

**UNIVERSIDAD TECNICA PARTICULAR DE LOJA**

**“CIENCIAS DE LA COMPUTACIÓN”**

**PROYECTO FINAL**

**Prácticum 1.2**

**Integrantes:**

-Hernán Sánchez

- Joselito Ordoñez

-Andrés Jiménez

- Luis Granda

**Fecha:**

27/07/2021

**Docentes:**

ING. Eduardo Encalada,

ING. Jorge López

**Tema:**

Análisis, interpretación y visualización de datos

**Grupo:**  **F**

Abril 2021 - agosto 2021

Contenido

[1. Introducción 3](#_Toc78329197)

[2. Planificación del trabajo 3](#_Toc78329198)

[2.1 Descarga del archivo csv 3](#_Toc78329199)

[2.2 Normalización de datos del archivo csv 3](#_Toc78329200)

[2.3 Creación del modelo lógico y Conceptual 3](#_Toc78329201)

[2.4 Carga de datos a nuestro esquema de MYSQL 3](#_Toc78329202)

[2.5 Conexión de nuestra base de datos a la herramienta Zeppelin 3](#_Toc78329203)

[2.6 Visualización de consultas 3](#_Toc78329204)

[3. Herramientas 4](#_Toc78329205)

[3.1 EXCEL 4](#_Toc78329206)

[3.2 MYSQL WORKBENCH 4](#_Toc78329207)

[3.3 ZEPPELIN 4](#_Toc78329208)

[3.3.1 Data Frame SPARK 4](#_Toc78329209)

[3.3.2 Data Frame JDBC 4](#_Toc78329210)

[3.3.3 Visualización a través de plugin Helium 4](#_Toc78329211)

[3.3.4 Visualización a través de Angular más Framework JS (VEGA LITE) 4](#_Toc78329212)

[4. Fuentes de datos 5](#_Toc78329213)

[5. Componente base de datos 5](#_Toc78329214)

[5.1 Normalización de la data 5](#_Toc78329215)

[5.2 Creación del diseño conceptual de la base de datos 5](#_Toc78329216)

[5.3 Esquema de la base de datos 6](#_Toc78329217)

[5.4 Consultas 6](#_Toc78329218)

[6. Componentes Programación 7](#_Toc78329219)

[6.1 Conexión a Zeppelin mediante el laboratorio virtual 7](#_Toc78329220)

[6.2 Carga y desarrollo de consultas 8](#_Toc78329221)

[6.3 Visualización de datos en diferentes tipos de graficos 8](#_Toc78329222)

[6.4 Conexión a la base de datos 8](#_Toc78329223)

[7. Resultados Obtenidos 9](#_Toc78329224)

[7.1 Conexión y visualización de datos 9](#_Toc78329225)

[8. Conclusión 10](#_Toc78329226)

[9. Bibliografía 11](#_Toc78329227)

PROYECTO FINAL

# Introducción

El Proyecto Final Desarrollado tiene como finalidad poner en práctica los conocimientos adquiridos durante el ciclo académico en las materias integradas (Base de datos Avanzada, Programación Avanzada), Consiste en aplicar un análisis y manejo exploratorio de datos con el propósito de visualizar e interpretar los datos de covid 19 previamente normalizados los cuales fueron extraídos de una fuente compartida en la plataforma del GITHUB.

# Planificación del trabajo

## Descarga del archivo csv

Accedemos a la fuente compartida en la plataforma de GITHUB y realizamos la descarga del archivo en formato csv

## Normalización de datos del archivo csv

Realizamos el análisis de los datos, y los dividimos en varios archivos csv para realizar una mejor interpretación.

## Creación del modelo lógico y Conceptual

Diseñamos nuestro modelo lógico y conceptual para la creación de nuestra base de datos teniendo en cuenta todas las variables encontradas en el csv realizando la cardinalidad entre las tablas.

## Carga de datos a nuestro esquema de MYSQL

Una vez implementada nuestra base de datos procedemos a realizar la carga de los datos mediante la importación de archivos

## Conexión de nuestra base de datos a la herramienta Zeppelin

Creamos el intérprete de modelo jdbs, para realizar la conexión a nuestra base de datos de mysql y realizar consultas con el lenguaje propio de sql

## Visualización de consultas

Usando frameworks propia de Zeppelin mostramos diferentes tipos de graficas que nos ayudan a interpretar y analizar a través de visualizaciones de las consultas previamente realizadas

# Herramientas

## 3.1 EXCEL

Excel es un programa informático desarrollado y distribuido por Microsoft Corp. Se trata de un software que permite realizar tareas contables y financieras gracias a sus funciones, desarrolladas específicamente para ayudar a crear y trabajar con hojas de cálculo.

