



PRIMERA PRACTICA CALIFICADA SISTEMA DE INTELIGENCIA DE NEGOCIOS (SI 807-U)

Profesor: Dr. Ing. Aradiel Castañeda, Hilario

Ayudante de Cátedra: García Atuncar, Fernando

Ciclo: 2025 B

Fecha: 22-09-25

Github: [David1712uni/Contralor-a-SIN](https://github.com/David1712uni/Contralor-a-SIN)

INTEGRANTES:

Código UNI	Apellidos y Nombres	Correo Electrónico	Tareas realizadas
20220008E	Andrade Saavedra, Navhi Giordano	navhi.andrade.s@uni.pe	<ul style="list-style-type: none">• Instalar Hortonworks Sandbox en una máquina virtual (VirtualBox).• Comprobar servicios activos (HDFS, Hive, Spark) mediante capturas de pantalla y comandos (hdfs dfs -ls /, spark-shell --version, etc.).• Subir archivos de datos (CSV/XLSX) al HDFS y mostrar evidencia.• Elaborar un diagrama de arquitectura inicial con flujo de datos (fuentes → ingesta → HDFS/Hive/Spark → usuarios).• Redactar el informe académico, con citas bibliográficas breves y estandarización de tablas.• Preparar y cargar en GitHub todos los entregables: informe, capturas, scripts/queries, diagrama y README.
20220122B	Caruzo Cieza, David	david.caruzo.c@uni.pe	<ul style="list-style-type: none">• Definir los objetivos del trabajo alineados a la estrategia del curso.• Describir el contexto de la empresa (nombre, misión, visión, productos, clientes, organigrama).
20200298H	Carhuas Romero Jhon Jesus	jhon.carhuas.r@uni.pe	<ul style="list-style-type: none">• Plantear las necesidades de información a nivel estratégico, táctico y operativo.• Formular al menos 10 KPI's iniciales, cada uno con nombre, definición, fórmula, unidad, frecuencia y nivel de decisión.

Objetivos

- Comprender el contexto estratégico de la empresa seleccionada.
- Identificar problemas de negocio y necesidades de información.
- Definir los KPI's iniciales alineados a la estrategia.



- Implementar y documentar la arquitectura base en Hortonworks Sandbox (VM VirtualBox).
- Entregar evidencia de un entorno Hadoop & Spark funcional.

Alcance

La práctica abarca desde el análisis del negocio hasta la puesta en marcha de la arquitectura técnica mínima:

- Análisis organizacional (misión, visión, cadena de valor, procesos).
- Identificación de problemas y necesidades de información.
- Definición de al menos 10 KPI's iniciales.
- Instalación de Hortonworks en VM (VirtualBox).
- Evidencia de servicios habilitados (HDFS, Hive, Spark).
- Diagrama de arquitectura inicial.

1. ENTORNO DE LA EMPRESA SELECCIONADA

1.1 Generalidades de la Empresa

Nombre o razón social	SEGURO SOCIAL DE SALUD
Giro de la empresa	Prestación de servicios de salud pública y aseguramiento social.
Ubicación	Sede central: Lima; red nacional de hospitales y establecimientos en todo el Perú.
Misión y visión	<p>Misión: Brindamos prestaciones de salud, económicas y sociales a nuestros asegurados con una gestión eficiente e innovadora que garantiza la protección financiera de las prestaciones integrales.</p> <p>Visión: Ser una institución moderna y en mejora continua, centrada en los asegurados, que garantiza el acceso a la seguridad social en salud con ética, oportunidad y calidad.</p>
Productos y clientes	<p>Productos: Servicios hospitalarios, prevención de enfermedades</p> <p>Clientes: asegurados, población regional</p>
Organigrama	ANEXO 1



1.1. Identificación de Problemas del Negocio

1. Falta de integración de datos entre hospitales.

- Justificación:** Cada hospital maneja sistemas de información distintos o aislados, sin interoperabilidad ni un repositorio centralizado. Esto dificulta consolidar datos en tiempo real.
- Impacto:** Genera duplicidad de información, retrasos en la atención coordinada de pacientes, imposibilidad de tener una visión nacional de la salud y decisiones basadas en datos fragmentados.

2. Desabastecimiento de medicamentos y falta de entrega a pacientes.

- Justificación:** No se cuenta con la cantidad necesaria de medicamentos en inventario para poder abastecer a todos los pacientes que lo necesitan
- Impacto:** Evita tratamientos tempranos de enfermedades y baja la calidad del servicio en total.

