**HITO GRUPAL PROGRAMACIÓN**

Alejandro Garcia Barbero y Jose Corral Muñoz

INDICE

* **Esta fase es una explicación “teórica” de cuatro puntos fundamentales**
* **En esta segunda fase se realiza la implementación de la investigación. En concreto sería acceder a un volumen de datos y mostrarlo. Podríamos utilizar Scala o Python y mostrar el resultado en PowerBI o Tableau. La idea es que sea algo muy impactante por la calidad de contenido tratado, velocidad de acceso, volumen de datos....**
* **Para finalizar, realizamos una evaluación o consideraciones de cómo han evolucionado el acceso a datos en los últimos años. Desde acceso a ficheros, pasando por base de datos y consumiendo APIs.**

**FASE 1**

**Esta fase es una explicación “teórica” de cuatro puntos fundamentales:**

**1. Hablamos de fuentes de datos. De grandes volúmenes de datos. Por ejemplo, data Lake o similar. También es importante tratar la diferencia entre datos estructurados y no estructurados en relación al bigData.**

**2. Entre las herramientas más interesantes a la hora de gestionar grandes volúmenes de datos nos encontramos con Hadoop y Spark. Habría que tratar sus características y finalidad.**

**3. Existen lenguajes de programación “recomendables” para gestionar datos. Entre ellos, están Python y Scala. Sería explicar brevemente por qué.**

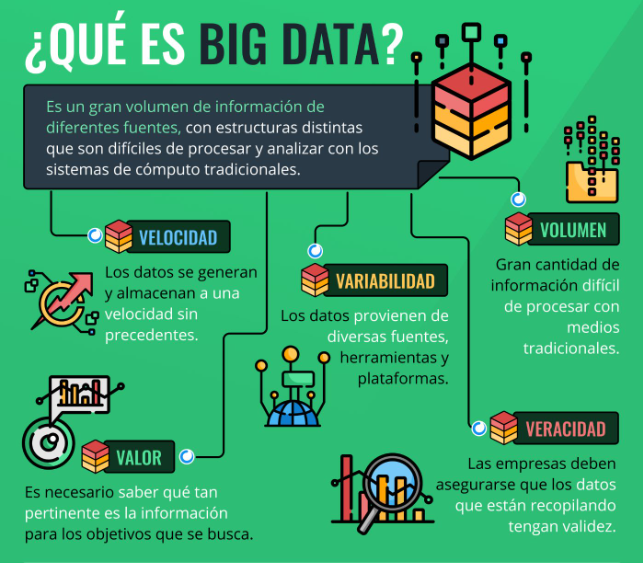
¿Cómo tratar un gran volumen de datos?

* **Big Data**

 Big Data es un término que describe el gran volumen de datos, tanto estructurados como no estructurados, que inundan los negocios cada día.

**Lo que importa con el Big Data es lo que las organizaciones hacen con los datos**. **Big Data se puede analizar para obtener ideas** que conduzcan a mejores decisiones y movimientos de negocios estratégicos.

**Se define como, conjuntos de datos o combinaciones de conjuntos de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales**, tales como base de datos relacionales y estadísticas convencionales o paquetes de visualización, dentro del tiempo necesario para que sean útiles.



**El análisis de Big Data ayuda a las organizaciones a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades**.

* **Reducción de coste**.
* **Mejor toma de decisiones**.
* **Nuevos productos y servicios**. la capacidad de medir las necesidades de los clientes y la satisfacción

* **Data Lake**

Un **data lake es un entorno de datos compartidos en su formato original que comprende múltiples repositorios y aprovecha las tecnologías de big data**

**Un data lake es capaz de proporcionar datos a la organización para una gran variedad de procesos analiticos diferentes:**

* **Descubrimiento y exploración de datos**
* **Análisis ad hoc simple**
* **Análisis complejo para toma de decisiones**
* **Informes**
* **Análisis en tiempo real**

**es un repositorio de almacenamiento que contienen una gran cantidad de datos en bruto y que se mantienen allí hasta que sea necesario**

**El principal beneficio de un data lake es la centralización de fuentes de contenido dispares.**

.

* **Datos estructurados y no estructurados, y su relación en big data**

Los **datos estructurados** son modelos de **datos** predefinidos, generalmente solo texto y fáciles de buscar y analizar, mientras que los **datos** no **estructurados** no son modelos de **datos** predefinidos, pueden venir en texto, imágenes, sonido, vídeos u otros formatos, y su búsqueda y análisis es más difícil.

Son datos estructurados, por ejemplo:

* Hoja de Excel
* Bases de datos relacionales o SQL
* Formularios web
* Fichas de clientes estandarizadas

Son ejemplos de datos no estructurados:

* Los archivos de imágenes
* Los archivos de audio
* Los PDF
* Los datos de redes sociales
* La mensajería instantánea
* Los datos de geolocalización

2.- Hadoop y Spark características y finalidad.

**Hadoop** es un proyecto de código abierto. Se trata de un framework de software que permite programar aplicaciones distribuidas, es decir, aplicaciones capaces de trabajar con enormes cantidades de nodos en red y de datos.

Este se basa en distribuir los datos y paralelizar el procesamiento de los mismos, de modo que cada nodo de una red de máquinas se encargue de procesar una parte de la totalidad de los datos durante el procesamiento del big data. A este método se le conoce como procesamiento distribuido o computación distribuida, que es, básicamente, distribuir las cargas de trabajo entre varios nodos de una misma red computacional.

Es útil para:

* **Entender mejor a sus clientes.**
* **Optimizar los procesos de negocio.**
* **Mejorar la productividad del personal y de las maquinarias.**

**Apache Spark está especialmente diseñado para su implementación en big data y machine learning.** Pues su potencia de procesamiento agiliza la detección de patrones en los datos, la clasificación organizada de la información, la ejecución de cómputo intensivo sobre los datos y el procesamiento paralelo en clústers.

Al ser código abierto puede ser modificado para crear versiones personalizadas dirigidas a problemas específicos o industriales.

**Apache Spark surge como una evolución de lo que es Hadoop para optimizar los procesos de big data.**

Python:

*-Código Simple:*

Una de las principales ventajas de usar Python para el procesamiento de big data es que es una solución amigable con el código que le permite desarrollar soluciones usando menos líneas de código que otros lenguajes de programación. El hecho de que no existen límites para el procesamiento de la información, lo que permite realizarlo en todo tipo de dispositivos y entornos, ya sea en el escritorio o en la nube., y el hecho de no tener una limitación en el procesamiento de información, lo que permite que sea ejecutado en todo tipo de dispositivos y entornos, tanto de escritorio como en la nube.

*-Velocidad en el procesamiento de datos:*

Actualmente, la programación en Python le permite desarrollar comandos y prototipos que facilitan la creación y ejecución de código de manera rápida y eficiente. Todo ello manteniendo la máxima transparencia entre el código desarrollado y el proceso. Esta es una de las principales razones por las que Python ya es considerado uno de los lenguajes de desarrollo de software más populares y destaca por su rendimiento y alta velocidad de escritura y ejecución.

*-Buen soporte de biblioteca:*

Por otro lado, Python incluye una gran cantidad de bibliotecas que lo convierten en una herramienta útil en varios campos.

*-Compatibilidad con diversos entornos:*

Su estructura y filosofía de código abierto hacen que Python sea compatible con muchas plataformas y pueda ejecutarse en diferentes sistemas operativos como Windows o Linux.

*-Aprendizaje rápido:*

Es más fácil que usar otros lenguajes de programación, por lo que es una de las mejores opciones tanto para principiantes como para usuarios avanzados.

*-Visualización sin límites:*

Otro factor que lleva a muchos usuarios a usar Python para big data es su capacidad para visualizar datos.

*-Codificación simple:*

La programación en Python implica menos líneas de código que otros lenguajes de programación. Es capaz de ejecutar programas con líneas mínimas de código. Además, Python proporciona automáticamente ayuda para identificar y vincular tipos de datos.

*-Código abierto:*

Desarrollado con un modelo comunitario, Python es un lenguaje de programación de código abierto. Como lenguaje de código abierto, Python admite una variedad de plataformas. También puede ejecutarse en diferentes entornos, como Windows y Linux.