(DE, 2008)

## 3.2 MYSQL WORKBENCH

MySQL Workbench permite diseñar visualmente, modelar, generar y administrar bases de datos. Incluye todo lo que necesita un modelador de datos para crear modelos complejos de ER, ingeniería directa e inversa y también ofrece características clave para realizar tareas difíciles de gestión de cambios y documentación que normalmente requieren mucho tiempo y esfuerzo. (A||Ubuntu, s.f.)

## 3.3 ZEPPELIN

Zeppelin es una moderna plataforma de ciencia de datos compatible de serie con Spark y Hive. De hecho, Zeppelin puede trabajar con back-ends que admiten varios lenguajes, por lo que es compatible con un ecosistema de fuentes de datos cada vez mayor. (CLOUDERA, 2021)

### 3.3.1 Data Frame SPARK

Es un módulo de Spark para el procesamiento de datos estructurados. A diferencia de la API básica de Spark RDD, las interfaces proporcionadas por Spark SQL brindan a Spark más información sobre la estructura tanto de los datos como del cálculo que se está realizando. Internamente, Spark SQL usa esta información adicional para realizar optimizaciones adicionales. (3.2.1, s.f.)

### 3.3.2 Data Frame JDBC

Spark SQL también incluye una fuente de datos que puede leer datos de otras bases de datos utilizando JDBC. Esta funcionalidad debe preferirse al uso de [JdbcRDD](https://spark.apache.org/docs/latest/api/scala/org/apache/spark/rdd/JdbcRDD.html) . Esto se debe a que los resultados se devuelven como un DataFrame y pueden procesarse fácilmente en Spark SQL o unirse con otras fuentes de datos. La fuente de datos JDBC también es más fácil de usar desde Java o Python, ya que no requiere que el usuario proporcione una ClassTag. (3.1.2, s.f.)

### 3.3.3 Visualización a través de plugin Helium

HELIUM es una plataforma para la automatización de expedientes corporativos que facilitat su implantación, reduciendo el tiempo de desarrollo y ofreciendo a los usuarios finales una vista homogénea entre los diferentes expedientes. La plataforma está diseñada mediante una arquitectura modular que permite la integración con sistemas externos de soporte como por ejemplo sistemas de gestión documental, de firma, de tramitación telemática y otros sistemas de gestión interna. (PROGRAMARLLIURE, s.f.)

### 3.3.4 Visualización a través de Angular más Framework JS (VEGA LITE)

* Angular es un framework de diseño de aplicaciones y plataforma de desarrollo para crear aplicaciones de una sola página eficientes y sofisticadas. Esta documentación de Angular te ayuda a aprender y usar el framework y la plataforma de desarrollo, desde tu primera aplicación hasta la optimización de aplicaciones complejas de una sola página, para empresas. (A, 2020)
* Vega-Lite es una gramática de visualización de alto nivel. Proporciona una sintaxis JSON concisa para admitir la generación rápida de visualizaciones para respaldar el análisis. Las especificaciones de Vega-Lite se pueden compilar según las especificaciones de [Vega](http://vega.github.io/vega) . (VEGA-LITE, s.f.)

