



Data Growth
Community

"Potenciando el crecimiento colectivo"

EXPOSITOR

Jefferson Lennart Campos Segovia



[linkedin.com/in/jlcampossegovia/](https://www.linkedin.com/in/jlcampossegovia/)



+51 967778310

Sobre mi

- Estudiante de pregrado en ciencias de la computación de la UNSAAC
- Miembro estudiantil IEEE CS
- Embajador IEEEExtreme 17.0
- Entusiasta por el mundo de los datos



Universidad Nacional de
San Antonio Abad del Cusco
IEEE Student Branch



TEMA:

Migración, Procesamiento y Análisis de Datos en un Entorno Hadoop



DESCRIPCION DEL NEGOCIO

El Ministerio de Salud en Perú es la entidad gubernamental encargada de liderar y gestionar las políticas de salud en el país. Su función principal es promover la salud y el bienestar de la población peruana a través de la formulación de estrategias, la supervisión de la calidad de los servicios de atención médica y la implementación de programas de salud pública. Además, se dedica a garantizar el acceso a servicios de atención médica asequibles y de calidad para todos los ciudadanos, abordando tanto las necesidades de las zonas urbanas como las rurales. También desempeña un papel fundamental en la respuesta a emergencias de salud pública y la cooperación internacional en asuntos de salud.

En su búsqueda de mejorar la calidad de vida de los peruanos, el Ministerio de Salud trabaja en estrecha colaboración con otras instituciones gubernamentales, organizaciones no gubernamentales y organismos internacionales, contribuyendo así al desarrollo y la promoción de la salud en el país. Su labor se extiende desde la promoción de estilos de vida saludables y la prevención de enfermedades hasta la gestión de la atención médica en el sistema de salud peruano, con el fin de brindar servicios médicos eficientes y equitativos para toda la población.

SITUACION PROBLEMÁTICA

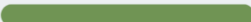







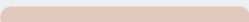
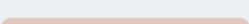

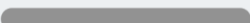

Dada la situación sanitaria que pasa el país por la pandemia del COVID-19, es crucial para el Ministerio de Salud saber disponibilidad de camas de hospitalización y UCI en zonas COVID y no COVID a nivel nacional puesto que plantea una problemática crucial en la gestión de la atención médica durante la pandemia. La capacidad de camas es un recurso crítico en la atención de pacientes infectados con COVID-19, y la variación en la disponibilidad de camas UCI y de hospitalización en distintas regiones y subsistemas de salud podría exponer disparidades significativas en la respuesta a la emergencia sanitaria. La falta de camas podría sobrecargar hospitales, impactando la calidad de la atención y aumentando la mortalidad, mientras que un exceso de camas disponibles podría resultar en recursos infrautilizados y gastos innecesarios. Por lo tanto, el análisis de la data histórica del registro de camas diarias es esencial para identificar patrones, tendencias y desafíos en la asignación y gestión de camas de hospitalización y UCI, lo que permitirá una toma de decisiones más informada y eficiente para abordar la demanda de atención médica en el contexto de la COVID-19.

REQUERIMIENTOS DEL CLIENTE

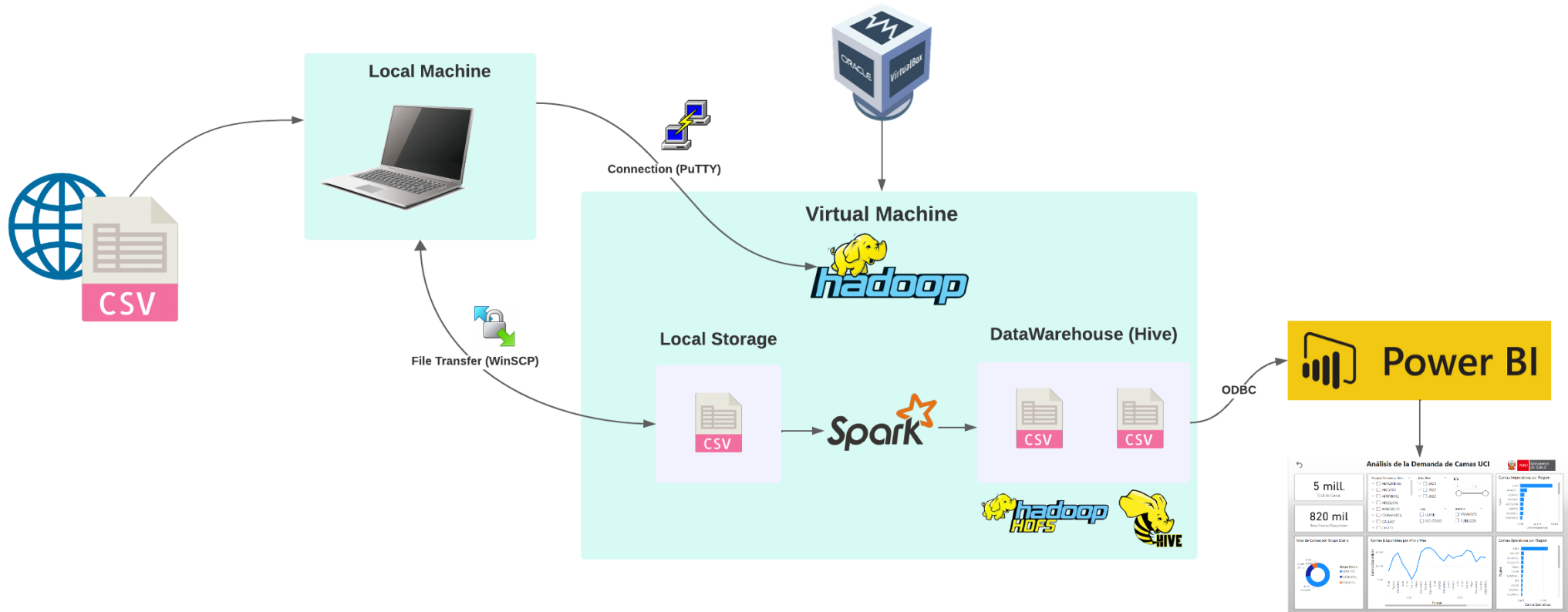
Desde la perspectiva del cliente, aplicar el proyecto Hadoop en el contexto de la gestión de camas de hospitalización y UCI es una respuesta a una problemática apremiante. Se traduce en una toma de decisiones más precisa y eficiente en la atención de pacientes, especialmente en situaciones críticas como la actual. Esta solución ofrece la capacidad de anticipar y satisfacer de manera óptima las necesidades de los pacientes, al tiempo que contribuye a una reducción significativa de los costos operativos, un desafío clave en la gestión de la salud.

