

Andres Felipe Rivera Diaz
Julian David Cuellar
Jose David Santa
Brayan Stiven Tigreros

Proyecto Final Infraestructura

ÍNDICE



1. INTRODUCCION

2. ANALISIS

3. DISEÑO

4. IMPLEMENTACION

5. CONCLUSION

INTRODUCCIÓN



Este proyecto integra dos partes clave:

- **Empaque y despliegue de una API usando Docker**
 - **Procesamiento distribuido de datos con Apache Spark**
-

Objetivo: Crear una solución eficiente y escalable que combine contenedores y análisis de datos. Se aplican herramientas modernas para simular un entorno real de infraestructura tecnológica.

DATASET SELECCIONADO

ANALISIS

ESTE DATASET CONTIENE INFORMACIÓN DETALLADA SOBRE LAPTOPS DISPONIBLES PARA LA VENTA, INCLUYENDO ESPECIFICACIONES TÉCNICAS Y PRECIOS FINALES [1]. ESTÁ DISEÑADO PARA ANÁLISIS DE PRODUCTOS TECNOLÓGICOS EN EL MERCADO.

PROPOSITO

PARA RESOLVER PROBLEMÁTICAS COMO:

- ¿CUÁNTAS LAPTOPS DE DETERMINADA MARCA HAY EN EL DATASET?
- ¿CUÁLES SON LAS LAPTOPS QUE TIENE UNA PANTALLA TÁCTIL?
- ¿CUÁL ES EL PROMEDIO DEL PRECIO DE UNA MARCA DETERMINADA DE LAPTOPS?

Información de cada columna

- Status: Estado del producto (Nuevo, Usado o Renovado)
- Brand: Nombre del fabricante (Asus, Acer, HP, Lenovb, MSI, Toshiba, etc)
- Model: Modelo específico del producto
- CPU: Tipo y modelo de procesador
- RAM: Cantidad de memoria RAM
- Storage: Capacidad de almacenamiento
- Storage Type: Tipo de almacenamiento
- GPU: Tarjeta Gráfica de la laptop
- Screen: Tamaño de la pantalla en pulgadas
- Touch: Indica si la pantalla es táctil ("Yes" or "No")
- Final Price: Precio de la laptop

Tamaño y estructura del Dataset

Número de filas: 12.099

Número de columnas: 11

ALTERNATIVAS DE DESPLIEGUE



Opciones para empaquetado y despliegue:

- Docker
 - Aislamiento y portabilidad
 - Requiere familiarización
- Maquina virtual de Gestión Manual
 - Configuras todo según tus necesidades
 - Difícil de replicar en otras maquinas virtuales
- Maquina virtual con Imágenes Preconfiguradas
 - Configuración empaquetada en la imagen
 - Las imágenes pueden ser grandes y difíciles de transferir



DATA

SERVICIOS

GESTIONADOS

EN LA NUBE

Uso de Ansible, Puppet o Chef

-  Fácil de usar en múltiples maquinas
-  Curva de aprendizaje alta

Empaquetado no binario o standalone

-  No requiere dependencias adicionales
-  Problemas al ejecutarse en diferentes sistemas

ALTERNATIVAS DE DESPLIEGUE



Opciones para clúster de procesamiento distribuido:

- Apache Spark Standalone
 - Fácil de configurar, ligero
 - No es el más escalable
- Hadoop YARN
 - Escalable, bien integrado con HDFS
 - Configuración compleja, consumo alto de recursos
- Kubernetes
 - Alta disponibilidad, ideal para la nube
 - Requiere conocimientos avanzados y configuración detallada

DATA SERVICIOS GESTIONADOS EN LA NUBE

Amazon EMR

-  Configuración rápida, integración con AWS
-  Dependencia de AWS, puede ser costoso

Google Cloud Dataproc

-  Escalable, integración con herramientas de Google
-  Dependencia de Google Cloud, costos similares a EMR

Azure HDInsight

-  Compatible con ecosistema Azure, soporta varios frameworks
-  Costoso, dependiente de Azure

Databricks

-  Interfaz amigable, potente para análisis y ML
-  Requiere licencia, elevado costo

Arquitectura del Sistema



La arquitectura completa que escogimos para este sistema para la parte del empaquetado y despliegue de la aplicación en contenedores fue el empaquetado con Docker ya que según sus pros y contras es la que mejor se adapta a la aplicación

Para nuestra aplicación de procesamiento de datos distribuidos hemos escogido apache Spark Standalone Cluster ya que Spark procesa y conserva los datos en la memoria RAM, sin escribir ni leer en el disco, lo que da como resultado velocidades de procesamiento mucho más rápidas, aparte de su de implementación y uso en máquinas virtuales como lo aprendimos es bastante fácil y cómodo

Pipeline del Sistema

Diagrama de Componentes

- HTML
- Docker y Docker Swarm
- CSV
- Spark DF Reader
- Spark SQL
- Spark Standalone Cluster
- CSV de Salida
- Visualizacion