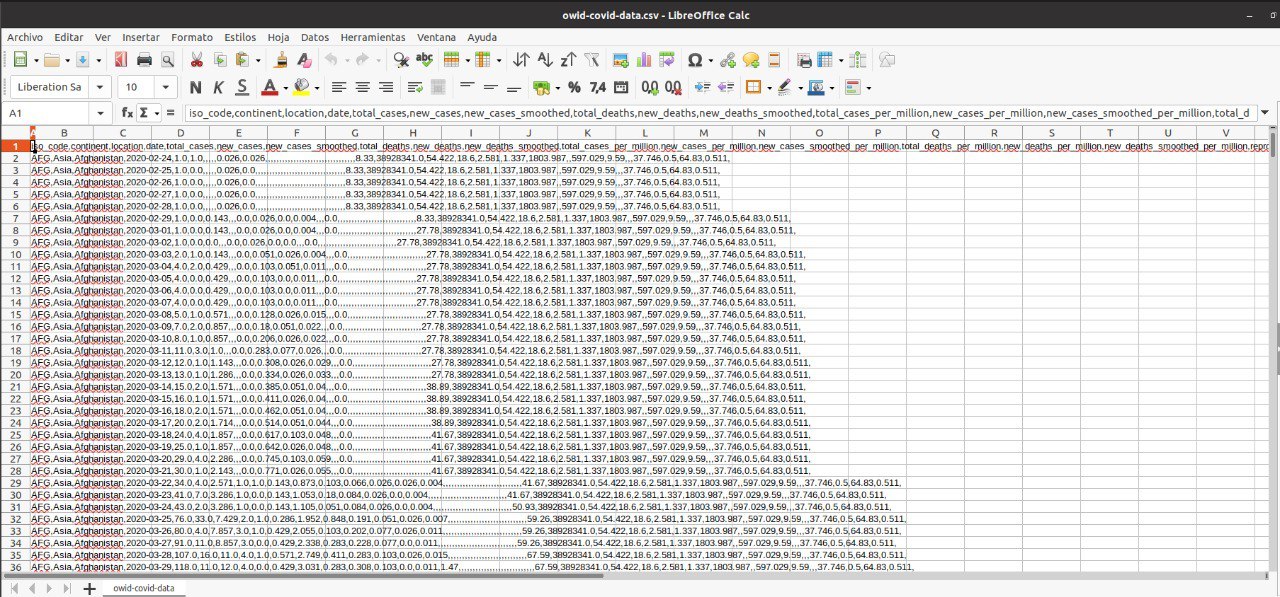
# Fuentes de datos

La Fuente de donde extrajimos los datos del covid es de la plataforma de GITHUB <https://github.com/owid>, proporcionada por el Ingeniero Eduardo Encalada donde fue descargada la data en archivo tipo csv

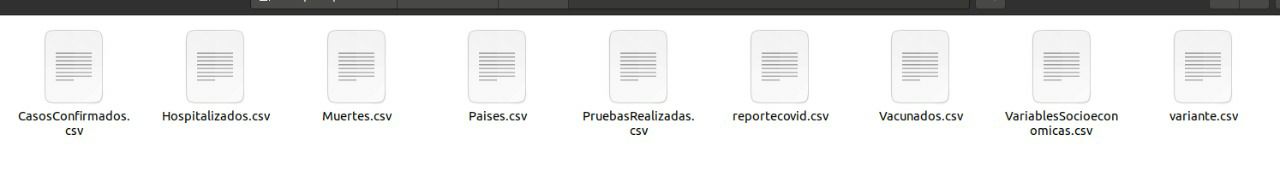
# Componente base de datos

## 5.1 Normalización de la data

Primero descargamos el archivo csv con datos crudos del covid 19

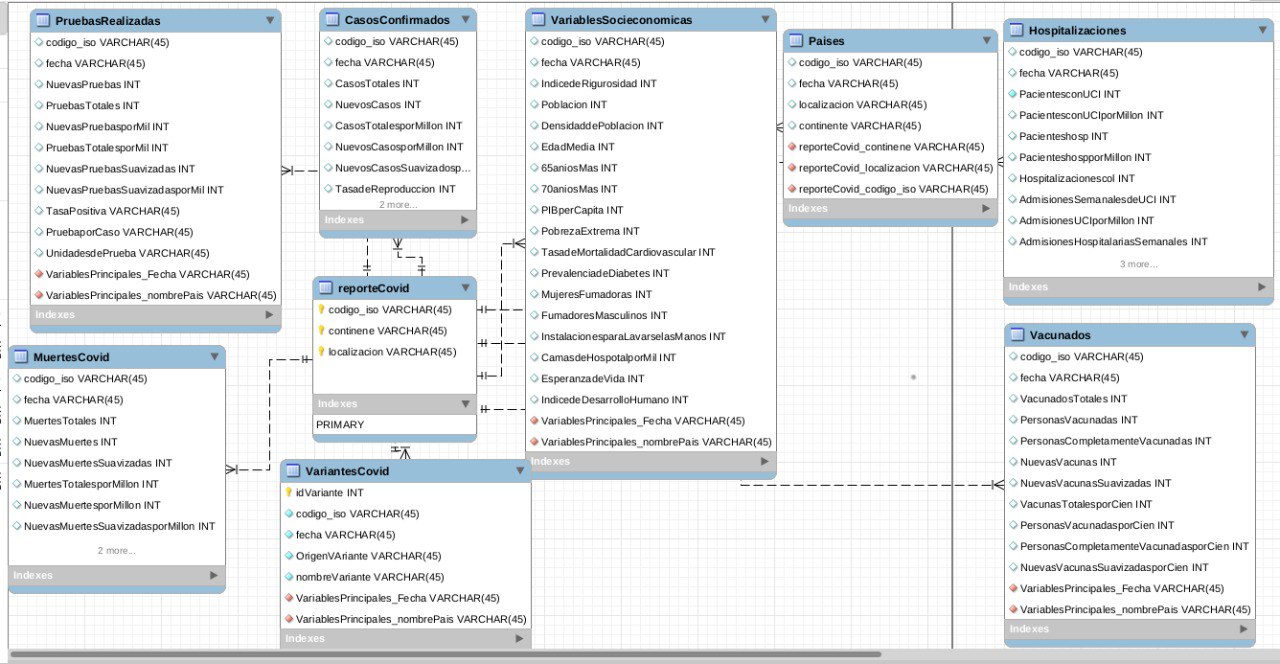


Luego de realizar la normalización así quedaron los archivos csv, listo para la importación a las tablas de nuestra base de datos en MYSQL