La elección de Hadoop se justifica por su capacidad probada para lidiar con grandes volúmenes de datos y es una herramienta esencial para abordar esta problemática dado que la data es abundante y además nos brinda otras herramientas para hacer el procesamiento de datos que son necesarias para generar una data más limpia y lista para el análisis.

HITOS O FASES DEL PROYECTO

HITOS	DIA 1	DIA 2	DIA 3	DIA 4
Ingesta de Datos <ul style="list-style-type: none">• Buscar la data• Descargar la data y almacenarlo de manera local	 			
Almacenamiento de Datos <ul style="list-style-type: none">• Llevar la data al almacenamiento local de Hadoop• Crear el entorno para almacenar la data en hdfs• Almacenar la data en hdfs		  		
Procesamiento de Datos <ul style="list-style-type: none">• Analizar la data para definir la ruta de procesamiento• Desarrollar un script en PySpark para procesar la data• Ejecutar el script		 		
Carga de los Datos <ul style="list-style-type: none">• Almacenar la data procesada en HDFS• Crear la base de datos y una tabla en Hive que haga referencia a la data procesada			 	
Análisis de Datos <ul style="list-style-type: none">• Hacer la conexión ODBC entre PowerBI y Hive• Utilizar Power Query para dar formato adecuado a las columnas• Crear nuevas medidas y realizar un dashboard en PowerBI				  

ARQUITECTURA



FUENTES DE DATOS

Los datos fueron obtenidos de la pagina web de [datosabiertos.Gob.pe](https://datosabiertos.gob.pe) que presenta datos del Estado Peruano. La data originalmente es un archivo .csv, pesa un total de 219MB, delimitado por '|', tiene 429878 filas (incluido el índice) y 206 columnas.

•Fuente de datos: [data](#)

