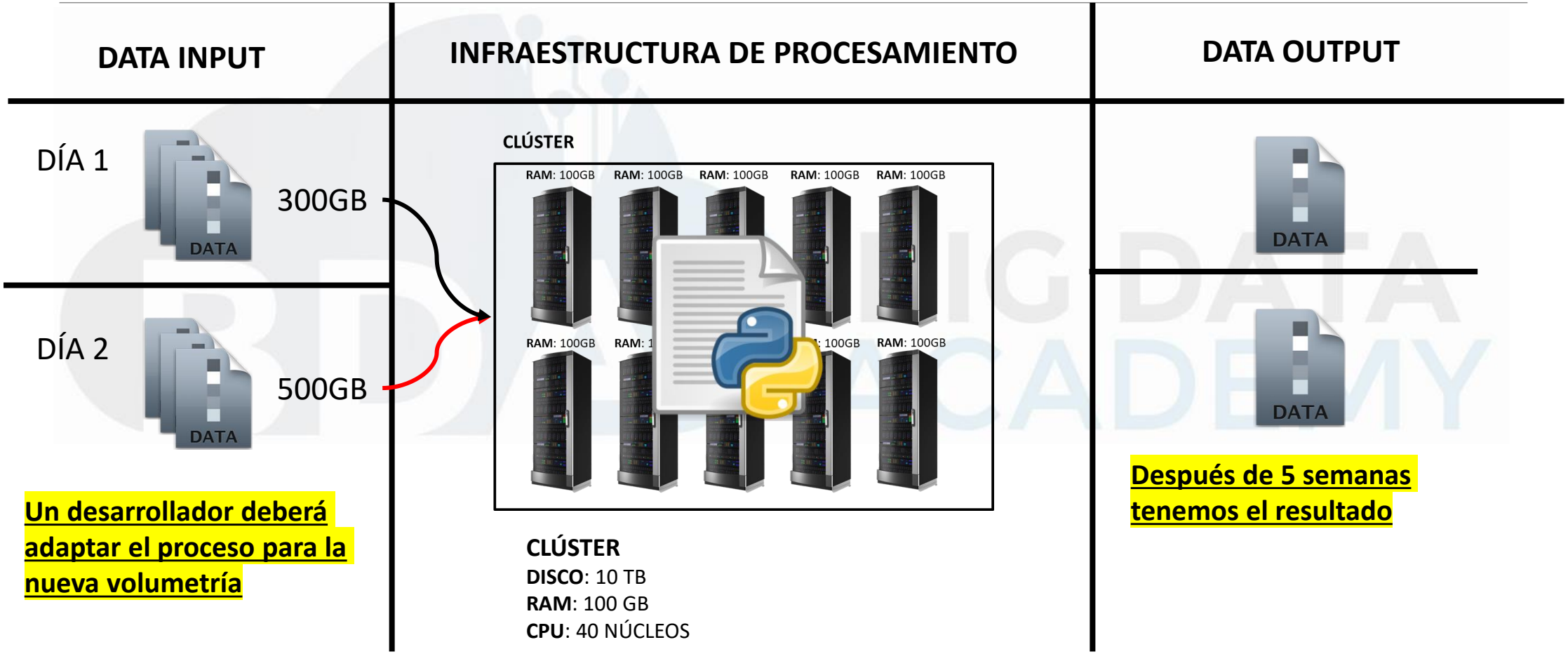
- Reserva de recursos sobre servidores [1 semana]
- Distribución de la carga de trabajo sobre servidores [1 semana]
- Coordinación de ejecución [1 semana]
- Gestión de excepciones [1 semana]
- Pruebas [1 semana]

PROCESO ADAPTADO

Necesidad de negocio + Necesidad técnica (P.E.: 3000 líneas de código)