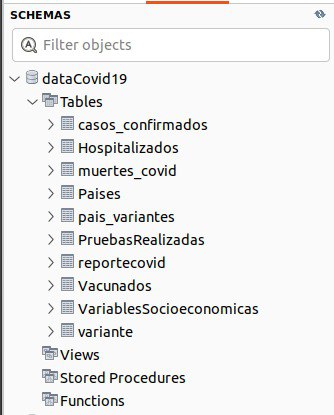
## 5.2 Creación del diseño conceptual de la base de datos

Creamos el diseño de nuestra base de datos a partir de los datos normalizados



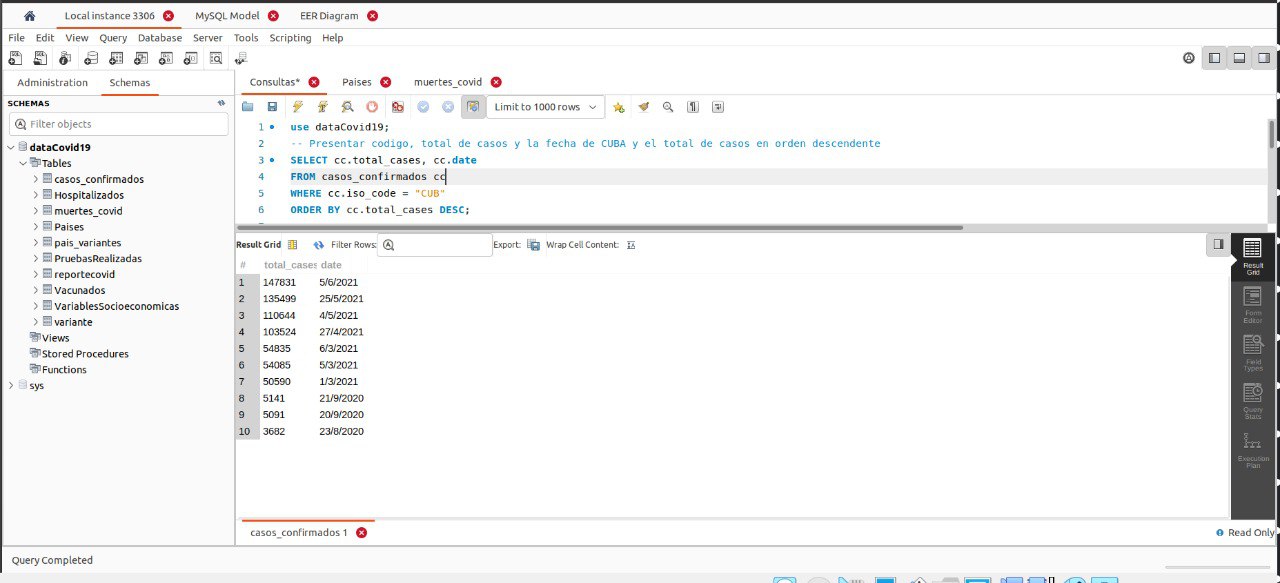
## 5.3 Esquema de la base de datos

Una vez importado los datos a las tablas así quedaría nuestro esquema de la base de datos



## 5.4 Consultas

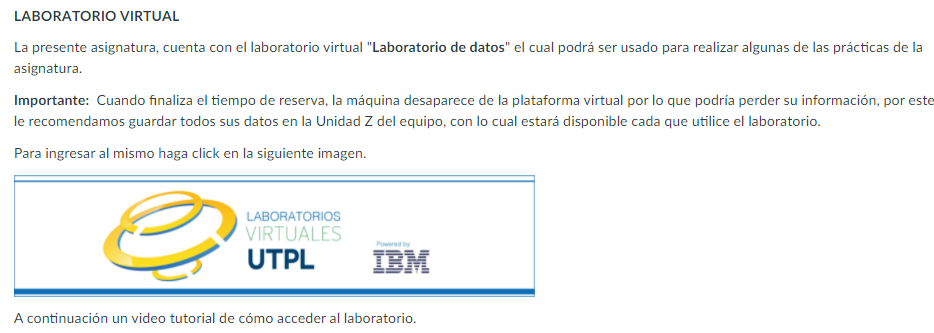
Realizamos diversas consultas, para comprobar la correcta importación de los datos a nuestras tablas de la base de datos