•Muestra de los datos:

```
HORA_CARGA|FECHACORTE|FECHAREGISTRO|CODIGO|NOMBRE|CATEGORIA|NIVEL|INSTITUCIÃ“N|GRUPO|SUB_GRUPO|MACROREGIONES|UBIGEO|REGION|PROVINCIA|DISTRITO|CUENTA_TRIAJE|
2023-06-20T06:11:05.683|2023-03-28|2023/03/28 11:17:43 AM|1|HOSPITAL IQUITOS "CESAR GARAYAR GARCIA"|II-2|Nivel 2|GOBIERNO REGIONAL|PÃŠBLICOS|MINSAGR|Zona Oriente|160101|LORE
2023-06-20T06:11:05.683|2023-04-06|2023/04/05 11:39:55 AM|1|HOSPITAL IQUITOS "CESAR GARAYAR GARCIA"|II-2|Nivel 2|GOBIERNO REGIONAL|PÃŠBLICOS|MINSAGR|Zona Oriente|160101|LORE
2023-06-20T06:11:05.683|2023-04-29|2023/04/29 7:50:54 AM|1|HOSPITAL IQUITOS "CESAR GARAYAR GARCIA"|II-2|Nivel 2|GOBIERNO REGIONAL|PÃŠBLICOS|MINSAGR|Zona Oriente|160101|LORE
2023-06-20T06:11:05.683|2023-03-08|2023/03/08 4:25:17 PM|1|HOSPITAL IQUITOS "CESAR GARAYAR GARCIA"|II-2|Nivel 2|GOBIERNO REGIONAL|PÃŠBLICOS|MINSAGR|Zona Oriente|160101|LORE
2023-06-20T06:11:05.683|2023-06-03|2023/02/10 6:35:18 PM|66|HOSPITAL II-1 SANTA CLOTILDE|II-1|Nivel 2|GOBIERNO REGIONAL|PÃŠBLICOS|MINSAGR|Zona Oriente|160107|LORETO|MAYNAS|N
2023-06-20T06:11:05.683|2023-01-06|2023/01/06 5:28:50 PM|308|HOSPITAL REGIONAL DOCENTE DE MEDICINA TROPICAL DR. JULIO CESAR DEMARINI CARO|II-2|Nivel 2|GOBIERNO REGIONAL|PÃŠBL
2023-06-20T06:11:05.683|2023-01-29|2023/01/29 4:22:34 PM|308|HOSPITAL REGIONAL DOCENTE DE MEDICINA TROPICAL DR. JULIO CESAR DEMARINI CARO|II-2|Nivel 2|GOBIERNO REGIONAL|PÃŠBL
2023-06-20T06:11:05.683|2022-12-10|2022/12/10 8:00:16 AM|210|HOSPITAL SANTA GEMA DE YURIMAGUAS|II-2|Nivel 2|GOBIERNO REGIONAL|PÃŠBLICOS|MINSAGR|Zona Oriente|160201|LORETO
2023-06-20T06:11:05.683|2023-05-21|2023/05/21 6:53:37 PM|520|DE APOYO FELIX MAYORCA SOTO|II-2|Nivel 2|GOBIERNO REGIONAL|PÃŠBLICOS|MINSAGR|Zona Centro|120701|JUNIN|TARMA|T/
2023-06-20T06:11:05.683|2023-05-16|2023/05/16 6:57:25 PM|519|DE APOYO JUNIN|II-1|Nivel 2|GOBIERNO REGIONAL|PÃŠBLICOS|MINSAGR|Zona Centro|120501|JUNIN|JUNIN|JUNIN|SI|56|SI|0|
2023-06-20T06:11:05.683|2023-06-12|2023/06/12 6:17:37 PM|432|DE APOYO MANUEL HIGA ARAKAKI|II-1|Nivel 2|GOBIERNO REGIONAL|PÃŠBLICOS|MINSAGR|Zona Centro|120601|JUNIN|SATIPO|S
2023-06-20T06:11:05.683|2023-03-27|2023/03/27 5:42:17 PM|520|DE APOYO FELIX MAYORCA SOTO|II-2|Nivel 2|GOBIERNO REGIONAL|PÃŠBLICOS|MINSAGR|Zona Centro|120701|JUNIN|TARMA|T/
2023-06-20T06:11:05.683|2022-12-30|2022/12/30 5:52:03 PM|979|DR.DANIEL ALCIDES CARRION GARCIA. |II-1|Nivel 2|GOBIERNO REGIONAL|PÃŠBLICOS|MINSAGR|Zona Centro|190113|PASCO|PASC
2023-06-20T06:11:05.683|2023-04-30|2023/04/30 6:38:02 PM|932|HOSPITAL DE TINGO MARIA (HOSPITAL DE CONTINGENCIA)|II-1|Nivel 2|GOBIERNO REGIONAL|PÃŠBLICOS|MINSAGR|Zona Oriente|
2023-06-20T06:11:05.683|2023-06-19|2023/06/19 6:54:24 PM|754|HOSPITAL REGIONAL HERMILIO VALDIZAN|II-2|Nivel 2|GOBIERNO REGIONAL|PÃŠBLICOS|MINSAGR|Zona Oriente|100101|HUANU
2023-06-20T06:11:05.683|2023-02-05|2023/02/05 6:40:11 PM|979|DR.DANIEL ALCIDES CARRION GARCIA. |II-1|Nivel 2|GOBIERNO REGIONAL|PÃŠBLICOS|MINSAGR|Zona Centro|190113|PASCO|PASC
2023-06-20T06:11:05.683|2023-04-29|2023/04/29 7:56:19 AM|26473|OBRA DE SAN CAMILO|II-E|Nivel 2|PRIVADO|PRIVADOS|PRIVADOS||150101|LIMA|LIMA|LIMA|NO|0|NO|0|0|0|0|0|0|0|0|0|
2023-06-20T06:11:05.683|2023-03-26|2023/03/26 6:18:54 PM|21508|SMQ QUIRUMEDIC SAC|II-E|Nivel 2|PRIVADO|PRIVADOS|PRIVADOS|Zona Norte|130101|LA LIBERTAD|TRUJILLO|TRUJILLO|NO|
2023-06-20T06:11:05.683|2023-04-20|2023/04/15 7:47:16 AM|18404|CLINICA SAN PABLO TRUJILLO|II-2|Nivel 2|PRIVADO|PRIVADOS|PRIVADOS|Zona Norte|130101|LA LIBERTAD|TRUJILLO|TRUJILLO|NO|
2023-06-20T06:11:05.683|2023-04-08|2023/04/08 5:15:26 PM|18580|GUILLERMO KAEIN DE LA FUENTE|II-2|Nivel 2|PRIVADO|PRIVADOS|PRIVADOS||150143|LIMA|LIMA|VILLA MARIA DEL TRIUNFO
2023-06-20T06:11:05.683|2023-04-08|2023/04/08 5:15:26 PM|18580|GUILLERMO KAEIN DE LA FUENTE|II-2|Nivel 2|PRIVADO|PRIVADOS|PRIVADOS||150143|LIMA|LIMA|VILLA MARIA DEL TRIUNFO
```

DICCIONARIO DE DATOS

El diccionario de datos explica el significa que tiene cada una de las columnas de la data original y lo presenta en un tabla de formato [columna, significado].