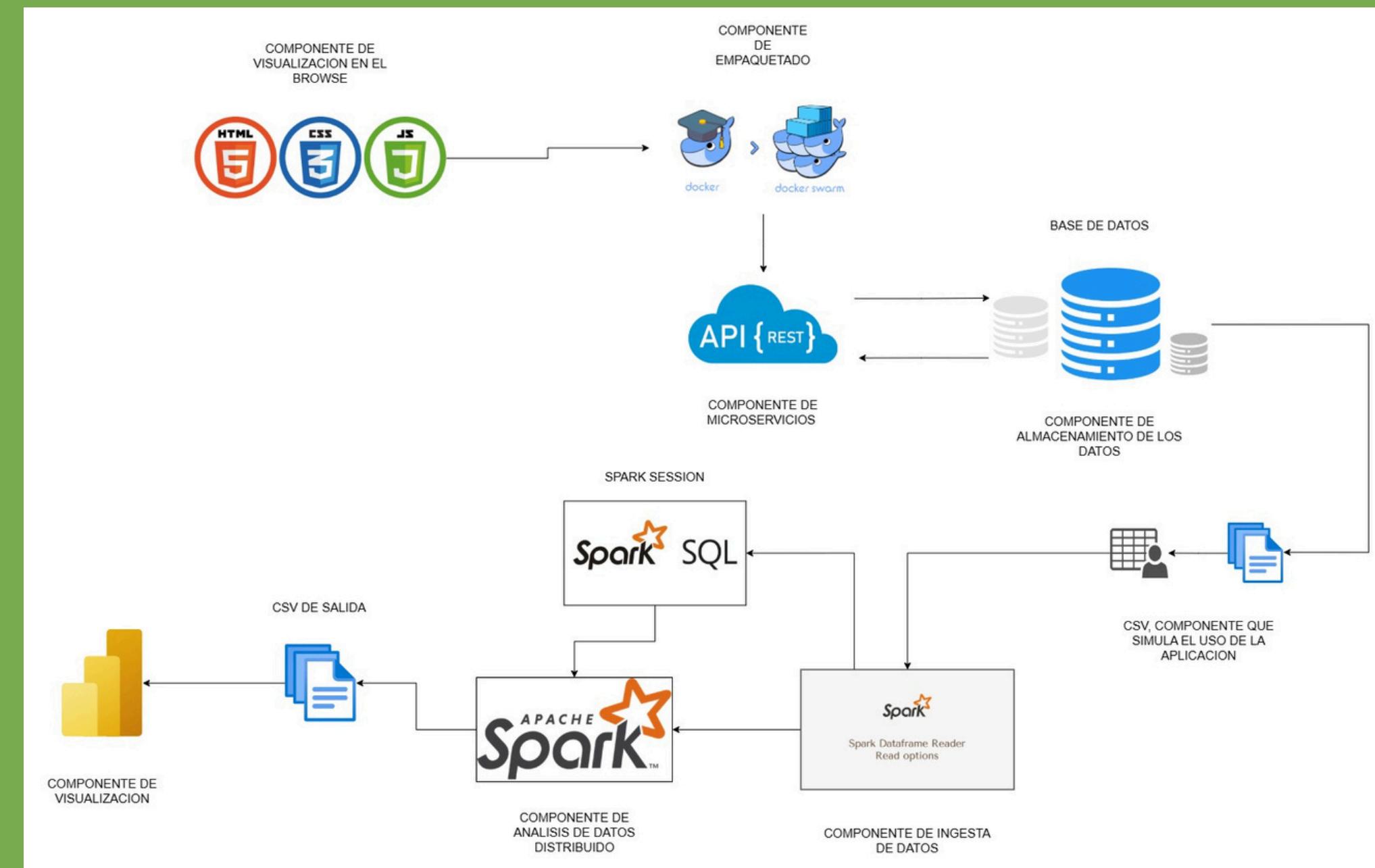
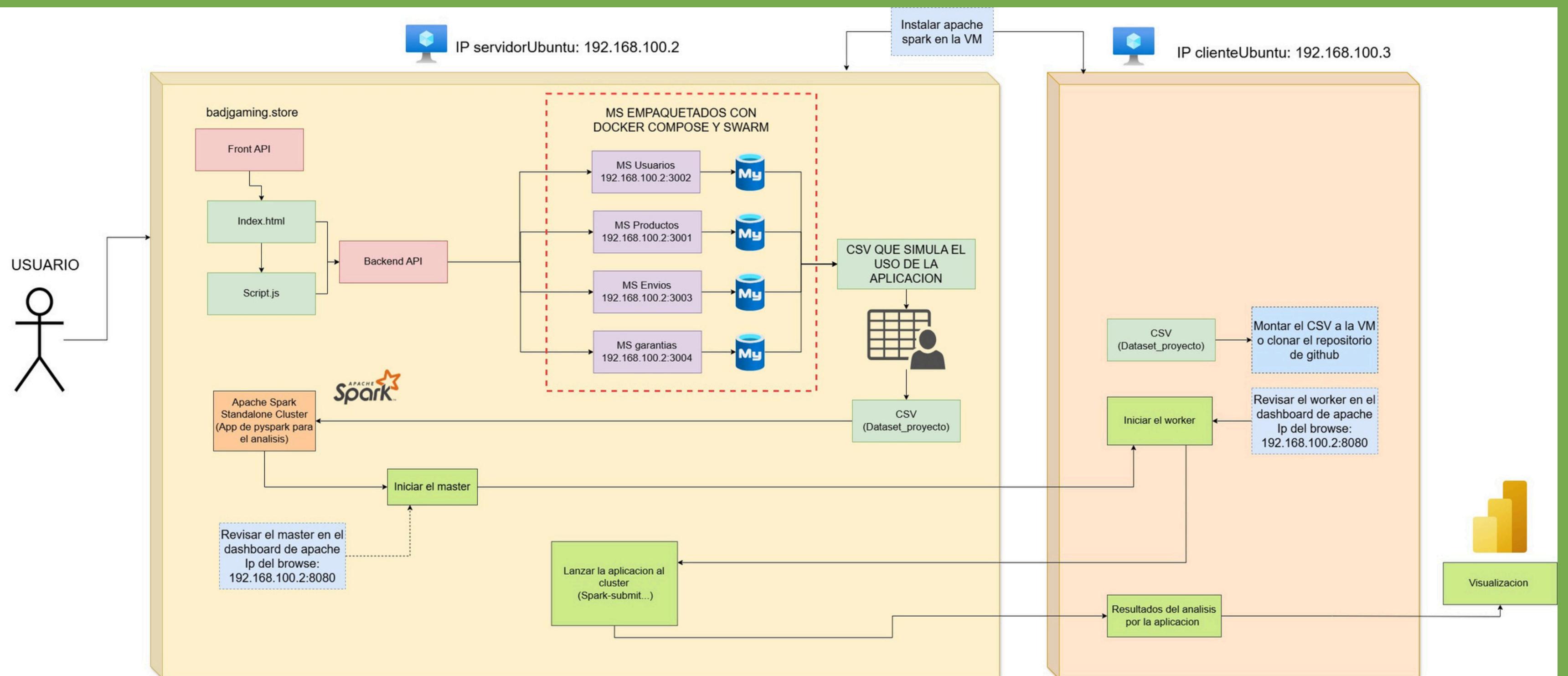