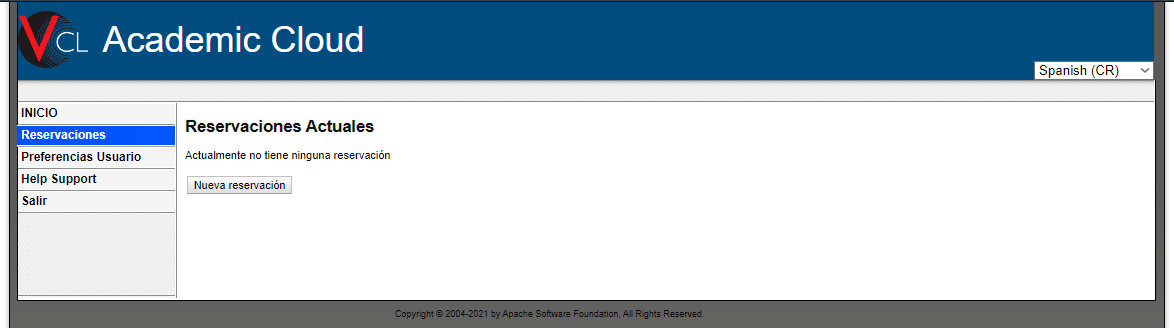


# Componentes Programación

## 6.1 Conexión a Zeppelin mediante el laboratorio virtual

Una de la Principales herramientas a utilizar es la herramienta de Spark Zeppelin, que en su s versiones no es compatible con Windows por lo que tuvimos problemas ya que somos usuarios Windows, buscamos diferentes formas y opciones para tener una conexión a la herramienta, la única forma fue mediante el uso de un laboratorio Virtual, que nos ofrece máximo 6 horas de uso.

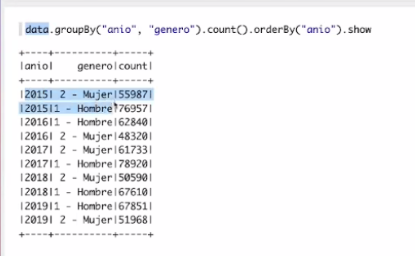




## 6.2 Carga y desarrollo de consultas

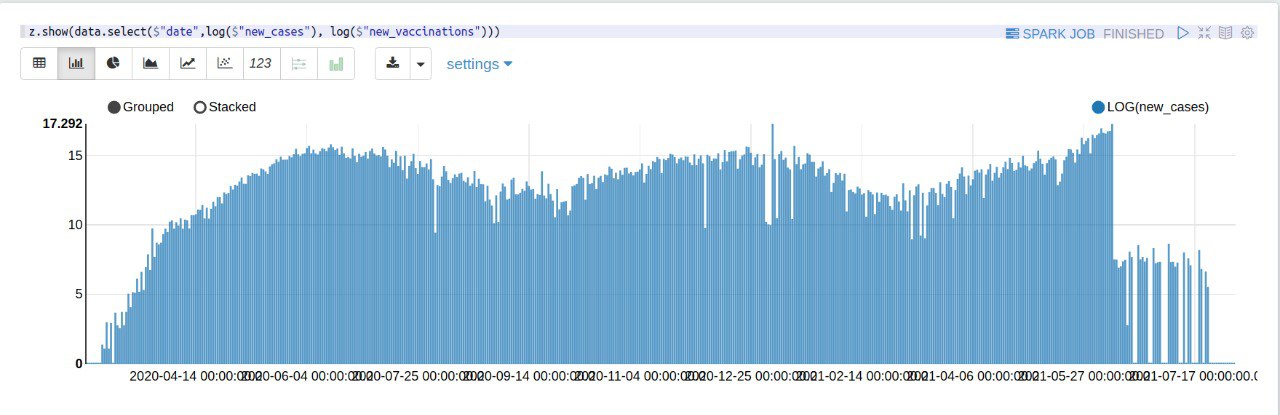
Una vez que logramos tener conexión a la herramienta, realizamos la carga de un archivo csv, seguimos los diferentes pasos para la carga y manejo de datos una vez echo eso realizamos diferentes consultas

* En este ejemplo es una consulta donde presentamos un conteo, donde también presentamos el año y el genero que esta ordenado según el año en forma descendente