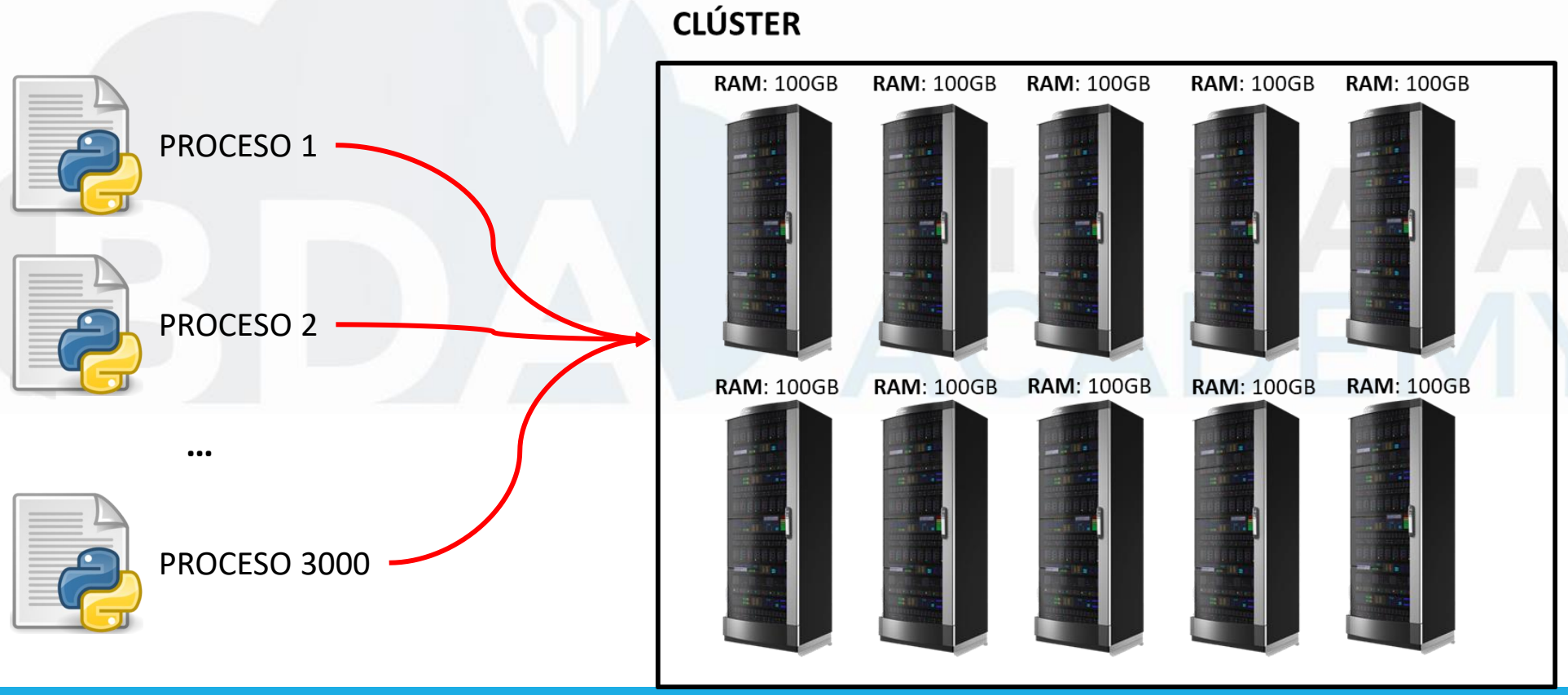
Adaptar el proceso para que funcione en un clúster toma mucho tiempo

Después de 5 semanas: Procesamiento sobre clúster



La realidad empresarial

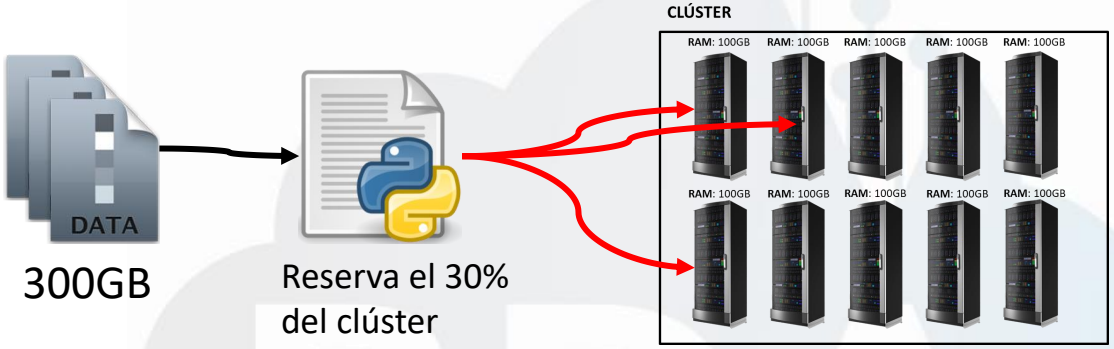
En la empresa hay miles de procesos, mientras más volumetría procese cada proceso, en algún momento colapsarán y habrá que adaptarlos, **la adaptación toma tiempo (varias semanas), por lo tanto los procesos no son escalables y varios procesos de negocio no funcionarán por varias semanas.**