Diagrama de Despliegue



BIG DATA

EMPAQUETADO DOCKER

```
vagrant@servidorUbuntu:~/proyecto-REI$ sudo docker-compose ps
```

Name	Command	State	Ports
envios_ms	docker-entrypoint.sh node ...	Up	0.0.0.0:3003->3003/tcp,:::3003->3003/tcp
garantias_ms	docker-entrypoint.sh node ...	Up	0.0.0.0:3004->3004/tcp,:::3004->3004/tcp
productos_ms	docker-entrypoint.sh node ...	Up	0.0.0.0:3001->3001/tcp,:::3001->3001/tcp
proyecto_mysql	docker-entrypoint.sh mysqld	Up (healthy)	0.0.0.0:3306->3306/tcp,:::3306->3306/tcp, 33060/tcp
usuarios_ms	docker-entrypoint.sh node ...	Up	0.0.0.0:3002->3002/tcp,:::3002->3002/tcp

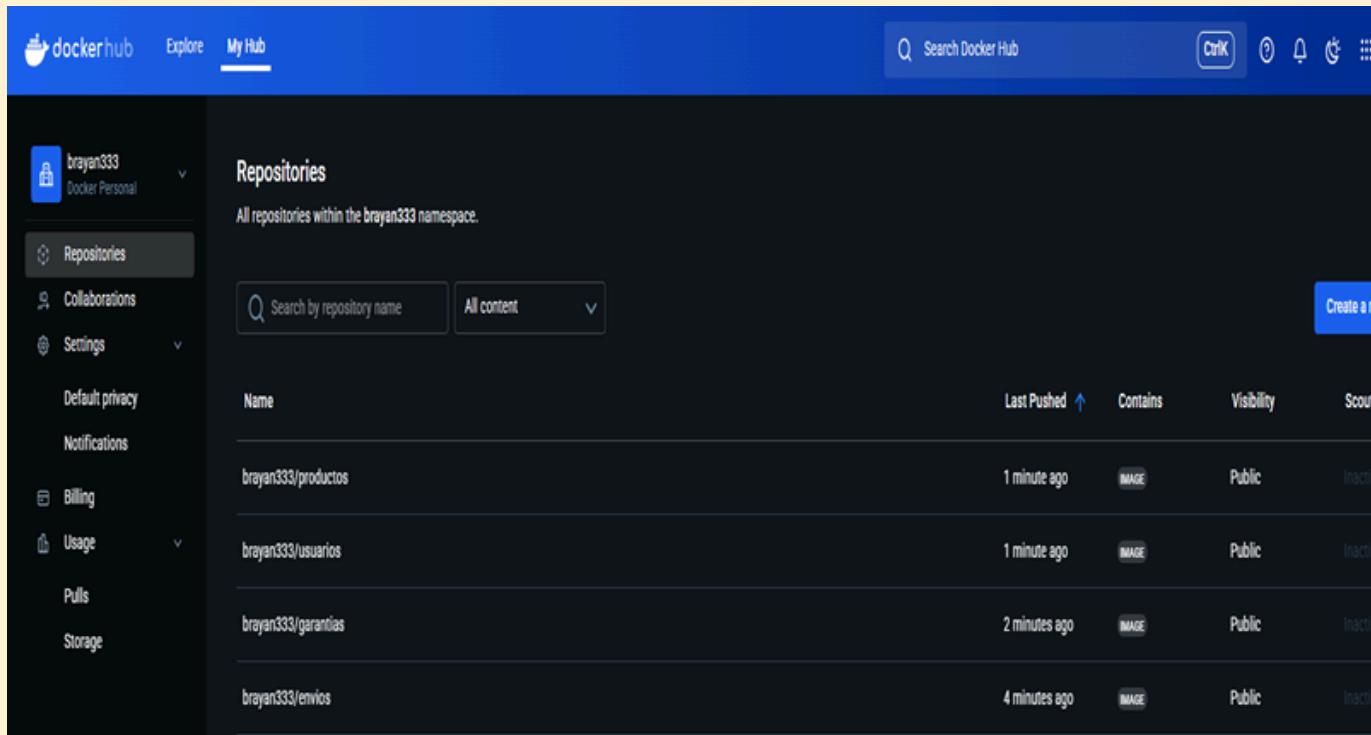
```
vagrant@servidorUbuntu:~/proyecto-REI$ sudo docker stack services proyecto-rei
```

ID	NAME	MODE	REPLICAS	IMAGE	PORTS
b2qmefy1mkbp	proyecto-rei_envios	replicated	1/1	brayan333/envios:latest	*:3003->3003/tcp
qoiq8qsnge2u	proyecto-rei_garantias	replicated	1/1	brayan333/garantias:latest	*:3004->3004/tcp
l82b7ucd7g1l	proyecto-rei_mysql	replicated	1/1	mysql:8.0	*:3306->3306/tcp
gl4639nmeoqu	proyecto-rei_productos	replicated	1/1	brayan333/productos:latest	*:3001->3001/tcp
gftw4llxgvzx	proyecto-rei_usuarios	replicated	1/1	brayan333/usuarios:latest	*:3002->3002/tcp

```
vagrant@servidorUbuntu:~/proyecto-REI$ |
```

BIG DATA

EMPAQUETADO DOCKER



The screenshot shows the Docker Hub interface for a user named 'brayan333'. The left sidebar includes options like 'Repositories', 'Collaborations', 'Settings', 'Default privacy', 'Notifications', 'Billing', 'Usage', 'Pulls', and 'Storage'. The main area displays a table of repositories under the heading 'Repositories'. The table has columns for 'Name', 'Last Pushed', 'Contains', 'Visibility', and 'Scout'. Four repositories are listed:

Name	Last Pushed	Contains	Visibility	Scout
brayan333/productos	1 minute ago	IMAGE	Public	Inactive
brayan333/usuarios	1 minute ago	IMAGE	Public	Inactive
brayan333/garantias	2 minutes ago	IMAGE	Public	Inactive
brayan333/envios	4 minutes ago	IMAGE	Public	Inactive

```
vagrant@servidorUbuntu:~/proyecto-REI$ sudo docker service scale proyecto-rei_usuarios=3
proyecto-rei_usuarios scaled to 3
overall progress: 3 out of 3 tasks
1/3: running [=====>]
2/3: running [=====>]
3/3: running [=====>]
verify: Service proyecto-rei_usuarios converged
```

EMPAQUETADO DOCKER



```
Calle 123 #45-67"}]}vagrant@servidorUbuntu:~/proyecto-REI$ curl 192.168.100.2:3002/usuarios
[{"nombre":"Ana Garcia","email":"ana.garcia@example.com","usuario":"anag23","password":"12345","telefono":"3101234567","cedula":"123456789012","direccion":"
Calle 123 #45-67"}, {"nombre":"jose","email":"jose@example.com","usuario":"jose23","password":"54321","telefono":"3101555555","cedula":"12345","direccion":"Colombia"}]}vagrant@servidorUbuntu:~/proyecto-REI$ curl 192.168.100.2:3001/productos
```