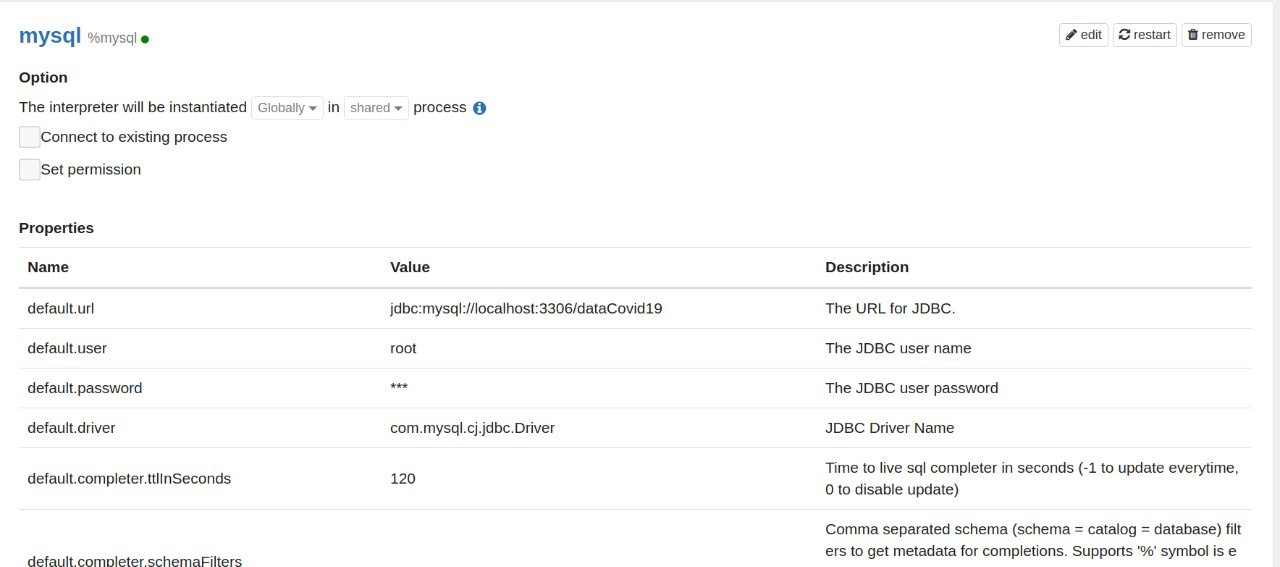
## 6.3 Visualización de datos en diferentes tipos de graficos

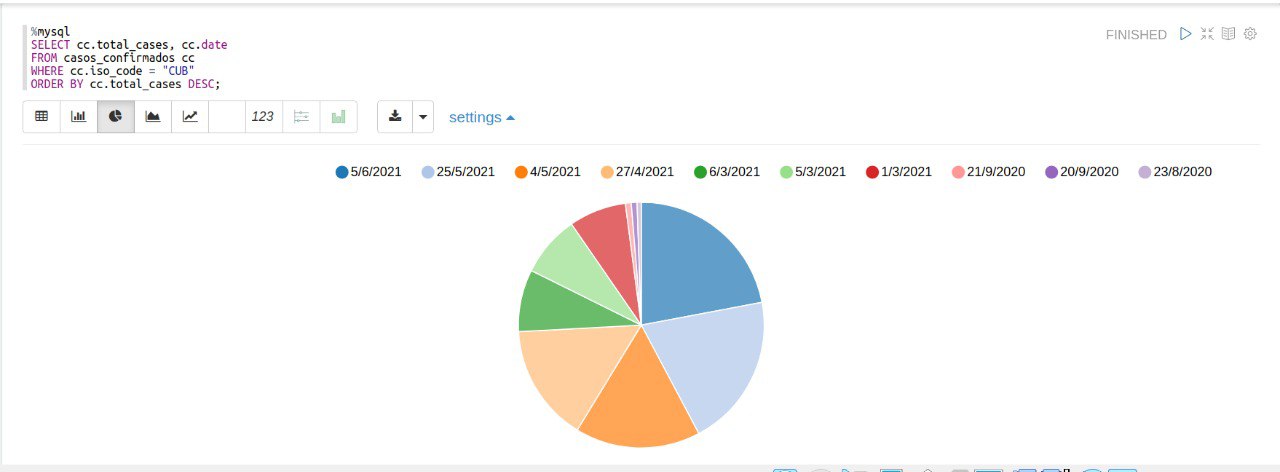
Aprendimos a realizar consultas mediante la visualización de los datos en diferentes frameworks del archivo csv previamente cargado



## 6.4 Conexión a la base de datos

Realizamos la conexión de la herramienta Zeppelin a nuestra base de datos en MYSQL mediante intérpretes como el JDBS y ya podemos realizar consultas y visualizaciones de los datos