*-Alcance:*

Python permite a los usuarios simplificar las operaciones de datos. Como Python es un lenguaje orientado a objetos, admite estructuras de datos avanzadas. Algunas de las estructuras de datos que maneja Python incluyen listas, conjuntos, tuplas, diccionarios y muchos más.

Scale:

-*Es un lenguaje de programación funcional:*

Este lenguaje de programación es funcional. Esto en el sentido de que cada función es un valor. Scala proporciona una sencilla sintaxis para que puedas dar con la definición de algunas funciones anónimas. Además, está configurado para soportar algunas funciones de orden superior.

*-Es un lenguaje de programación extensible:*

En la práctica, el desarrollo de aplicaciones de dominio específico a menudo requiere extensiones de lenguaje de dominio específico. Es por esto que el lenguaje de programación Scala proporciona una combinación única de mecanismos de lenguaje que hacen que sea simple añadir nuevos desarrollos en forma de bibliotecas.

*-El lenguaje de programación Scala es relativamente fácil de aprender:*

A un programador de Java le resulta difícil aprender cualquier lenguaje funcional. Es por ello que el uso de Scala es sencillo. Esto debido a su funcionalidad orientada a objetos. Tiene una sintaxis limpia, buenas bibliotecas y una extensa documentación en línea.

*-Se trata de un lenguaje de programación escrito estáticamente:*

Un lenguaje estáticamente escrito, evita errores en el código y ayuda a los desarrolladores a programar de forma adecuada. Además, puede ayudar a depurar el código de forma sencilla. En los lenguajes dinámicos, los errores solo son visibles cuando se ejecuta un programa o compilador. Es por esto que el uso del lenguaje de programación Scala ofrece lo mejor de los lenguajes estáticos y dinámicos.

**4. En la parte de visualización de datos, de mostrar dashboards nos encontramos con PowerBI y Tableau entre otros. Debemos explicar qué son.**

**Powerbi**

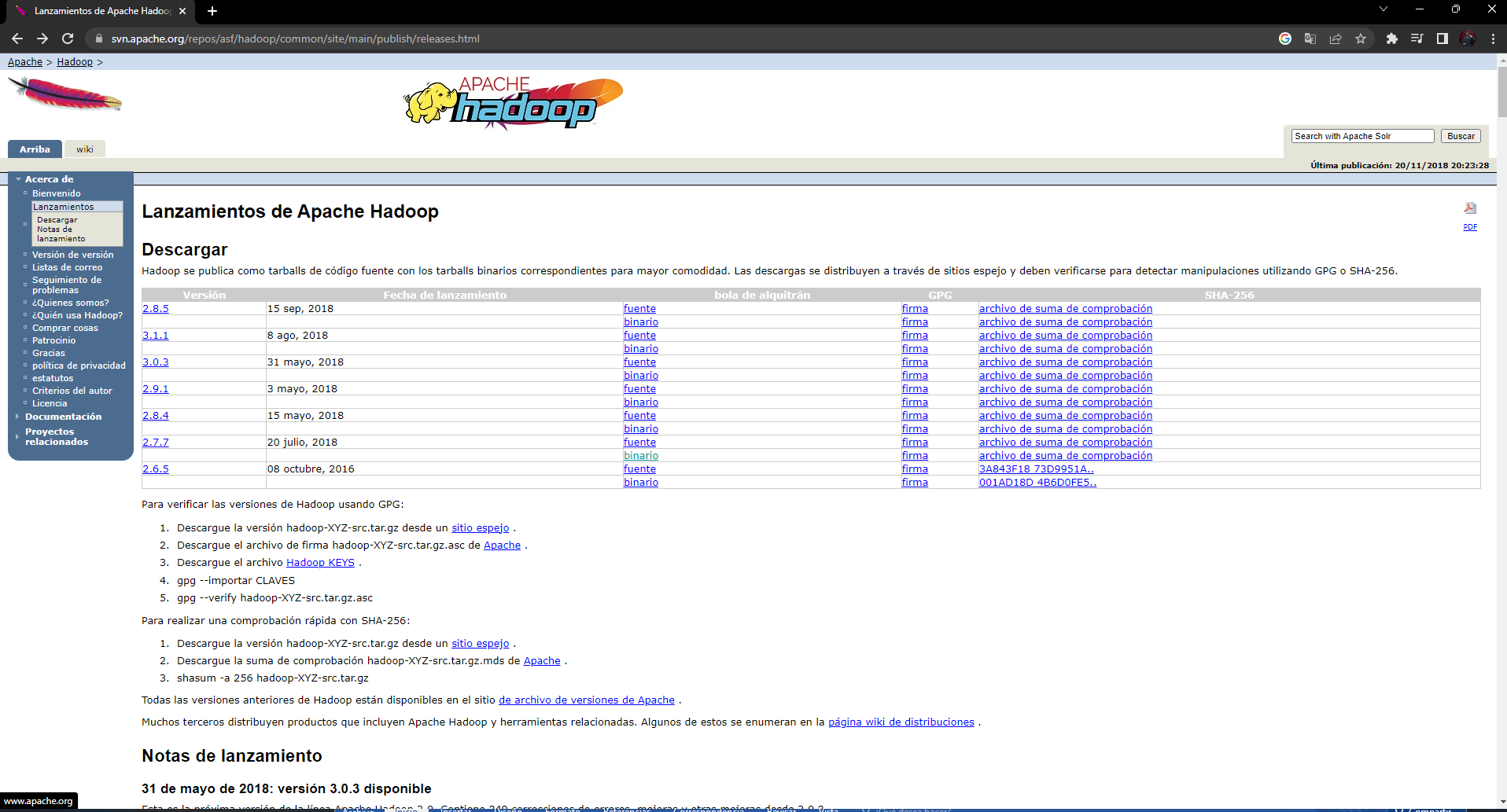
Es una colección de servicios. Software, aplicaciones y conectores que trabajan juntos para transformar dividir las fuentes de datos en consistentes, interactivas y Visualmente atractivo. Los datos pueden ser una hoja de cálculo de Excel o Una colección de almacenamiento híbrido local y en la nube. Power BI se conecta a fuentes de datos, visualizaciones y descubra lo que importa y compártalo con cualquiera o todos usuario preferido.

**Tableau**

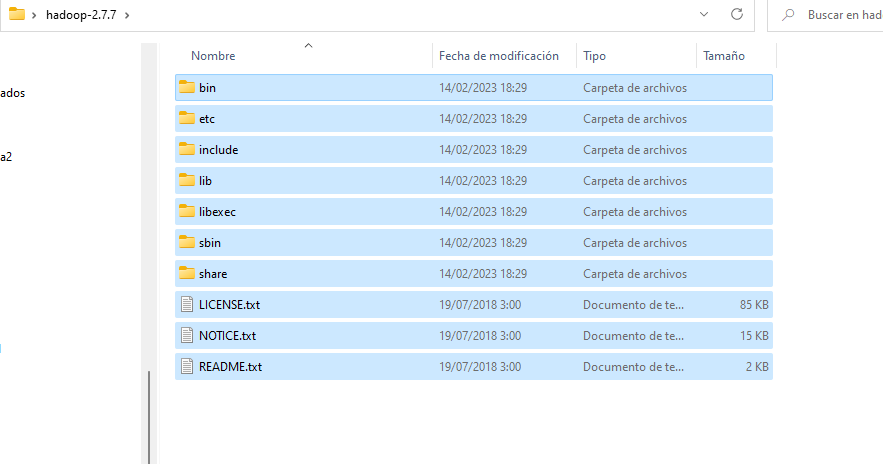
La plataforma Tableau es una elección inteligente negocio moderno líder en el mercado. más fácil de explorar y gestionar datos. También permite descubrir y compartir más información. Rápidamente.