Clústers de Big Data

Definición del clúster ideal

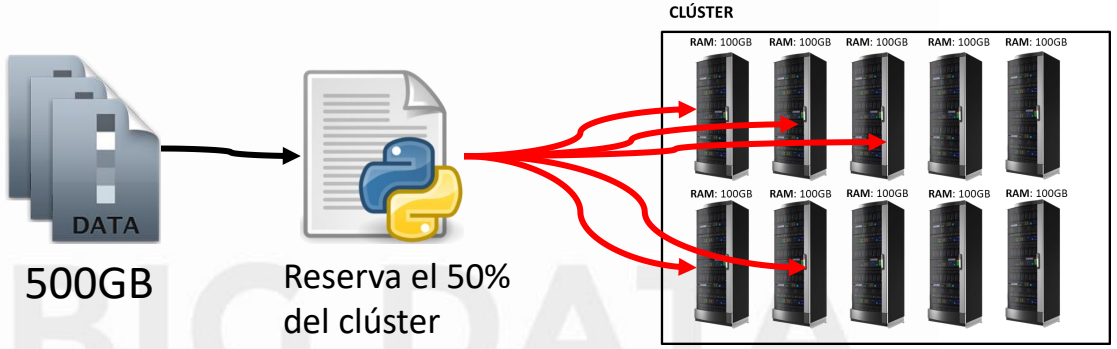
DÍA 1



300GB

Reserva el 30% del clúster

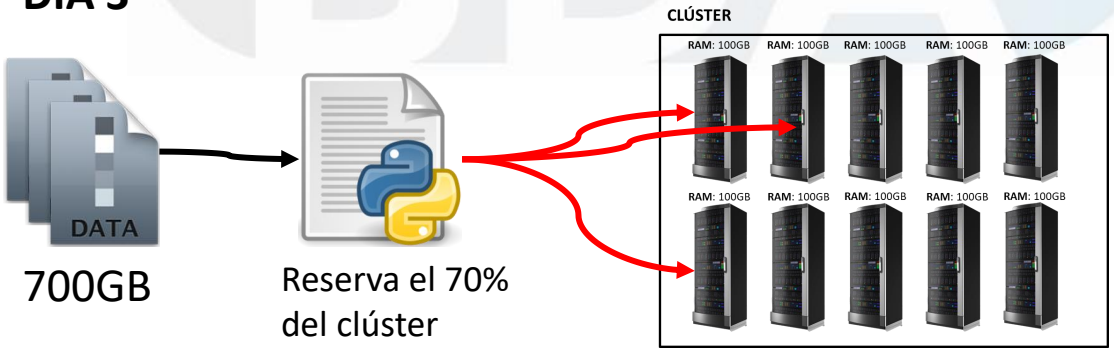
DÍA 2



500GB

Reserva el 50% del clúster

DÍA 3

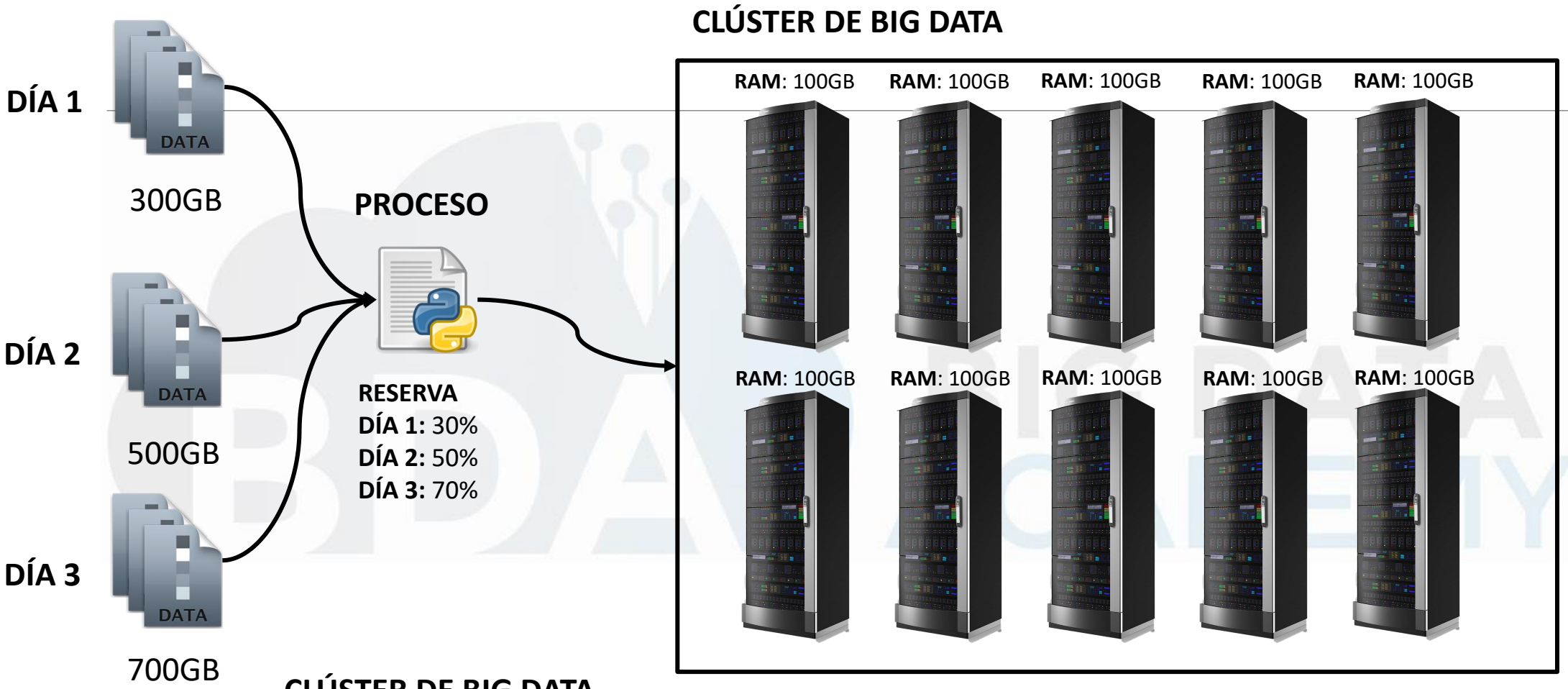


700GB

Reserva el 70% del clúster

Si aumenta la volumetría, aumentamos la reserva del clúster, el clúster reserva los servidores y **distribuye la carga de trabajo del código de manera automática**