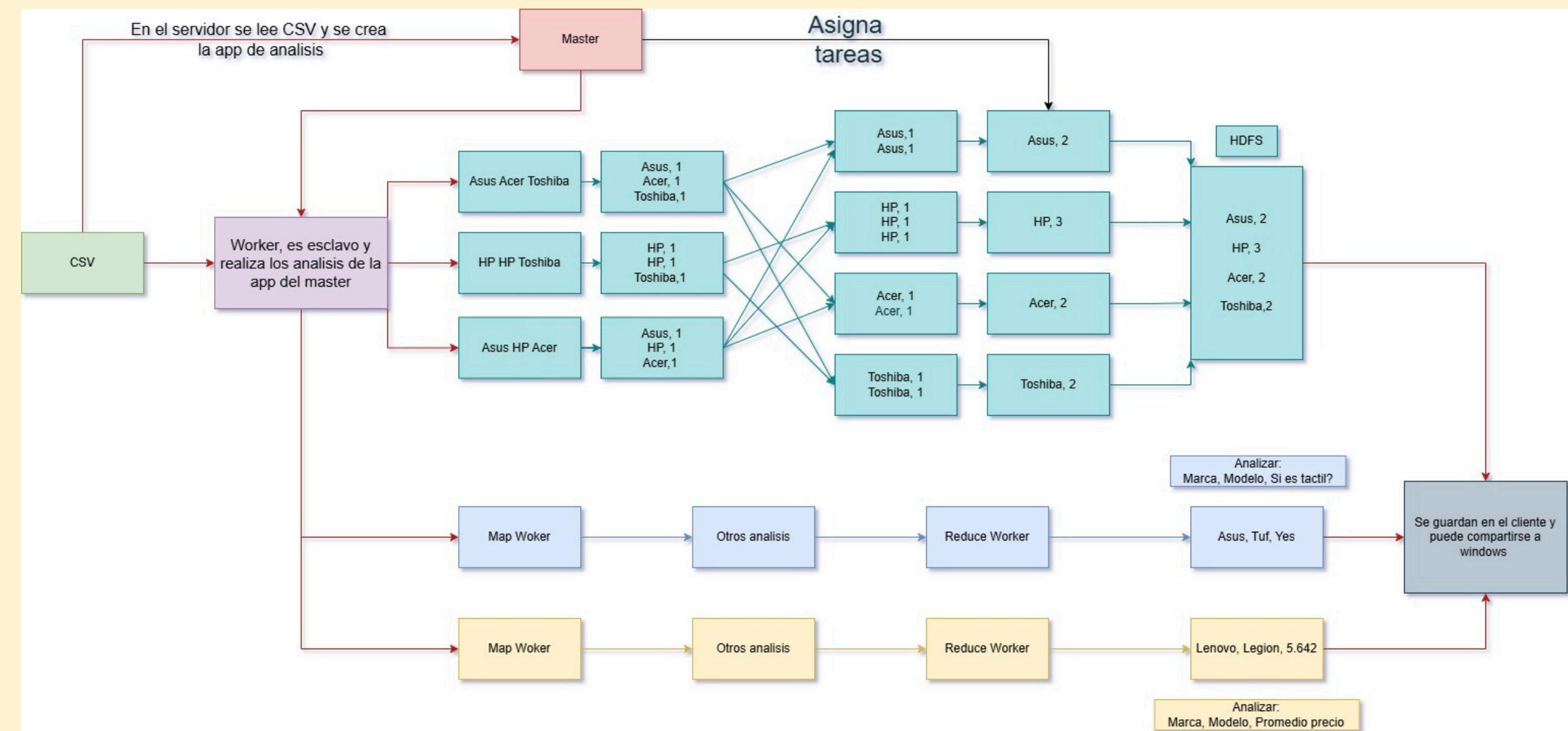
```
ombia"}]}vagrant@servidorUbuntu:~/proyecto-REI$ curl 192.168.100.2:3001/productos
[{"id":1,"nombre":"Laptop Gamer Pro","estado":"Nuevo","marca":"TechBrand","modelo":"X-5000","cpu":"Intel Core i9-12900H","ram":"32GB DDR5","gb_almacenamiento":"
"1000GB","tipo_almacenamiento":"SSD NVMe","gpu":"NVIDIA RTX 3080","pantalla":"15.6\" 4K OLED","es_tactil":"No","precio":1999.99,"garantia":"24 meses","stock":50}, {"id":2,"nombre":"MSI RAPTOR","estado":"Nuevo","marca":"TechBrand","modelo":"X-5000","cpu":"Intel Core i9-12900H","ram":"32GB DDR5","gb_almacenamiento":"
"1000GB","tipo_almacenamiento":"SSD NVMe","gpu":"NVIDIA RTX 3080","pantalla":"15.6\" 4K OLED","es_tactil":"No","precio":5000.99,"garantia":"24 meses","stock":5}]}vagrant@servidorUbuntu:~/proyecto-REI$ |
```