•**Diccionario de datos:** [diccionario](#)

•**Muestra del diccionario:**

FECHACORTE	Rango de tiempo de corte al que corresponde la información solicitada, deberán informar dos corte
FECHAREGISTRO	Fecha del registro realizado
CODIGO	CÓDIGO DE IPRESS: Código de 8 dígitos según el Registro Nacional de IPRESS - RENIPRESS administr
NOMBRE	Nombre de la IPRESS
CATEGORIA	Categoría vigente con la que cuenta el establecimiento de salud que registra los datos de camas; fu
NIVEL	Nilvel del establecimiento de salud que registra los datos de camas; fuente RENIPRESS.
INSTITUCIÓN	Institución o ámbito a la que pertenece el establecimiento de salud que registra los datos de camas
GRUPO	Grupo del establecimiento de salud que registra los datos de camas; fuente RENIPRESS.
SUB_GRUPO	Sub grupo del establecimiento de salud que registra los datos de camas; fuente RENIPRESS.
MACROREGIONES	Macroregión de la IPRESS
UBIGEO	Código de único del establecimiento salud que registra los datos de las camas, fuente RENIPRESS
REGION	Región de la IPRESS
PROVINCIA	Provincia de la IPRESS
DISTRITO	Distrito de la IPRESS
CUENTA_TRIAJE	¿CUENTA CON TRIAJE EN ZONA COVID-19?: Se entiende por Triaje en Zona Covid-19 al área de Tria
NU_ATENC_ULT	Es el número de atenciones en triaje diferenciado del servicio de emergencia para pacientes exclusi
CUENTA_ZC	IPRESS cuenta con Zona Diferenciada COVID 19
ZC_UCI_AACT_CAM_TOTAL	TOTAL - AMBIENTE DE ATENCIÓN CRITICO TEMPORAL (AACT) ZONA COVID-19: Es la suma de came
ZC_UCI_AACT_CAM_INOPERATIVOS	INOPERATIVAS - AMBIENTE DE ATENCIÓN CRITICO TEMPORAL (AACT) ZONA COVID-19: Es el núme
ZC_UCI_AACT_CAM_TOT_OPER	OPERATIVAS - AMBIENTE DE ATENCIÓN CRITICO TEMPORAL (AACT) ZONA COVID-19: Es el númerc
ZC_UCI_AACT_CAM_TOT_DISP	DISPONIBLES - AMBIENTE DE ATENCIÓN CRITICO TEMPORAL (AACT) ZONA COVID-19: Es el número
ZC_UCI_AACT_CAM_TOT_OCUP	OCUPADAS - AMBIENTE DE ATENCIÓN CRITICO TEMPORAL (AACT) ZONA COVID-19: Es el número d
ZC_UCI_AACT_COC_CAM_CONFIR	CONFIRMADO - AMBIENTE DE ATENCIÓN CRITICO TEMPORAL (AACT) ZONA COVID19: Es el número
ZC_UCI_AACT_COC_CAM_X_CONFIR	POR CONFIRMAR - AMBIENTE DE ATENCIÓN CRITICO TEMPORAL (AACT) ZONA COVID19: Es el núm
ZC_UCI_AACT_COO_CAM_CANULAS	CÁNULA DE ALTO FLUJO - AMBIENTE DE ATENCIÓN CRITICO TEMPORAL (AACT) ZONA COVID19: E
ZC_UCI_AACT_COO_CAM_SIN_VM	OCUPADAS SIN VENTILACIÓN MECÁNICA en AMBIENTE DE ATENCIÓN CRITICO TEMPORAL (AACT)
ZC_UCI_AACT_COO_CAM_CON_VM	OCUPADAS CON VENTILACIÓN MECÁNICA enAMBIENTE DE ATENCIÓN CRITICO TEMPORAL (AACT)
ZC_UCI_ADUL_CAM_TOTAL	TOTAL - UCI ADULTOS ZONA COVID-19: Es la suma de camas INOPERATIVAS más las camas OPERA'

TERMINOS IMPORTANTES

Dada la circunstancia de la pandemia y la demanda por las camas UCI se vio la necesidad de analizar solamente camas UCI de adultos, neonatales y pediátricos tanto en la zona covid y no covid. El objetivo es analizar cuantas camas están inoperativas, operativas y de estas cuantas disponibles y ocupadas. Este análisis ayudará al ministerio de salud a saber cuántas camas UCI inoperativas existen por región, provincia o distrito y tomar las acciones correspondientes. Además, podrá saber cuántas camas UCI están operativas y disponibles por región, provincia o distrito y con esto lograr una mejor gestión de las camas UCI disponibles para cubrir a la mayoría de pacientes necesitados. Toda esta información lo podrá saber para un año, mes o día concreto tanto en el sector público como privado.