Clúster de Big Data



CLÚSTER DE BIG DATA

- **Reserva la potencia** del clúster según la volumetría
- **Distribuye de manera automática** la carga de trabajo sobre los servidores reservador

Concepto de Big Data

Es un marco de trabajo (conceptos + tecnologías) que permite implementar **procesos escalables** para **procesar grandes volúmenes** de datos

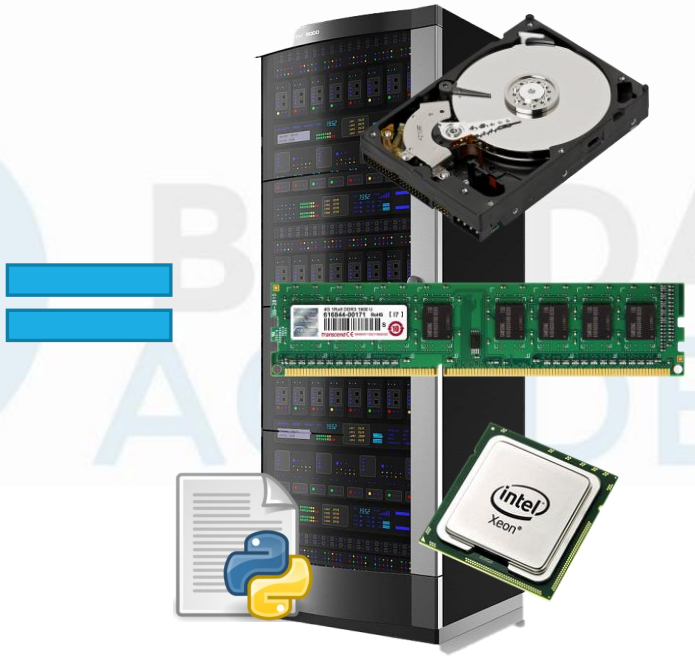
Hadoop como ecosistema tecnológico estándar de Big Data