BIG DATA

DESARROLLO DEL CLUSTER

Apache Spark

Como funciona la app de análisis



DESARROLLO DEL CLUSTER

Apache Spark



E7. 2023-03 Cluster Spark.pdf X Spark Master en spark://192.168.1.11 X +

← C ▲ No seguro 192.168.100.2:8080 vagrant@clienteUbuntu:~/labSpark\$ cd resultado_analisis vagrant@clienteUbuntu:~/labSpark/resultado_analisis\$ ls

Spark 3.5.5 Spark Master en spark://192.168.100.2:7077 contar_marca laptops_Premium promedio_precio_marca promedio_tactiles

Dirección URL: spark://192.168.100.2:7077 vagrant@clienteUbuntu:~/labSpark/resultado_analisis\$ ls

Trabajadores vivos: 1 contar_marca laptops_Premium promedio_precio_marca promedio_tactiles

Núcleos en uso: 2 en total, 0 Usados vagrant@clienteUbuntu:~/labSpark/resultado_analisis\$ cd contar_marca

Memoria en uso: 1024.0 MiB en total, 0.0 B Usado vagrant@clienteUbuntu:~/labSpark/resultado_analisis\$ ls _temporary

Recursos en uso:

Aplicaciones: 0 en ejecución, 1 Completado vagrant@clienteUbuntu:~/labSpark/resultado_analisis\$ cd _temporary

Controladores: 0 en funcionamiento, 0 Completado vagrant@clienteUbuntu:~/labSpark/resultado_analisis\$ ls 0

Estado: VIVO vagrant@clienteUbuntu:~/labSpark/resultado_analisis\$ cd 0

▼ Trabajadores (1) vagrant@clienteUbuntu:~/labSpark/resultado_analisis\$ vagrant@clienteUbuntu:~/labSpark/resultado_analisis\$ ls task_202505200039334285946117170621337_0004_m_000000 _temporary

Identificación del trabajador	Dirección	Estado	Corazones	Memoria	Recursos
trabajador-20250520003311-192.168.100.3:42105	192.168.100.3:42105	VIVO	2 (0 usados)	1024.0 MiB (0.0 B utilizados)	

▼ Aplicaciones en ejecución (0) vagrant@clienteUbuntu:~/labSpark/resultado_analisis\$ cat part-00000-3a55044f-1dd0-4052-bf54-512f96368dd4-c000.csv

Id. de aplicación	Nombre	Corazones	Memoria por ejecutor	Recursos por albacete	Tiempo de envío	Usuario	Estado	Duración
aplicación-20250520003913-0000	App_analisis_distribuido	2	1024.0 MiB		2025/05/20 00:39:13	vagabundo	TERMINADO	26 s

▼ Solicitudes completadas (1) vagrant@clienteUbuntu:~/labSpark/resultado_analisis\$

Id. de aplicación	Nombre	Corazones	Memoria por ejecutor	Recursos por albacete	Tiempo de envío	Usuario	Estado	Duración
aplicación-20250520003913-0000	App_analisis_distribuido	2	1024.0 MiB		2025/05/20 00:39:13	vagabundo	TERMINADO	26 s

vagrant@clienteUbuntu:~/labSpark/resultado_analisis\$ cat part-00000-3a55044f-1dd0-4052-bf54-512f96368dd4-c000.csv

Brand,count
Razer,1049
Millenium,2
Realme,1
Medion,31
HP,1386
Dell,1143
Jetwing,1
Primux,8
Dynabook Toshiba,19
Acer,1125
Asus,1379
Deep Gaming,8
Lenovo,1357
Vant,6
Samsung,1027
Thomson,4
LG,31

DESARROLLO DEL CLUSTER

Apache Spark



```
vagrant@clienteUbuntu:~$  
vagrant@clienteUbuntu:~$ cd labSpark  
vagrant@clienteUbuntu:~/LabSpark$ ls  
analisis_distribuido dataset population resultado_analisis resultsCluster resultsPopulation spark-3.5.5-bin-hadoop3 spark-3.5.5-bin-hadoop3.tgz  
vagrant@clienteUbuntu:~/LabSpark$ cd resultado_analisis  
vagrant@clienteUbuntu:~/LabSpark/resultado_analisis$ ls  
contar_marca gpu_rtx_laptops laptops_mas_grandes laptops_Premium promedio_precio_marca promedio_tactiles  
vagrant@clienteUbuntu:~/LabSpark/resultado_analisis$ |
```

En la maquina cliente se guardaron 6 directorios cada uno con los datos analizados