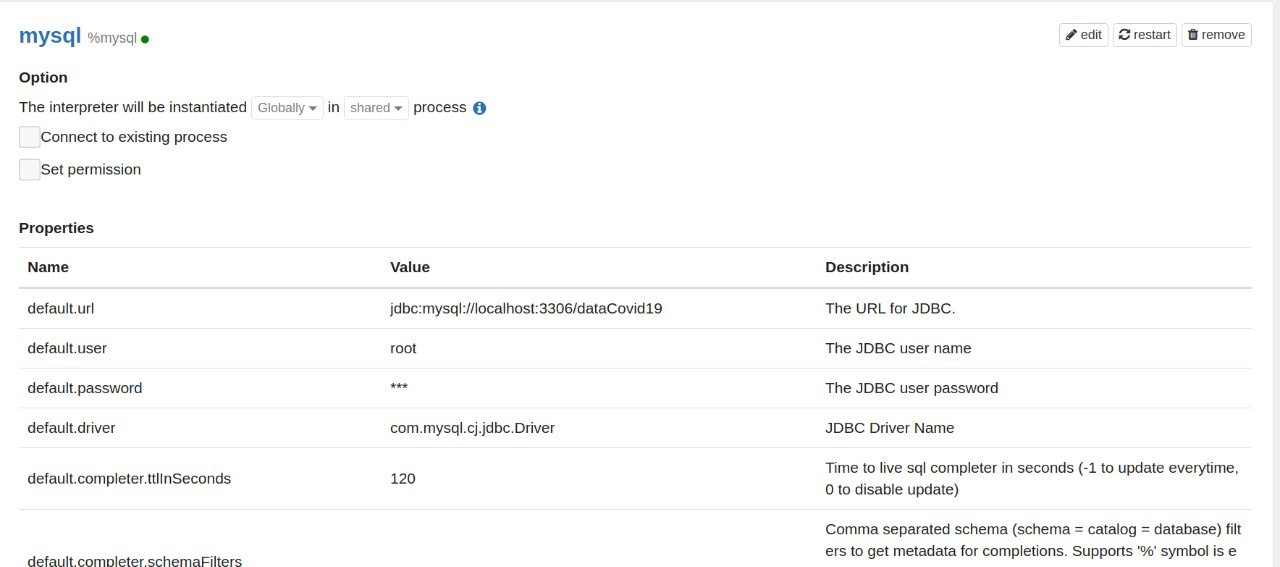




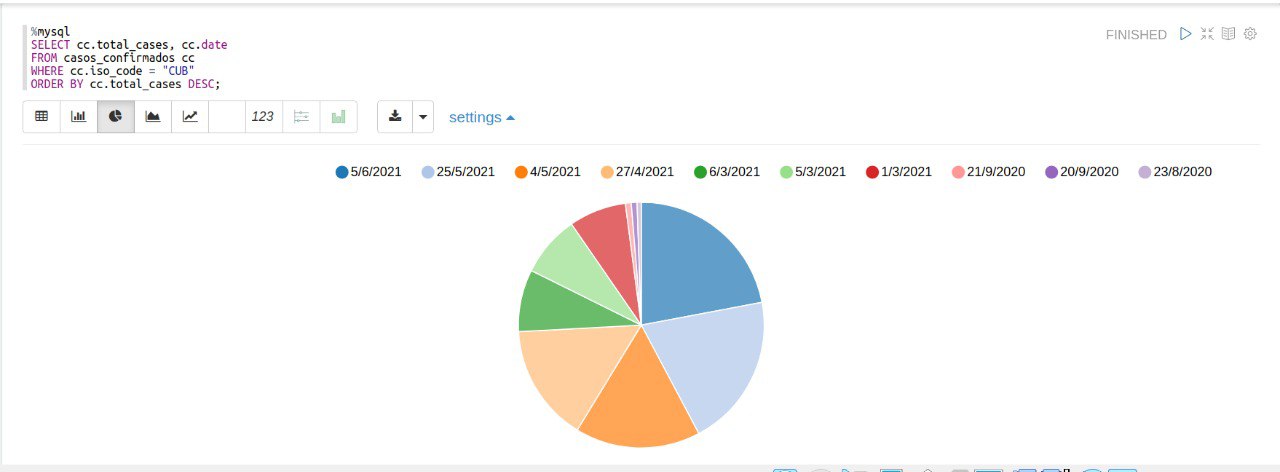
# Resultados Obtenidos

## 7.1 Conexión y visualización de datos

Una vez logrado la conexión de nuestra base de datos a la herramienta de Zeppelin, logramos realizar diferentes consultas con los que nos permitió tener diferentes visualizaciones de los datos con los distintos gráficos que nos ofrece