TERMINOS IMPORTANTES

Dado que solo se esta analizando las camas UCI en adultos, neonatales y pedriatico en la zona covid y no covid, se están considerando solo las 24 columnas siguientes para el análisis final del número de camas:

Nro	Columna
1	ZC_UCI_ADUL_CAM_INOPERATIVOS
2	ZC_UCI_ADUL_CAM_TOT_OPER
3	ZC_UCI_ADUL_CAM_TOT_DISP
4	ZC_UCI_ADUL_CAM_TOT_OCUP
5	ZC_UCI_NEONATAL_CAM_INOPERATIVOS
6	ZC_UCI_NEONATAL_CAM_TOT_OPER
7	ZC_UCI_NEONATAL_CAM_TOT_DISP
8	ZC_UCI_NEONATAL_CAM_TOT_OCUP
9	ZC_UCI_PEDIA_CAM_INOPERATIVOS
10	ZC_UCI_PEDIA_CAM_TOT_OPER
11	ZC_UCI_PEDIA_CAM_TOT_DISP
12	ZC_UCI_PEDIA_CAM_TOT_OCUP
13	ZNC_UCI_ADUL_CAM_INOPERATIVO
14	ZNC_UCI_ADUL_CAM_OPERATIVO
15	ZNC_UCI_ADUL_CAM_DISPONIBLE
16	ZNC_UCI_ADUL_CAM_OCUPADO
17	ZNC_UCI_NEONATAL_CAM_INOPERATIVO
18	ZNC_UCI_NEONATAL_CAM_OPERATIVO
19	ZNC_UCI_NEONATAL_CAM_DISPONIBLE
20	ZNC_UCI_NEONATAL_CAM_OCUPADO
21	ZNC_UCI_PEDIA_CAM_INOPERATIVO
22	ZNC_UCI_PEDIA_CAM_OPERATIVO
23	ZNC_UCI_PEDIA_CAM_DISPONIBLE
24	ZNC_UCI_PEDIA_CAM_OCUPADO

DESARROLLO DEL HITO 1: INGESTA DE DATOS

1. La data se obtuvo de [datosabiertos.Gob.pe](https://datosabiertos.gob.pe) del siguiente link: [data](#), donde se puede descargar la data y el diccionario de datos.
2. Se descargó y se almacenó en la máquina local.

 Data	24/06/2023 07:28 a. m.	Documento de tex...	224,127 KB
 Diccionario de datos	15/08/2023 11:01 a. m.	Hoja de cálculo d...	19 KB

DESARROLLO DEL HITO 2: ALMACENAMIENTO DE DATOS

1. Se llevo la data original y el script, para el procesamiento, al almacenamiento local de la máquina virtual donde esta Hadoop, esto se hizo con WinSCP.

C:\Users\jxtr\Downloads\				/home/maria_dev/					
Nombre	Tamaño	Tipo	Modificado	Nombre	Tamaño	Modificado	Permisos	Propieta...	
.		Directorio superior	25/10/2023 02:19:15 a. m.	.		18/06/2018 10:30:38 a. m.	rw-r--r--	root	
Data.txt	224,127 KB	Documento de tex...	24/06/2023 07:28:31 a. m.	Data.txt	224,127 KB	24/06/2023 07:28:31 a. m.	rw-rw-r--	maria_d...	
spark_proc.py	5 KB	Python File	25/10/2023 02:19:55 a. m.	spark_proc.py	5 KB	25/10/2023 02:19:55 a. m.	rw-rw-r--	maria_d...	

2. Se creo una carpeta en HDFS para almacenar la data sin procesar y data procesada.

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -mkdir /user/maria_dev/proyecto_de
```

3. Se llevo la data del almacenamiento local de Hadoop al HDFS

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put Data.txt /user/maria_dev/proyecto_de
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls proyecto_de
Found 1 items
-rw-r--r--  1 maria_dev hdfs  229505664 2023-10-25 08:29 proyecto_de/Data.txt
[maria_dev@sandbox-hdp ~]$
```

DESARROLLO DEL HITO 3: PROCESAMIENTO DE DATOS

1. Se analizo la data y para definir un ruta de procesamiento que se describe de manera general en la sección “Terminos Importantes” del desarrollo de este proyecto, pero resumiendo la ruta sería:
 1. Seleccionar columnas importantes
 2. Eliminar filas con valores nulos asociados solo a algunas columnas
 3. Llenar valores nulos con ‘0’ que representa el número de camas
 4. Generar columnas Año, Mes y Dia en base a la columna “FECHAREGISTRO”
 5. Realizar la operación unpivot (eliminar dinamización) a las columnas que representan cantidad de camas. Esto agrego 2 columnas extras: “ATRIBUTO” y “CAMAS”
 6. De la columna “ATRIBUTO” se genero 3 columnas extras: “USO”, “GRUPO ETARIO” y “ESTADO”
 7. Se reemplazo los valores de las 3 columnas generadas con valores adecuados. Ejm: ZC -> COVID, ZNC -> NO COVID, etc.
 8. Se guarda la data procesada en un archivo csv en HDFS

DESARROLLO DEL HITO 3: PROCESAMIENTO DE DATOS

2. El script de procesamiento se realizó con PySpark y está disponible en el repositorio público del proyecto: [repositorio](#)
3. Para ejecutar el script, este primero debe estar en el almacenamiento local de hadoop (se muestra en el paso 1 del desarrollo del hito 2), luego se ejecuta con el siguiente comando:

```
[maria_dev@sandbox-hdp ~]$ spark-submit --master yarn spark_proc.py
```