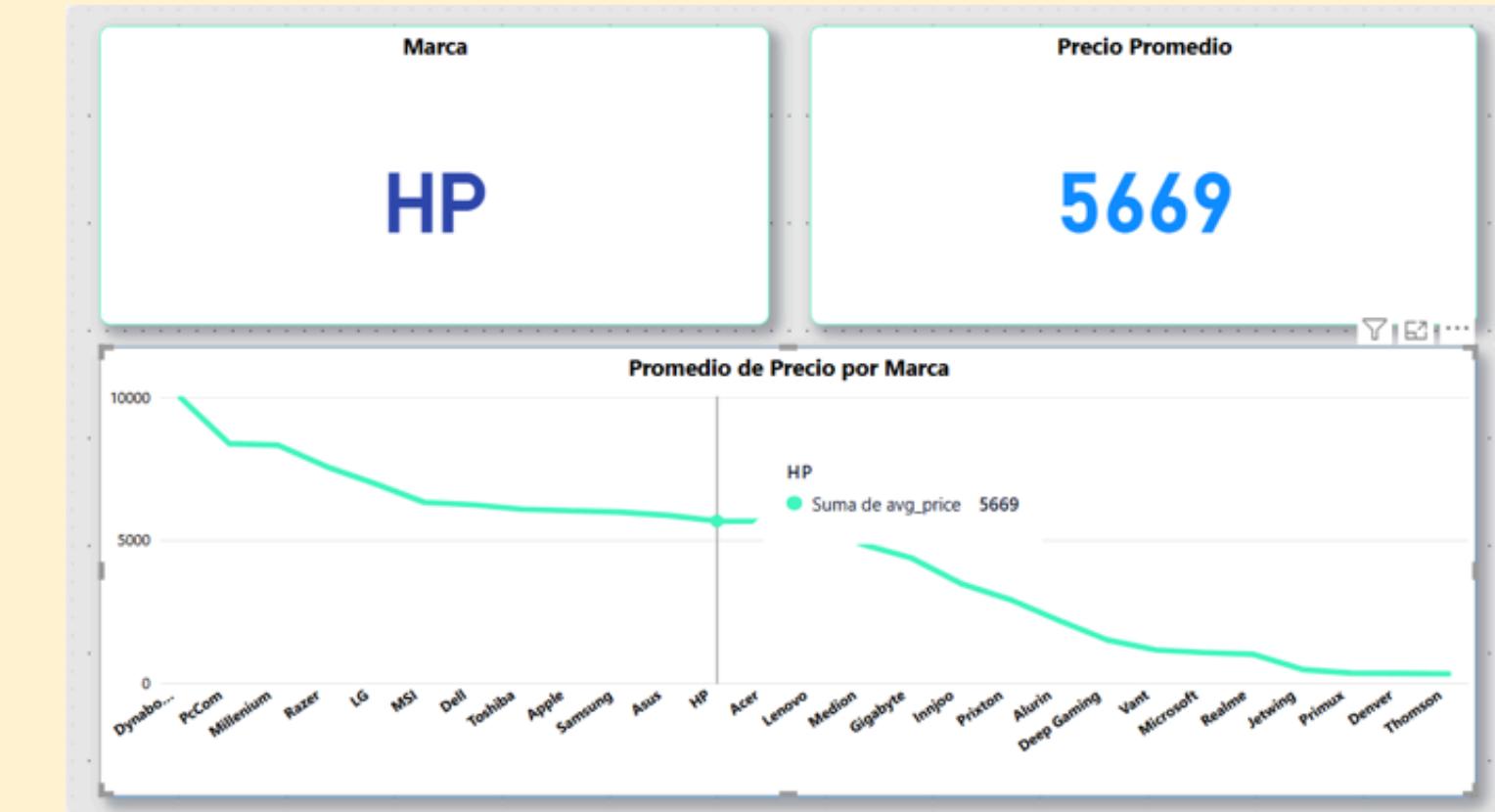