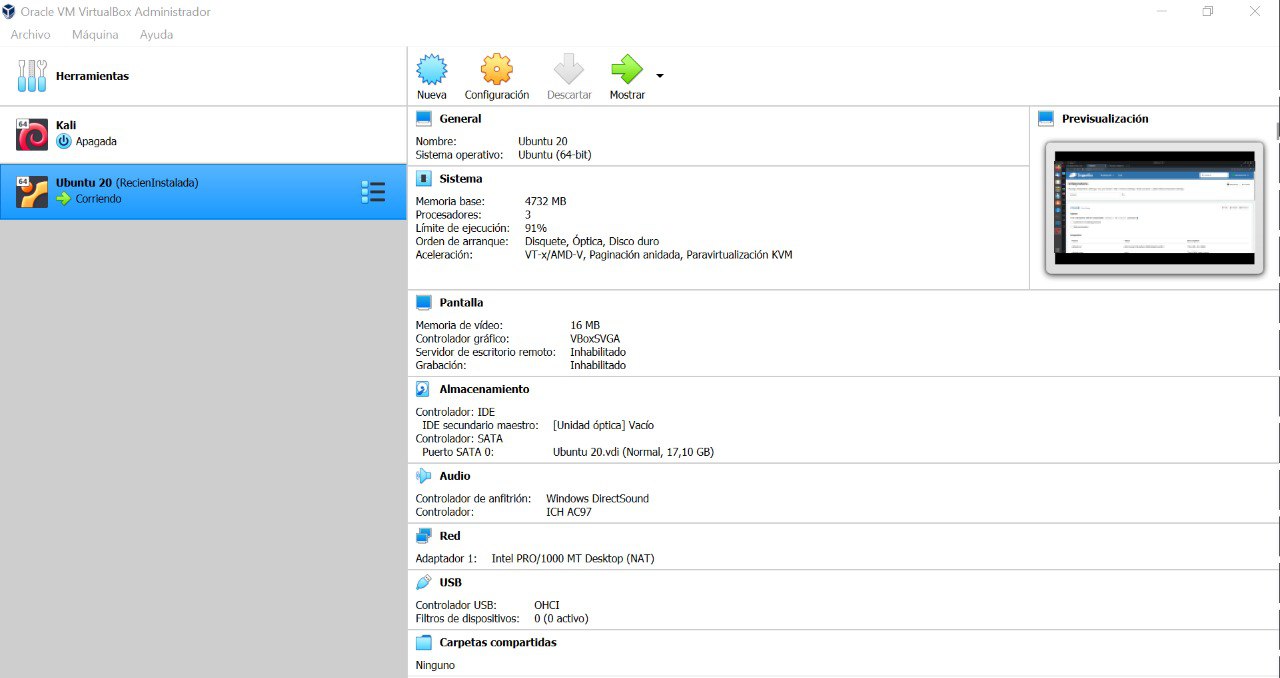


Logramos analizar y comprender el uso de las herramientas, para la manipulación de los datos, con el fin de presentar visualizaciones con la información, mediante las consultas realizadas.



# Conclusión

* En conclusión, el uso de las diversas herramientas para el análisis y manipulación de los datos, para poder realizar diversas consultas para lograr visualizaciones con la información que deseamos conocer.
* Tenemos que comprender los diferentes interpretes de las herramientas para lograr un mejor resultado en la conexión a nuestra base de datos mediante zeppelin para lograr un buen resultado en nuestras consultas
* La necesidad de tener conexión a la herramienta de zeppelin, buscamos diversas maneras de lograrlo, con lo cual la única solución fue instalar una máquina virtual(VIRTUAL BOX), donde instalamos el sistema operativo UBUNTU, y las diversa herramientas a utilizar en este proyecto tuvimos diferentes altercados en la realización, donde hasta toda la información se eliminaba y nos tocaba una y otra vez volver a realizar los mismo procedimientos para lograr culminar el proyecto.



# Bibliografía

3.1.2, S. (s.f.). *SPARK 3.1.2*. Obtenido de JDBC To Other Databases: https://spark.apache.org/docs/latest/sql-data-sources-jdbc.html

3.2.1, S. (s.f.). *SPARK 3.1.2*. Obtenido de Spark SQL, DataFrames and Datasets Guide: https://spark.apache.org/docs/latest/sql-programming-guide.html

A. (2020). *A*. Obtenido de Introducción a la Documentación de Angular: https://docs.angular.lat/docs

A||Ubuntu, D. (s.f.). *MySQL Workbench, herramienta visual para el diseño de bases de datos*. Obtenido de MySQL Workbench, herramienta visual para el diseño de bases de datos: https://ubunlog.com/mysql-workbench-bases-datos/

CLOUDERA. (2021). *CLOUDERA*. Obtenido de Apache Zeppelin: https://es.cloudera.com/products/open-source/apache-hadoop/apache-zeppelin.html#:~:text=Zeppelin%20es%20una%20moderna%20plataforma,de%20datos%20cada%20vez%20mayor.

DE, D. (2008). *DEFINICION DE*. Obtenido de DEFINICION DE EXCEL: https://definicion.de/excel/

PROGRAMARLLIURE. (s.f.). *PROGRAMARLLIURE*. Obtenido de HELIUM: https://www.caib.es/sites/programarilliure/es/28487/

VEGA-LITE. (s.f.). *VEGA-LITE*. Obtenido de VEGA-LITE: https://vega.github.io/vega-lite-v1/