DESARROLLO DEL HITO 4: CARGA DE LOS DATOS

1. El script para el procesamiento ya almacena la data procesada en la ruta HDFS donde se esta realizando el proyecto

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls proyecto_de/DataProcesada
Found 3 items
-rw-r--r--  1 maria_dev hdfs          0 2023-10-25 08:40 proyecto_de/DataProcesada/_SUCCESS
-rw-r--r--  1 maria_dev hdfs 418016630 2023-10-25 08:39 proyecto_de/DataProcesada/part-00000-8cdd211c-fe38-401e-903f-170688548775-c000.csv
-rw-r--r--  1 maria_dev hdfs 386889768 2023-10-25 08:40 proyecto_de/DataProcesada/part-00001-8cdd211c-fe38-401e-903f-170688548775-c000.csv
[maria_dev@sandbox-hdp ~]$
```

2. Crear la base de datos y la tabla que haga referencia donde se almaceno la data procesada (esto se realizo en la UI Web de Apache Ambari)

```
1 create database Proyecto_DE
```

DESARROLLO DEL HITO 4: CARGA DE LOS DATOS

DATABASE

Select or search database/schema

×

proyecto_de

```
1 create external table proyecto_de.Resultado_ETL (  
2   ANIO STRING,  
3   MES STRING,  
4   DIA STRING,  
5   AMBITO STRING,  
6   REGION STRING,  
7   PROVINCIA STRING,  
8   DISTRITO STRING,  
9   USO STRING,  
10  GRUPO_ETARIO STRING,  
11  ESTADO STRING,  
12  CAMAS INT)  
13  row format delimited  
14  fields terminated by ','  
15  location '/user/maria_dev/proyecto_de/DataProcesada'
```

✓ Execute

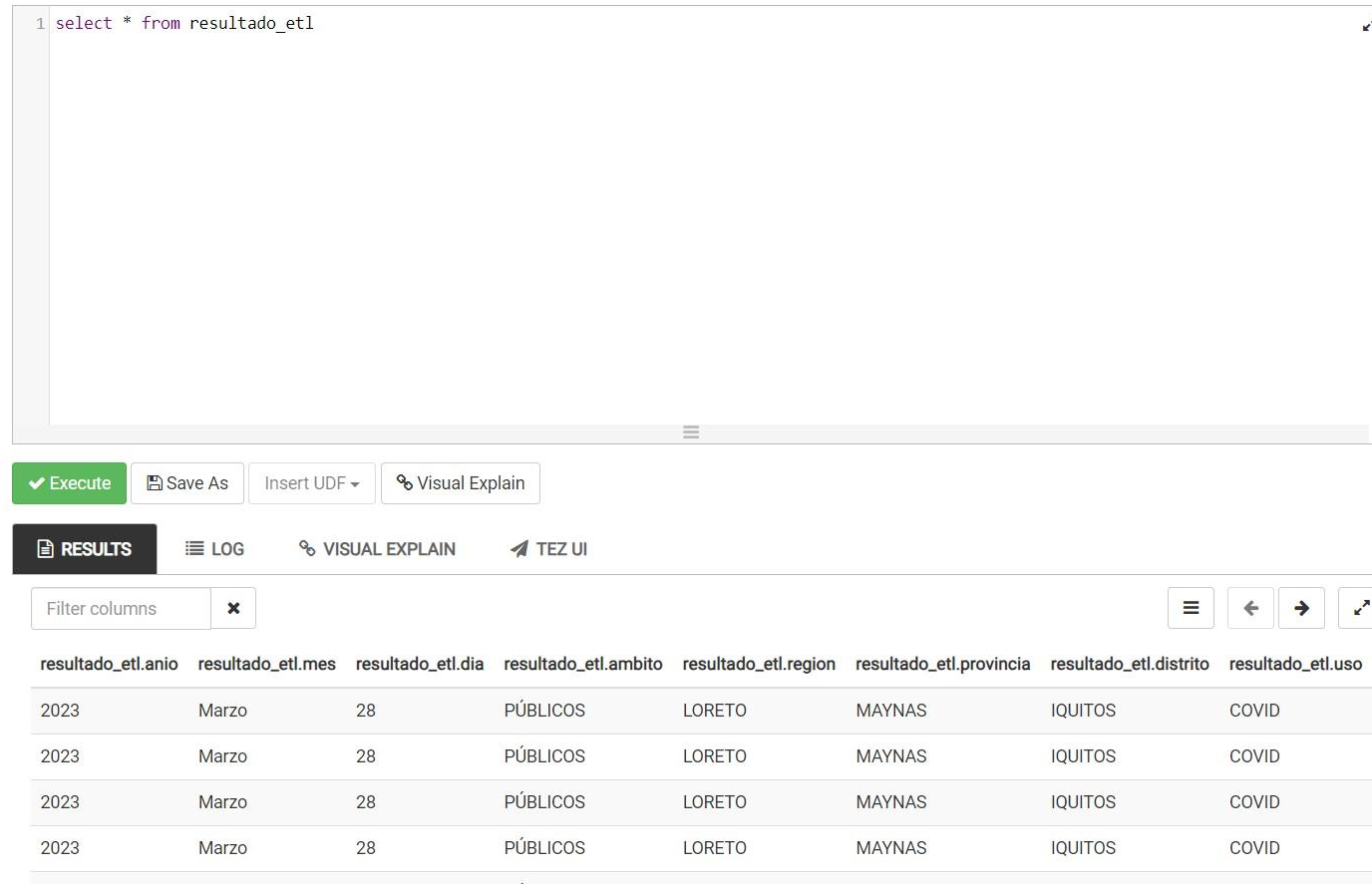
📄 Save As

Insert UDF ▾

🔗 Visual Explain

DESARROLLO DEL HITO 4: CARGA DE LOS DATOS

Ahora se va a verificar si la data procesada se encuentra en el datawarehouse (Hive)

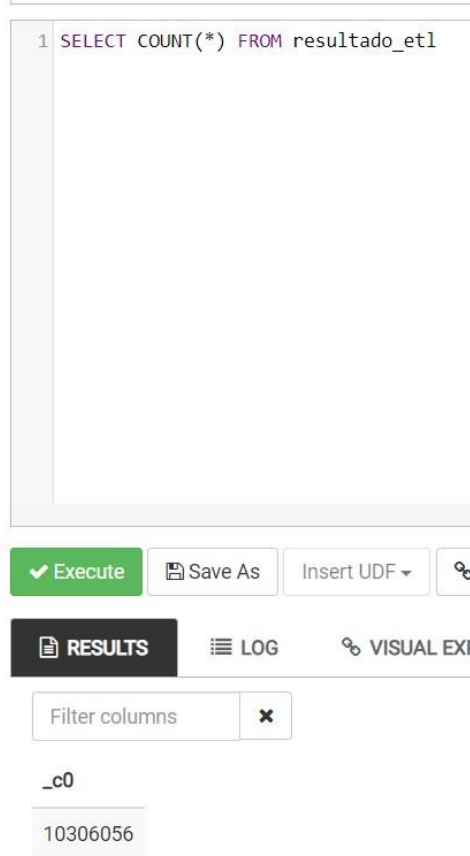


The screenshot shows a SQL query execution interface. At the top, a text area contains the query: `1 select * from resultado_etl`. Below the text area are buttons for **Execute** (green), **Save As**, **Insert UDF**, and **Visual Explain**. Below these buttons is a navigation bar with **RESULTS** (selected), **LOG**, **VISUAL EXPLAIN**, and **TEZ UI**. Below the navigation bar is a table with 8 columns: `resultado_etl.anio`, `resultado_etl.mes`, `resultado_etl.dia`, `resultado_etl.ambito`, `resultado_etl.region`, `resultado_etl.provincia`, `resultado_etl.districto`, and `resultado_etl.uso`. The table contains 4 rows of data, all with the same values: 2023, Marzo, 28, PÚBLICOS, LORETO, MAYNAS, IQUITOS, and COVID.