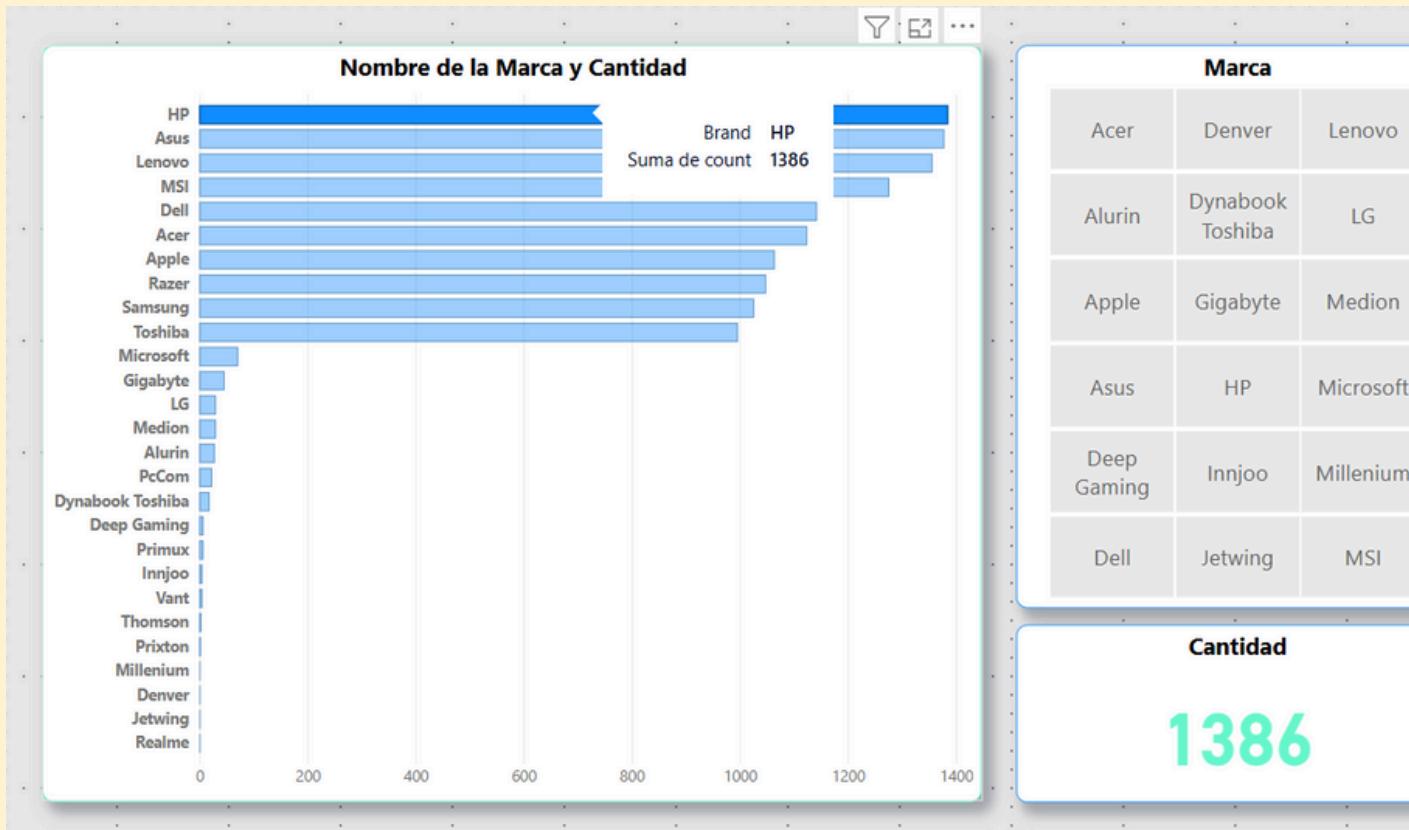
DESARROLLO DEL CLUSTER