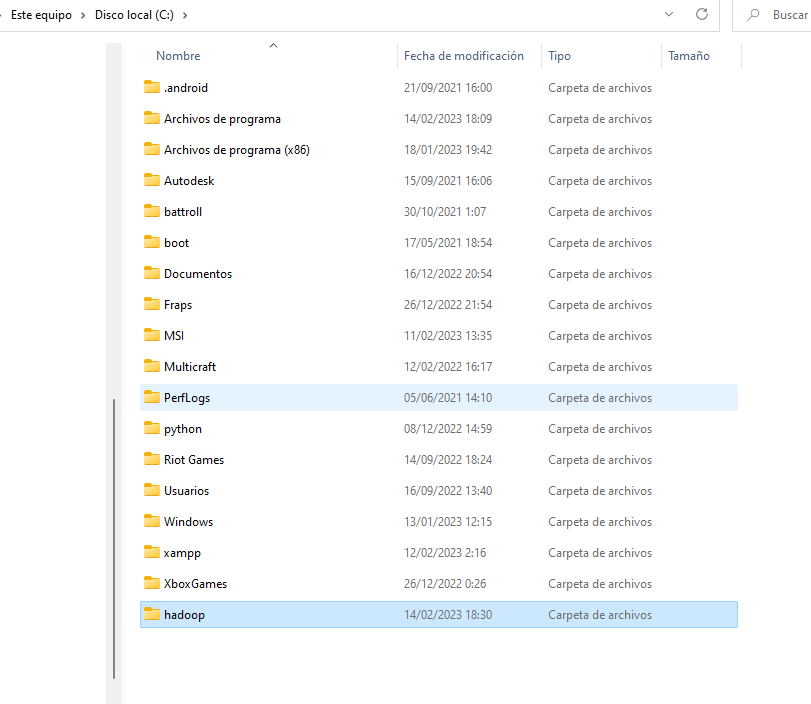
**Fase2:**

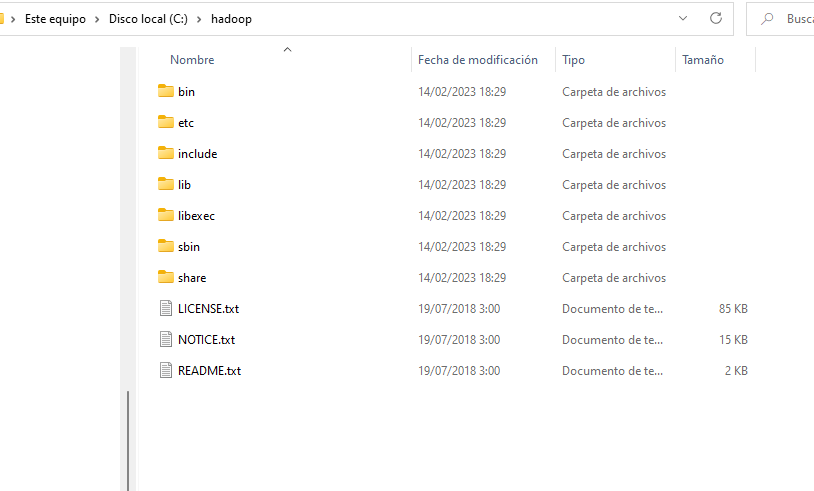
**En esta segunda fase se realiza la implementación de la investigación. En concreto sería acceder a un volumen de datos y mostrarlo. Podríamos utilizar Scala o Python y mostrar el resultado en PowerBI o Tableau. La idea es que sea algo muy impactante por la calidad de contenido tratado, velocidad de acceso, volumen de datos....**

-Primero descargamos el Hadoop,

-Extraemos el rar

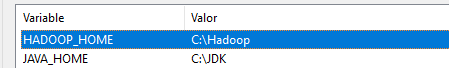


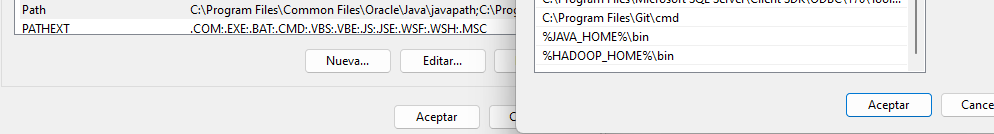
-Creamos una carpeta llamada Hadoop en C:, y metemos el contenido que hemos extraído



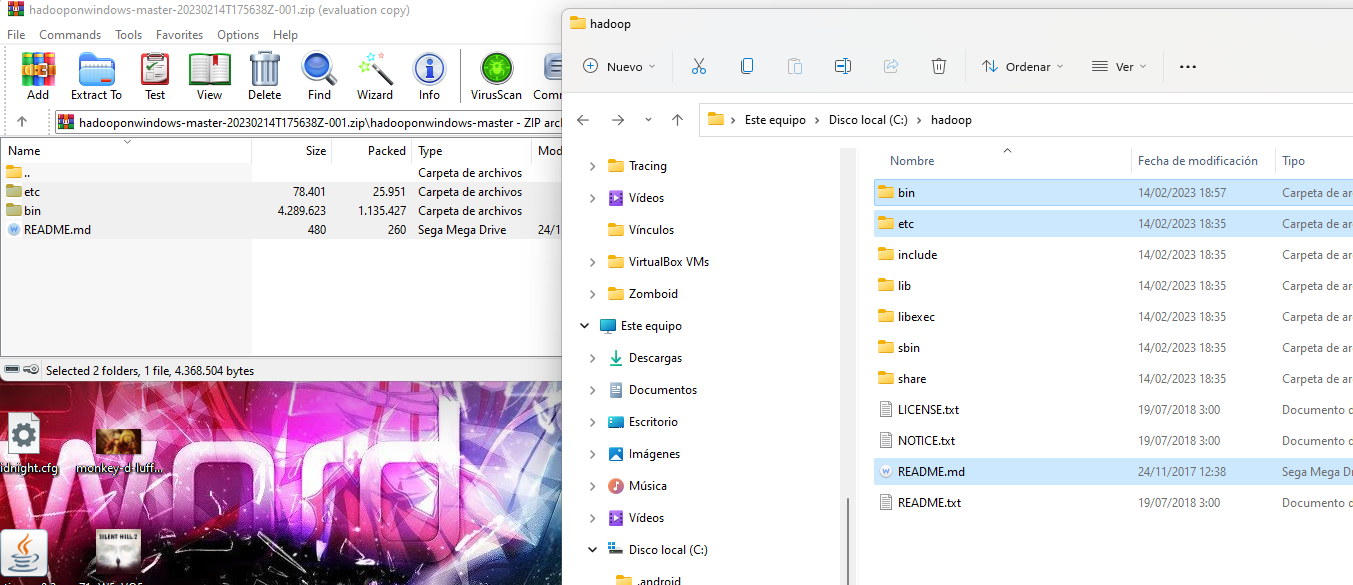
Creamos las siguientes variables de entorno:

JAVA\_HOME, la variable es la ruta al jdk.

HADOOP\_HOME, la variable es la ruta de Hadoop.

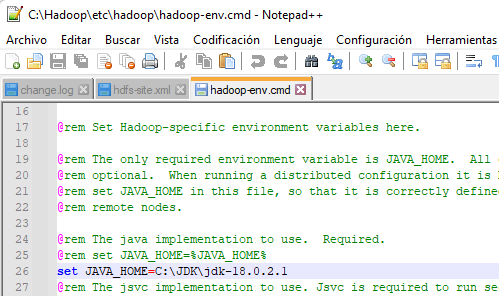
-Ahora en variables del sistema, añadimos 2 nueva a path

-Nos descargamos otro rar y sustituimos los archivos de la Carpeta Hadoop por esos