resultado_etl.anio	resultado_etl.mes	resultado_etl.dia	resultado_etl.ambito	resultado_etl.region	resultado_etl.provincia	resultado_etl.districto	resultado_etl.uso
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITOS	COVID
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITOS	COVID
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITOS	COVID
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITOS	COVID

DESARROLLO DEL HITO 4: CARGA DE LOS DATOS

La data procesada tiene un total de 10306056 filas



DESARROLLO DEL HITO 5: ANÁLISIS DE DATOS

1. Se hizo la conexión ODBC entre PowerBI y Hive

The screenshot displays the 'Obtener datos' (Get Data) window in Power BI. On the left, under 'Otras' (Others), the 'ODBC' connector is selected. The 'Navegador' (Navigator) pane on the right shows the hierarchy: ODBC (dsn=Hive_Data) [1] > HIVE [4] > proyecto_de [1] > resultado_etl. The 'resultado_etl' table is selected, and a preview is shown on the right. The preview table has 20 rows and 7 columns: año, mes, día, ambito, region, provincia, and distrito. All data in the preview is identical: 2023, Marzo, 28, PÚBLICOS, LORETO, MAYNAS, IQUITO.

Obtener datos

Buscar

Todo

Archivo

Base de datos

Microsoft Fabric (versión preliminar)

Power Platform

Azure

Servicios en línea

Otras

Otras

- Web
- Lista de SharePoint
- Fuente OData
- Active Directory
- Microsoft Exchange
- Archivo Hadoop (HDFS)
- Spark
- Hive LLAP
- Script de R
- Script de Python
- ODBC
- Importa datos de ODBC.
- BitSight Security Ratings
- BQE Core
- Celonis EMS (beta)

Navegador

Opciones de presentación

- ODBC (dsn=Hive_Data) [1]
 - HIVE [4]
 - foodmart
 - laboratorio
 - migracion
 - proyecto_de [1]
 - resultado_etl

resultado_etl

Vista previa descargada el lunes

año	mes	día	ambito	region	provincia	distrito
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO
2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS	IQUITO

Conectores certificados | Aplicaciones de plantilla

Conectar Cancelar

Seleccionar tablas relacionadas

Cargar Transformar datos Cancelar

DESARROLLO DEL HITO 5: ANÁLISIS DE DATOS