Apache Spark

Nombre	Fecha de modificación	Tipo	Tamaño
📁 .vagrant	17/02/2025 10:34 a. m.	Carpeta de archivos	
📁 2019-Dec	2/05/2025 7:18 p. m.	zip	75.546 KB
📁 laptops_csv_REI	19/05/2025 7:01 p. m.	zip	119 KB
📄 mynew.box	18/02/2025 10:14 p. m.	Archivo BOX	1.036.004 KB
✖️ part-00000-1c5cf02e-22a7-479a-ad00-e7...	19/05/2025 8:05 p. m.	Archivo de valores...	245 KB
✖️ part-00000-3e21bf72-72af-4bc8-b4fc-4e2...	19/05/2025 8:05 p. m.	Archivo de valores...	24 KB
✖️ part-00000-9a0c009c-a435-4079-aa20-d5...	19/05/2025 8:05 p. m.	Archivo de valores...	55 KB
✖️ part-00000-3555db01-da26-4cab-aeed-7...	19/05/2025 8:05 p. m.	Archivo de valores...	1 KB
✖️ part-00000-5105e26a-f5c7-49cd-9aae-2b...	19/05/2025 8:05 p. m.	Archivo de valores...	1 KB
✖️ part-00000-bb2e493e-9f4e-411f-9f9f-90b...	19/05/2025 8:05 p. m.	Archivo de valores...	1 KB
📄 Vagrantfile	17/02/2025 10:53 a. m.	Archivo	1 KB
📄 Vagrantfile	17/02/2025 10:53 a. m.	Documento de te...	1 KB
📁 world_population	2/05/2025 9:21 p. m.	zip	16 KB

Podemos guardarlos en la maquina anfitriona por medio del directorio compartido de vagrant

REPORTES DE INTERES





DATA

CONCLUSION

Este proyecto permitió reforzar todos los conocimientos adquiridos en el curso

- Permitiéndonos conocer y evaluar alternativas para empaquetado y análisis distribuido
- Crear diagramas para reforzar el entendimiento de como funcionan y como se relacionan componentes
- Llevar a cabo implementaciones en varios enfoques y con herramientas vistas a lo largo de la materia

En general el aprendizaje de estos conocimientos nos ayuda a tener nuevas con las que contar a lo largo de nuestra carrera profesional