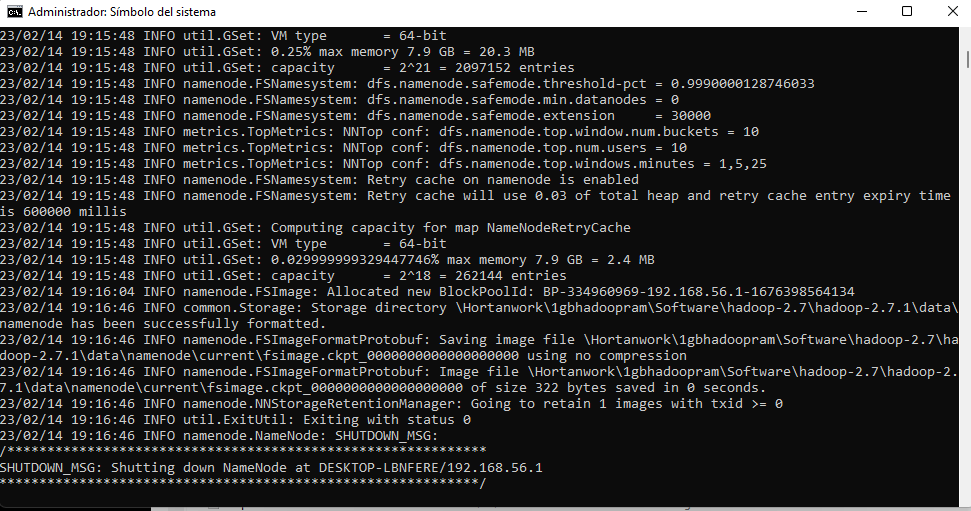


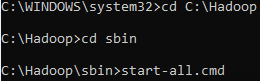
Abrimos la carpeta etc, y buscamos el siguiente archivo:

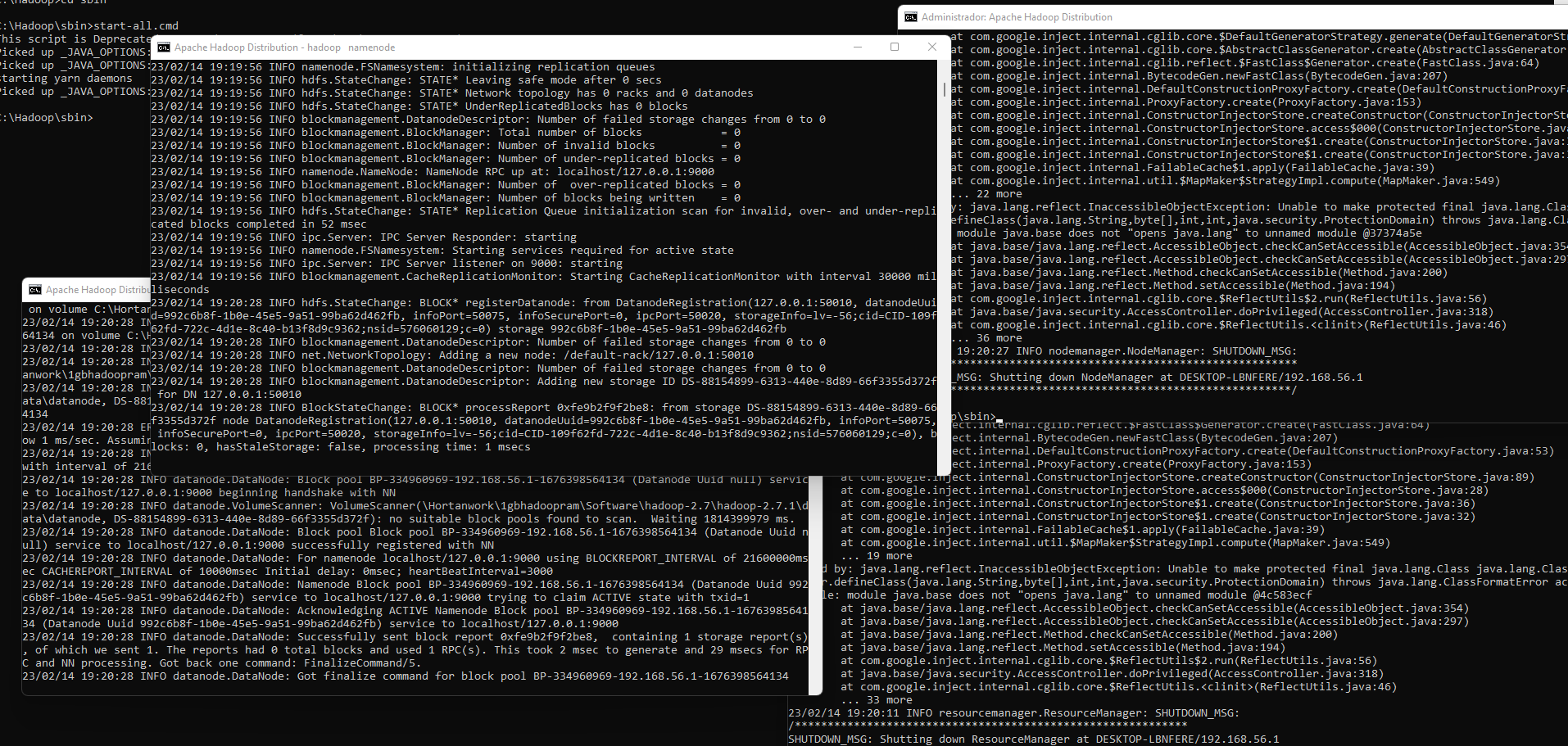
-Editamos el archivo y cambiamos la ruta del JAVA\_HOME:

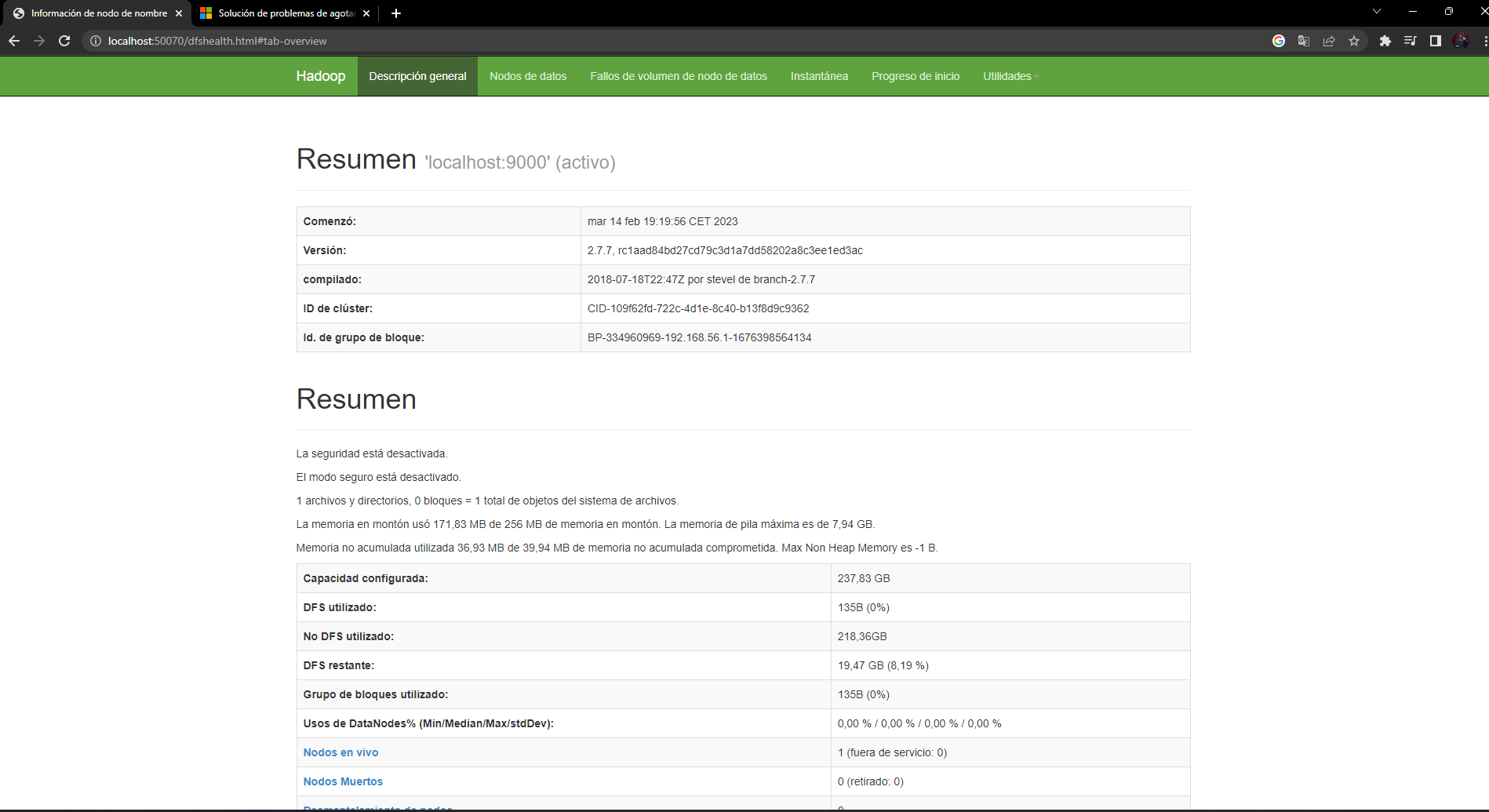
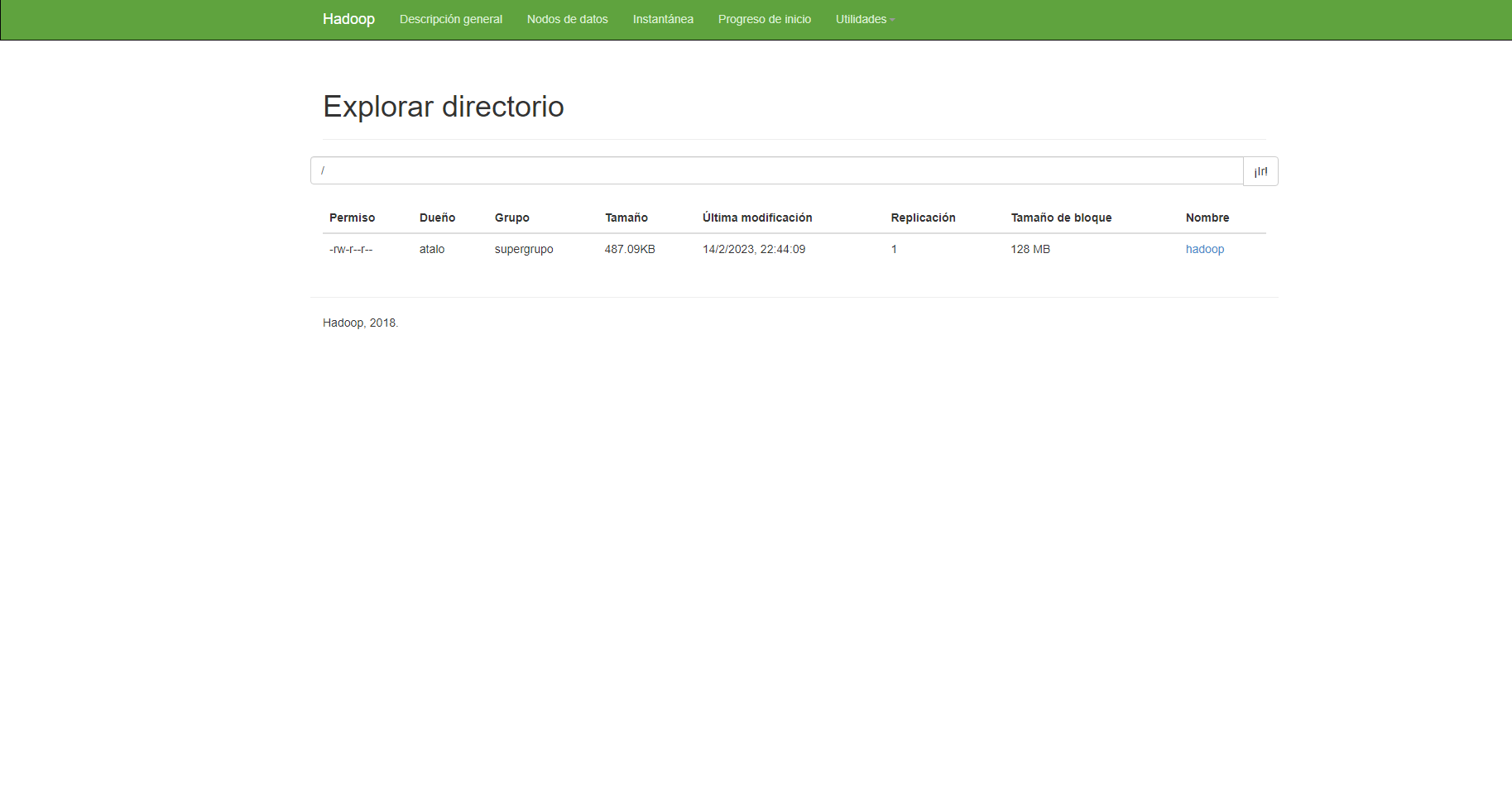


-Iniciamos cmd como admin y ponemos las siguientes ordenes:

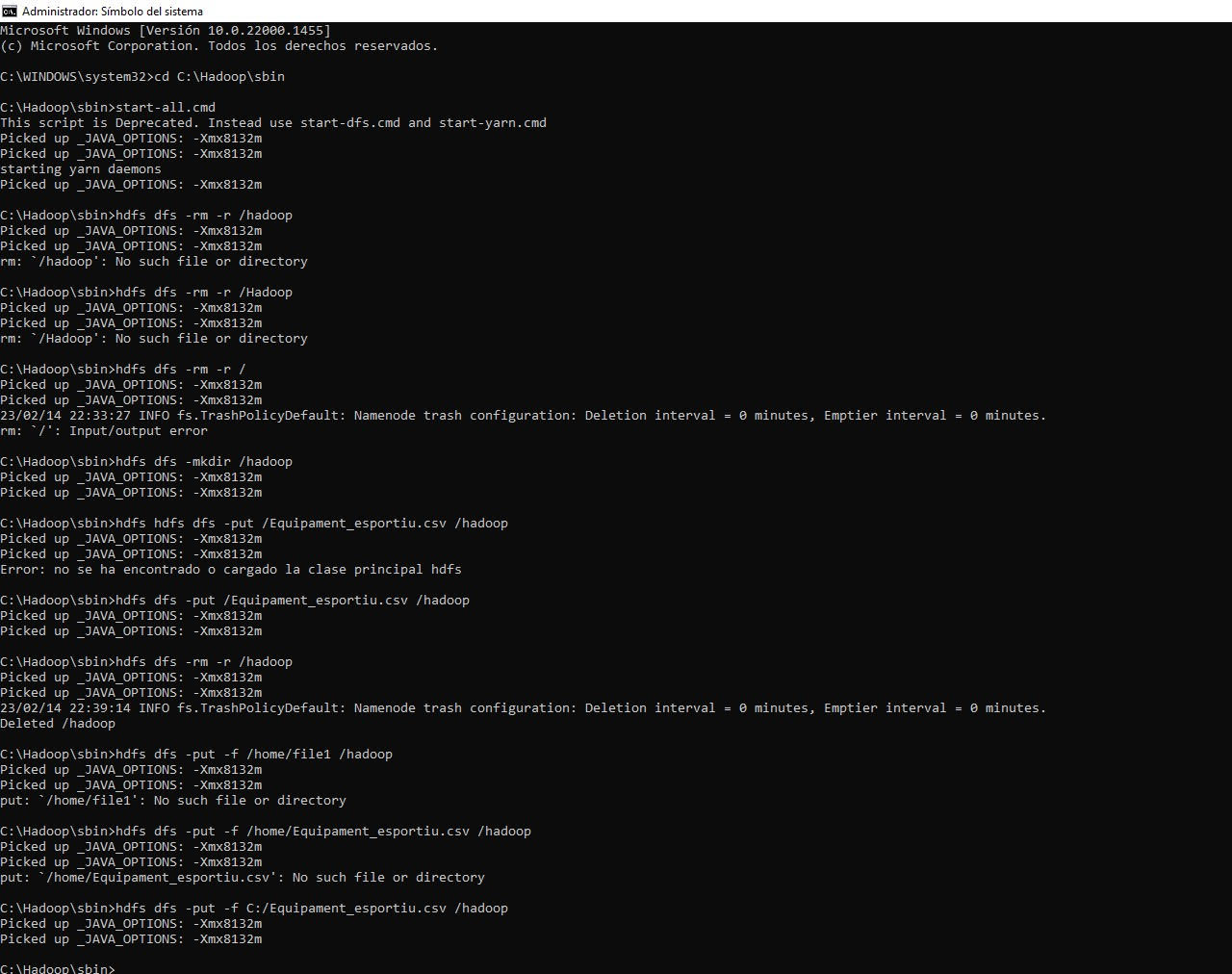


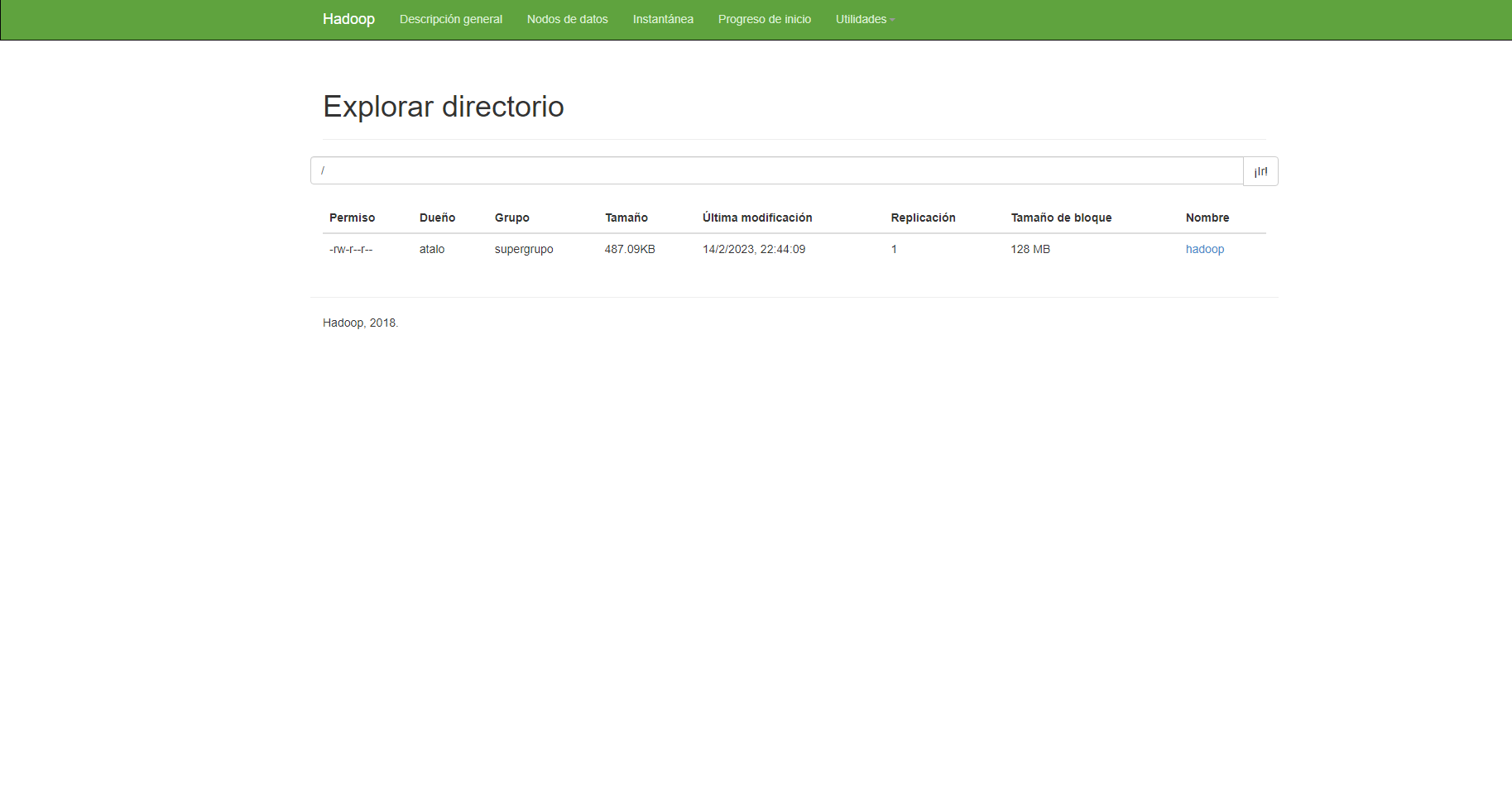




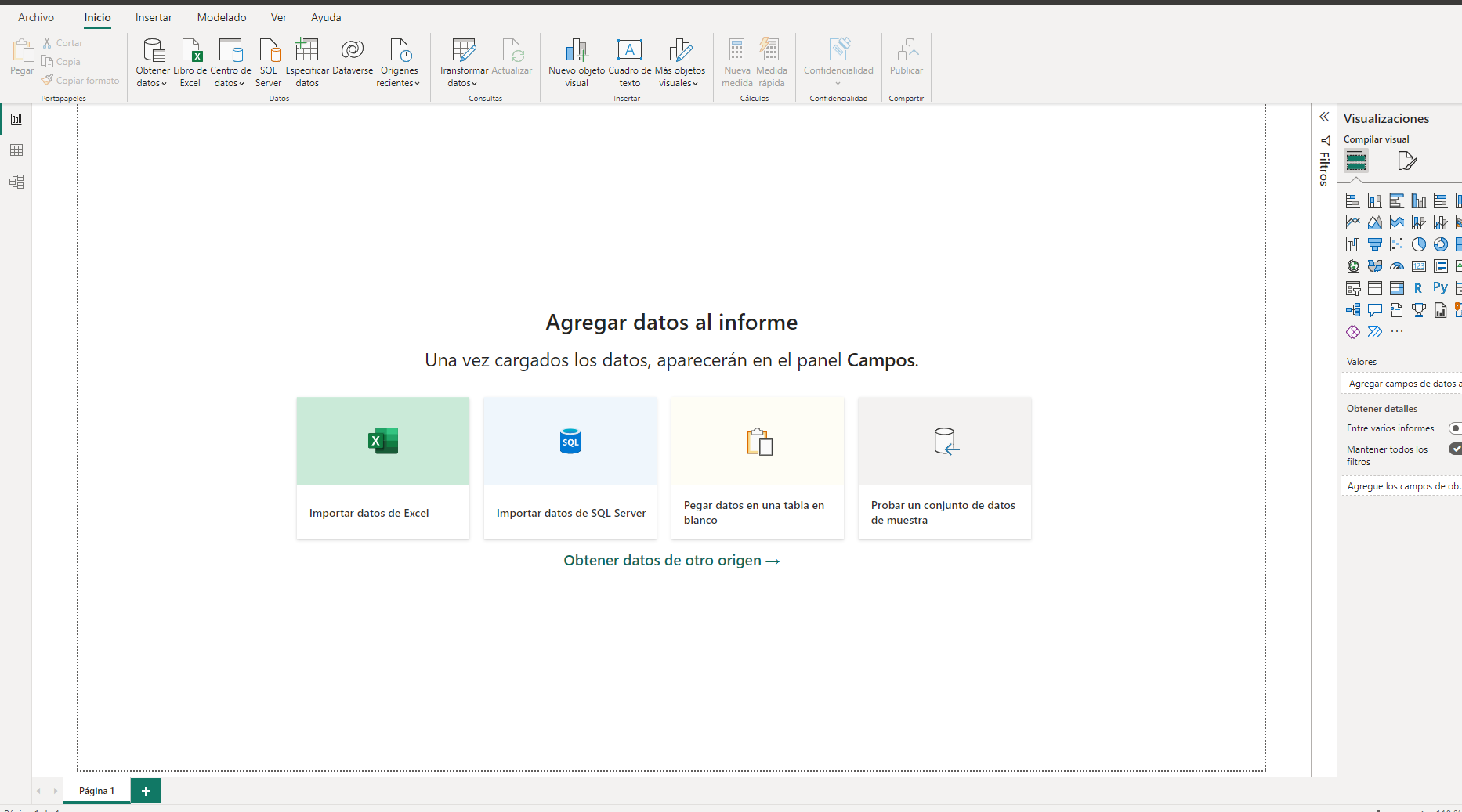
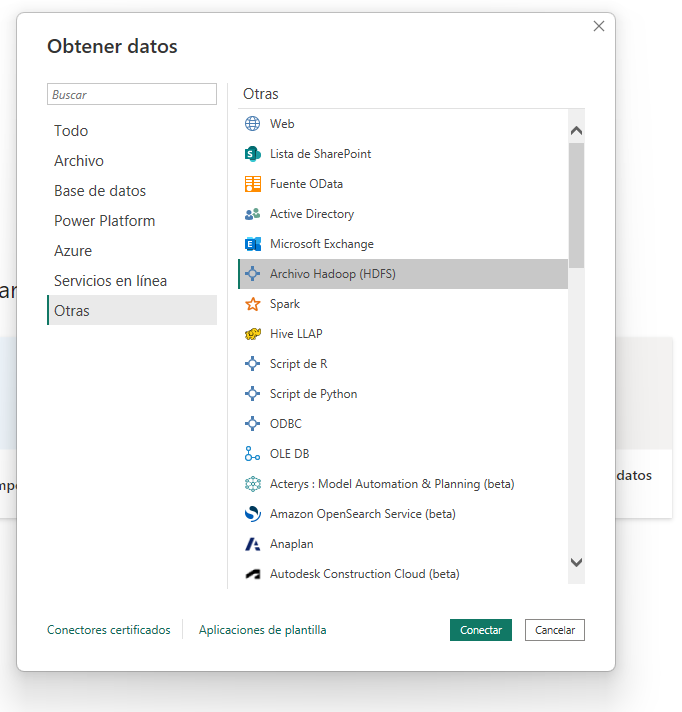
-Ahora en el navegador ponemos localhost:50070, y si están bien los pasos anteriores veremos esto:

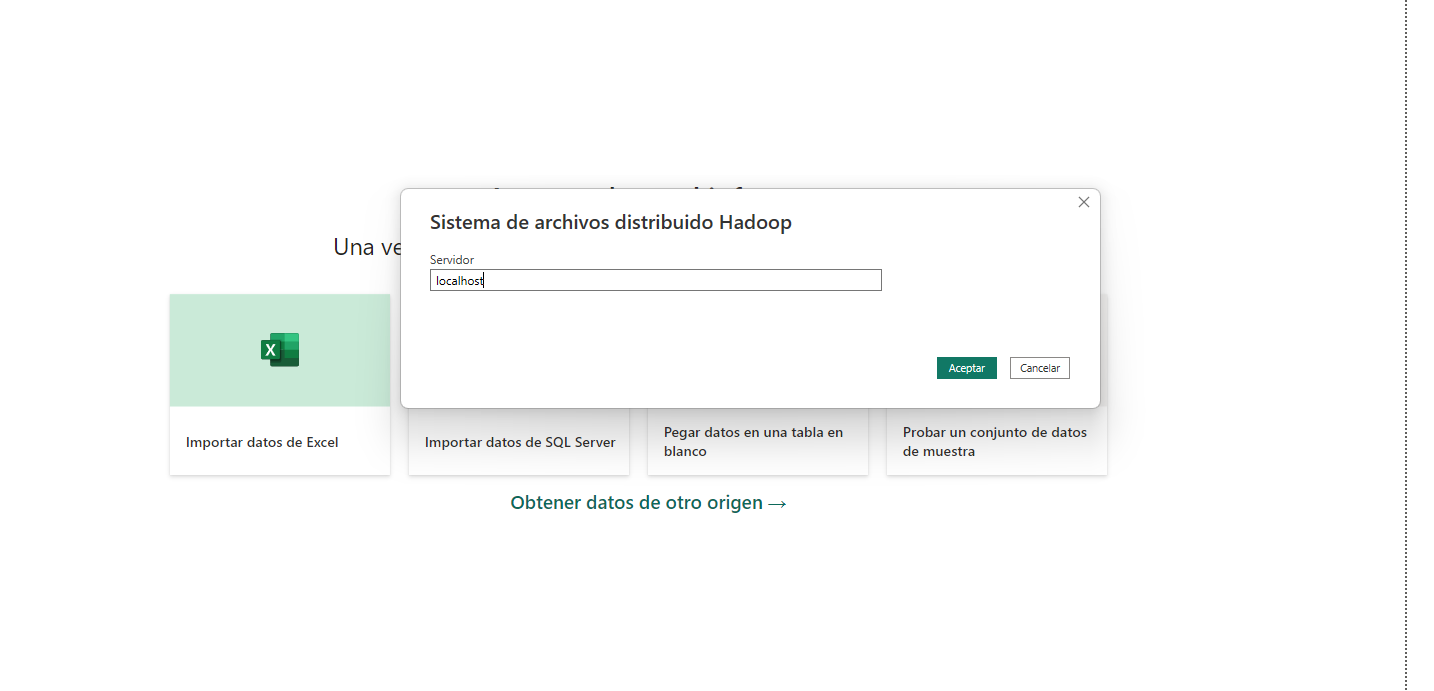
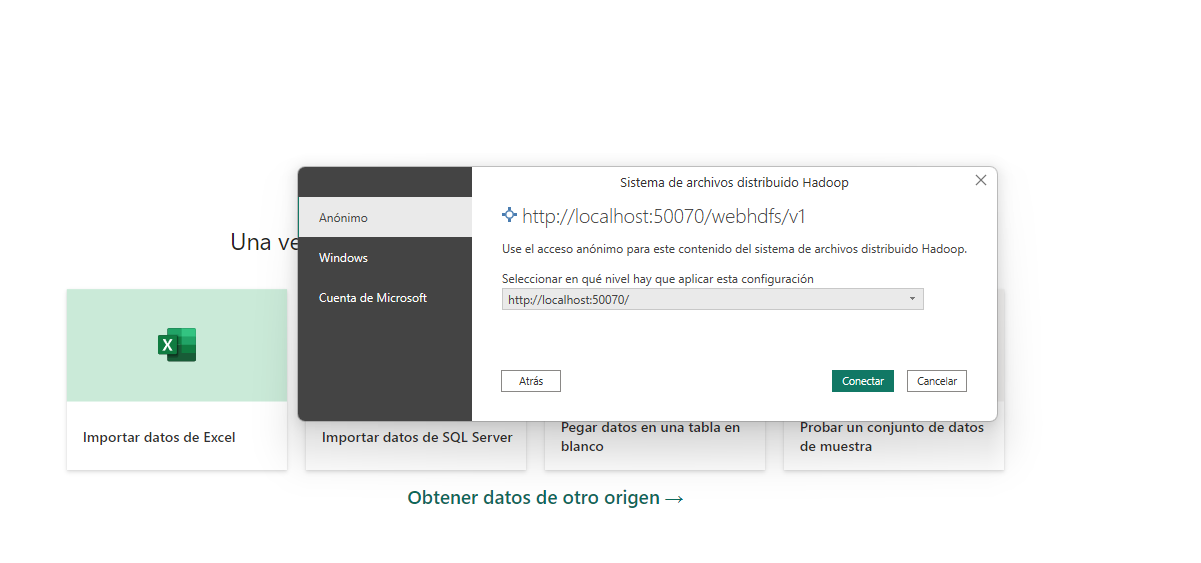
-Ahora en la consola ponemos las siguientes ordenes para poder subir el CSV que tendremos que descargar en nuestro caso en la página del gobierno.