2. Se utilizó Power Query solo para corregir el nombre de las columnas y darles el formato adecuado

Power Query Editor showing the M formula and the resulting data table.

M Formula: `= Table.RenameColumns(resultado_etl_Table,{{"anio", "Año"}, {"mes", "Mes"}, {"dia", "Día"}, {"ambito", "Ambito"}, {"region", "Region", "Provincia"}}`

	Año	Mes	Día	Ambito	Region	Provincia
1	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
2	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
3	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
4	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
5	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
6	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
7	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
8	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
9	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
10	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
11	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
12	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
13	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
14	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
15	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
16	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
17	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS
18	2023	Marzo	28	PÚBLICOS	LORETO	MAYNAS

Configuración de la consulta

PROPIEDADES

Nombre: resultado_etl

PASOS APLICADOS

- Origen
- Navegación
- Columnas con nombre cambiado
- Tipo cambiado

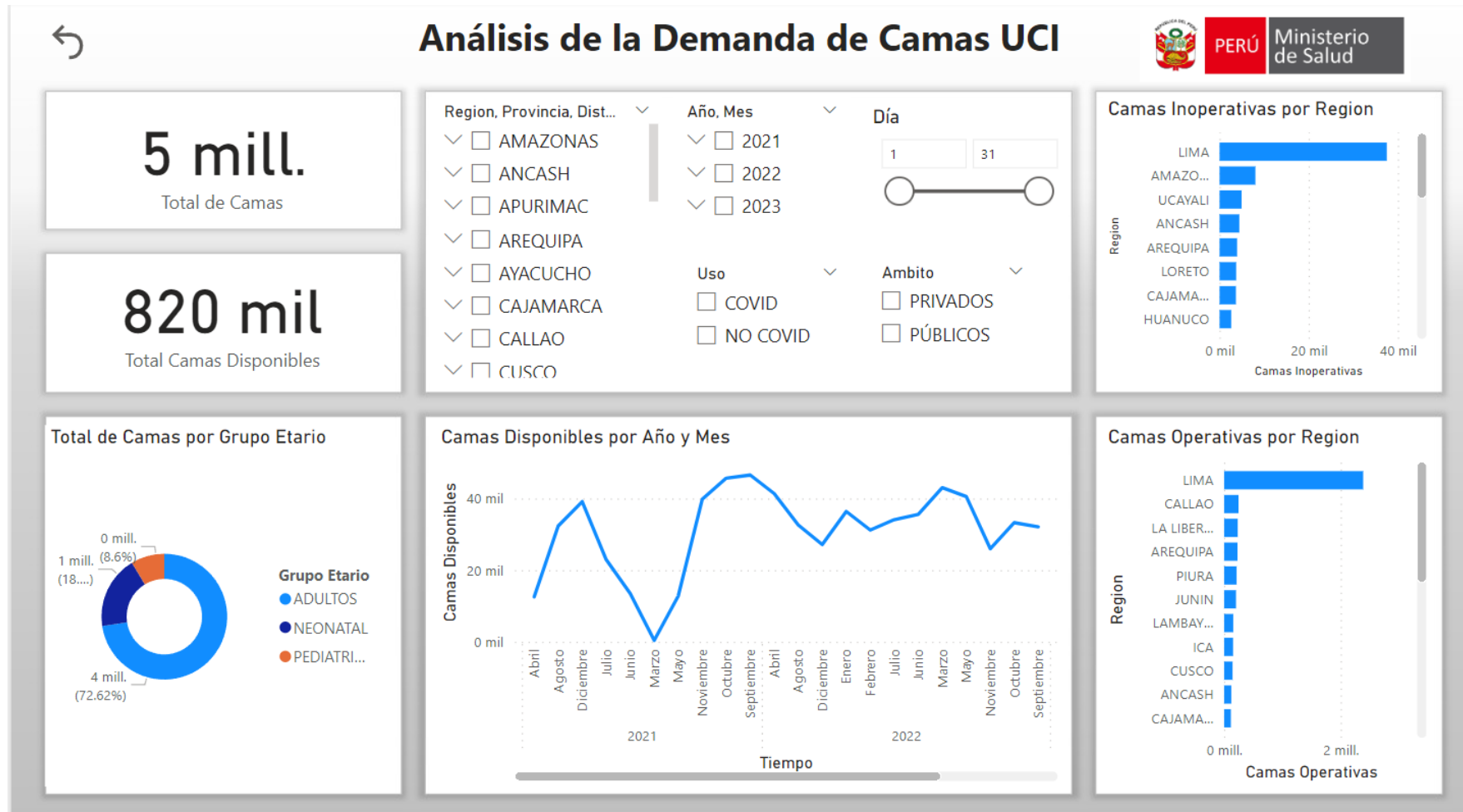
DESARROLLO DEL HITO 5: ANÁLISIS DE DATOS

3. Se crearon nuevas medidas y se hizo un dashboard



DESARROLLO DEL HITO 5: ANÁLISIS DE DATOS

El dashboard es público y lo puedes ver en el siguiente enlace: [Dashboard](#)



CONCLUSIONES

1. Se logro de manera satisfactoria concluir el proyecto y confirmar la rapidez de procesamiento de Spark para grandes volúmenes de datos en un entorno distribuido.
2. Para proyectos futuros se buscara trabajar con una mayor cantidad de datos de nivel de PetaBytes (PB) y automatizar todo el proceso ETL mediante un script de bash.

● Síguenos en :



Data Growth Community



Data Growth Community



Data Growth Community



Data Growth
Community

"Potenciando el crecimiento colectivo"