El clúster de Big Data como un “súper-servidor”

CLÚSTER DE BIG DATA

10 servidores de 10TB de Disco, 100GB de RAM y 40 Núcleos de CPU

SÚPER SERVIDOR



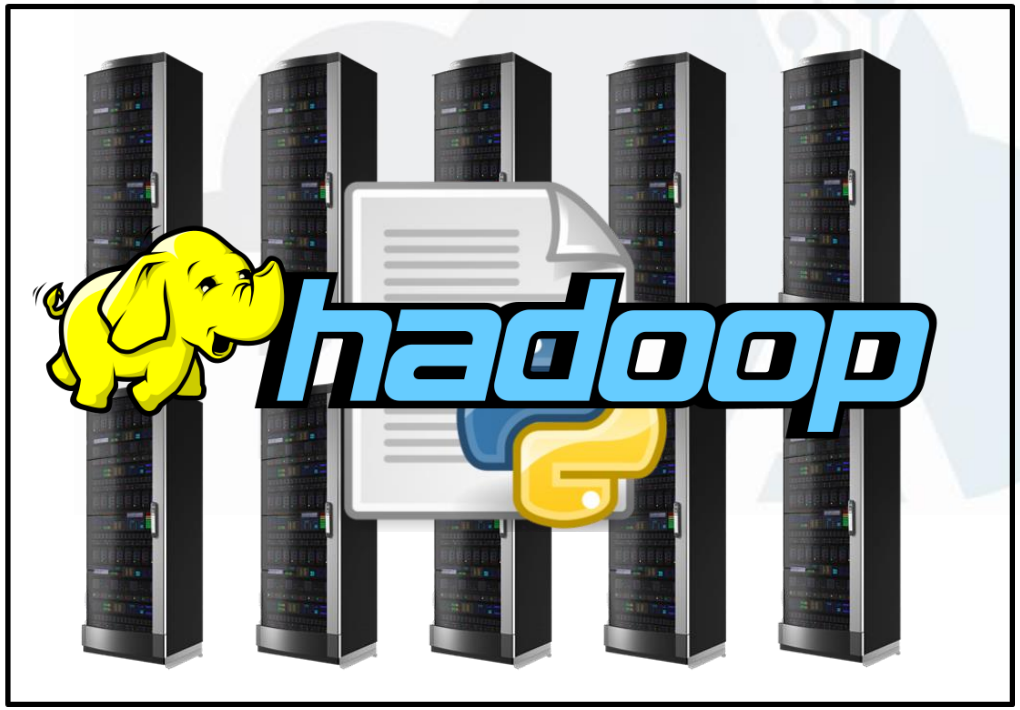
POTENCIA
DISCO: 100 TB
RAM: 1000 GB
CPU: 400 NÚCLEOS

Los desarrolladores ven al clúster como 1 “súper-servidor de gran capacidad”, ¿qué tecnología permite esto?

Hadoop

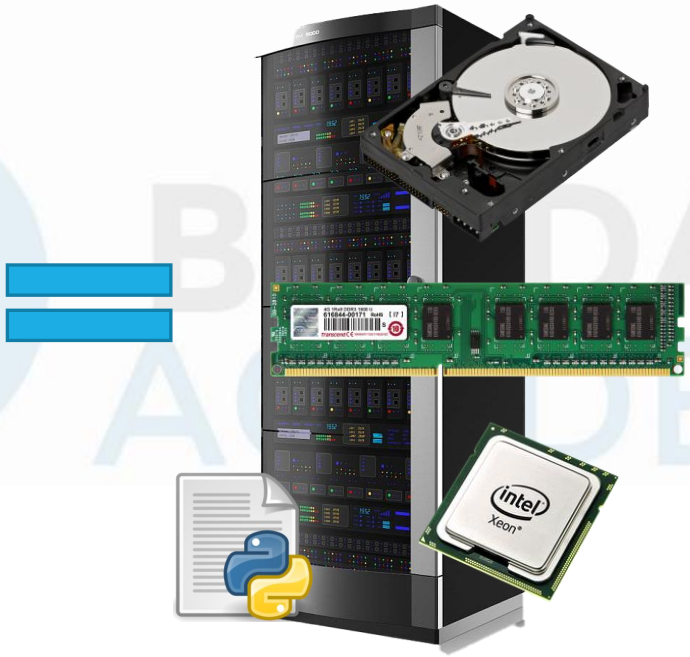
Hadoop es una tecnología de gestión de clústers de Big Data y tiene tres módulos principales