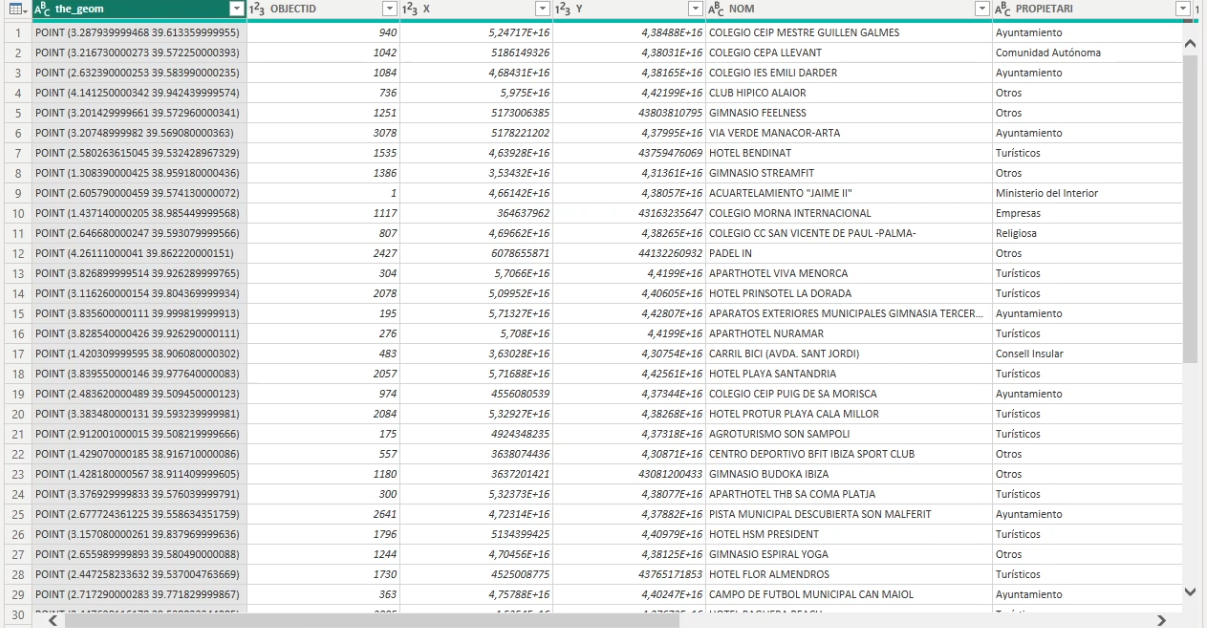
-

- Al subirse se verá así´:

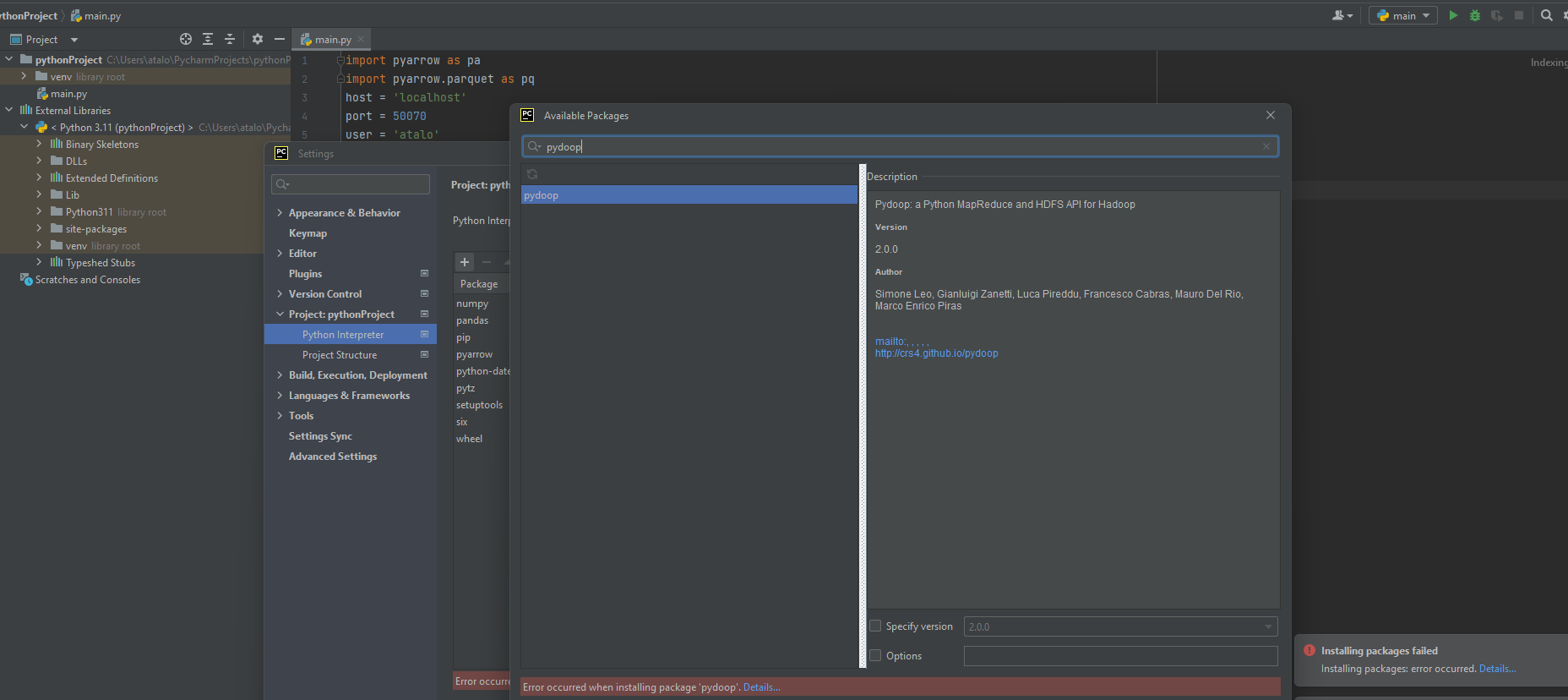
-Hacemos la parte de powerbi:



-Lo visualizamos



-En el apartado de Python, no lo hemos podido realizar debido a que nos saltaba repetidamente un error al intentar instalar las librerías.



Fase 3:

**Hadoop** es un software de procesamiento de grandes volúmenes de datos que ha demostrado ser muy efectiva en la gestión y análisis de datos. Está diseñado para trabajar con grandes conjuntos de datos. La tecnología permite la distribución de datos y el procesamiento de forma distribuida, lo que permite trabajar con conjuntos de datos que superan el tamaño de la memoria disponible en una sola máquina. Hadoop también es capaz de manejar el aumento de volumen de datos sin afectar el rendimiento.

También permite la gestión de diversos tipos de datos, desde estructurados a no estructurados, y se integra con otras tecnologías y herramientas para la realización de análisis de datos.

Es una tecnología de código abierto, por lo que el costo de licencias y soporte es mucho más bajo en comparación con otras soluciones comerciales.

**Power BI** es un software de análisis y visualización de datos de Microsoft. Destaca por su facilidad de uso y funcionalidades avanzadas. Se integra con una gran cantidad datos, desde bases de datos relacionales hasta servicios en la nube como Azure y Salesforce. Además, Power BI puede conectar con fuentes de datos locales, archivos de Excel, y servicios de streaming de datos. Además, sus datos pueden verse mediante distintos tipos de gráficos, diagramas... Estas visualizaciones pueden ser personalizadas y configuradas para adaptarse a las necesidades de cada usuario.

Power BI proporciona herramientas avanzadas de análisis de datos, como modelos de aprendizaje automático, análisis de tendencias, análisis de segmentación y otras funcionalidades avanzadas.

Cabe destacar la compatibilidad con otras aplicaciones de Microsoft, lo que facilita la colaboración y el intercambio de datos entre distintas aplicaciones.

**Bibliografía**

[**https://www.powerdata.es/big-data**](https://www.powerdata.es/big-data)

[**https://www.powerdata.es/data-lake**](https://www.powerdata.es/data-lake)

[**https://ayudaleyprotecciondatos.es/bases-de-datos/diferencias-datos-estructurados-no-estructurados/**](https://ayudaleyprotecciondatos.es/bases-de-datos/diferencias-datos-estructurados-no-estructurados/)

[**https://www.juanbarrios.com/que-son-hadoop-y-spark/**](https://www.juanbarrios.com/que-son-hadoop-y-spark/)

<https://sitiobigdata.com/2020/01/16/python-para-big-data-lenguaje-querido/>

<https://ceupe.com.ar/blog/python-lenguaje-de-programacion-aplicado-al-big-data/>