CLÚSTER DE BIG DATA



En todos los servidores del clúster se debe instalar Hadoop

SÚPER SERVIDOR



HDFS [Hadoop Distributed File System]
Gestiona los discos duros del clúster

YARN [Yet Another Resource Negotiator]
Reserva la RAM y las CPU del clúster

MAPREDUCE ENGINE [Motor de procesamiento]
Ejecuta el proceso sobre la RAM y CPU reservadas

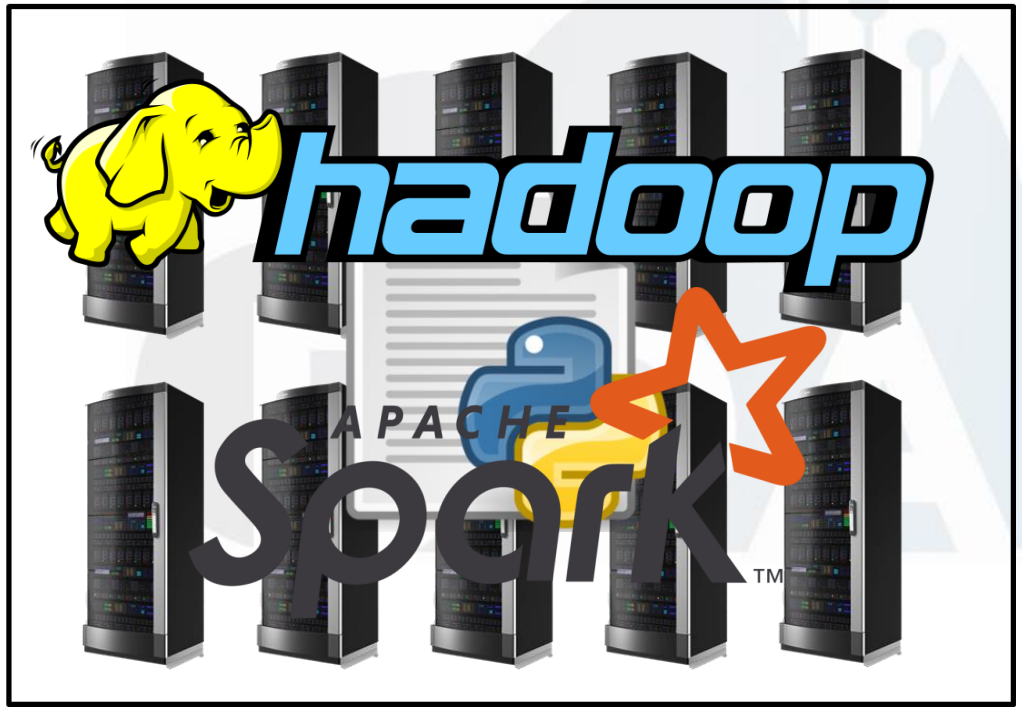
Spark como motor de procesamiento

SPARK como remplazo de MAPREDUCE

Es muy antiguo y está orientado sólo para procesos del tipo BATCH ESTRUCTURADOS (no permite muchos otros tipos de procesamiento como REAL-TIME, ANALÍTICOS, SEMI-ESTRUCTURADOS, etc)

Ecosistema estándar de Big Data: Hadoop + Spark

CLÚSTER DE BIG DATA



En todos los servidores del clúster se debe instalar Hadoop

=

SÚPER SERVIDOR

HDFS [Hadoop Distributed File System]
Gestiona los discos duros del clúster

YARN [Yet Another Resource Negotiator]
Reserva la RAM y las CPU del clúster

SPARK [Motor de procesamiento]
Ejecuta el proceso sobre la RAM y CPU reservadas