3. Infraestructura y equipamiento médico deficiente.

- Justificación:** Las sedes alrededor del país no se encuentran en condiciones óptimas para atender a la demanda presente.
- Impacto:** Disminuye la calidad del servicio y evita tratamientos específicos de cierto equipaje médico

4. Uso intensivo de procesos manuales que generan errores y pérdida de trazabilidad.

- Justificación:** La dependencia en registros físicos o digitación manual incrementa el margen de error humano y no permite rastrear el flujo de datos de forma transparente.
- Impacto:** Se producen inconsistencias en los registros médicos, pérdida de información valiosa para la gestión y reducción de la confianza en los reportes generados.

5. Falta de efectividad en las campañas de salud

- Justificación:** Las campañas no se sustentan en análisis predictivo ni segmentación de la población, y muchas veces se planifican de forma reactiva.
- Impacto:** Se desperdician recursos económicos y humanos, la población objetivo no recibe la atención adecuada y se reduce el impacto preventivo en la salud pública.

1.2. Necesidades de Información y Decisiones Críticas.

Nivel	Tipo de decisión	Necesidad de información	Problema Relacionado
Estratégico	Definir políticas de salud, asignar presupuesto nacional	Indicadores consolidados de incidencia y prevalencia, análisis de costos, cobertura de asegurados, efectividad de campañas de prevención	2, 3, 4, 5
Táctico	Planificación regional y asignación de recursos	Reportes comparativos por hospital/región, ocupación de camas, stock de medicamentos, tiempo de espera promedio, indicadores de desempeño regional	1, 2, 3, 4, 5
Operativo	Gestión diaria y respuesta a emergencias	Datos en tiempo real: atenciones nuevas por día, resultados de laboratorio, disponibilidad de UCI, personal disponible, trazabilidad de pacientes.	3, 4, 5

1.2.1. Problema elegido :

Se ha elegido el problema de predicción de personas que sufriran de las siguientes condiciones específicas de: **Diabetes, Hipertensión y Obesidad**



Facultad de Ingeniería Industrial y Sistemas Escuela Profesional de Ingeniería de Sistemas

en determinadas ubicaciones geográficas, dado que representa un punto crítico para mejorar la planificación preventiva y la preparación de EsSalud. Los principales motivos son:

1. Disponibilidad de data confiable para su análisis:

En la plataforma de Datos Abiertos del Gobierno del Perú (<https://datosabiertos.gob.pe/group/seguro-social-de-salud-essalud>

) se encuentran bases de datos oficiales relacionadas con atenciones médicas, diagnósticos y distribución demográfica. Estas fuentes permiten construir modelos predictivos basados en evidencia, identificar patrones epidemiológicos y anticipar la demanda futura de servicios de salud.

2. Impacto directo en la planificación y prevención

La capacidad de anticipar qué poblaciones y zonas tendrán mayor riesgo de enfermedades permite asignar recursos médicos, medicamentos e infraestructura de forma más eficiente. Esto reduce costos a largo plazo y mejora la calidad de la atención al prevenir la saturación hospitalaria.

3. Relación con otros problemas identificados

La falta de integración de datos entre hospitales, los reportes tardíos y el uso intensivo de procesos manuales afectan directamente la posibilidad de construir modelos predictivos confiables. Abordar este problema implica fortalecer la interoperabilidad de los sistemas y estandarizar indicadores de salud a nivel nacional.

4. Alta relevancia social y estratégica

La predicción de enfermedades no solo tiene un impacto en la gestión de EsSalud, sino también en la percepción ciudadana. Una institución capaz de anticiparse a los brotes y planificar campañas con base en datos genera mayor confianza, legitimidad institucional y contribuye a la salud pública del país en su conjunto. a positiva en la percepción pública y en la legitimidad institucional.

1.3. KPI's Iniciales

Nombre del KPI	Descripción	Fórmula	Unidad de medida	Frecuencia
Tasa de diagnósticos por 1000 habitantes	Mide la frecuencia de diagnósticos de estas patologías en la población.	$(N^{\circ} \text{ diagnósticos} \div \text{Población total}) \times 1000$	Casos por 1000 hab.	Mensual
Tendencia de crecimiento de casos	Mide la variación porcentual de diagnósticos en el tiempo para anticipar zonas de riesgo.	$((\text{Casos periodo actual} - \text{Casos periodo anterior}) \div \text{Casos periodo})$	%	Mensual



		anterior) $\times 100$		
Concentración geográfica de diagnósticos	Identifica regiones o distritos con mayor número de casos diagnosticados.	(Casos en región \div Total de casos) $\times 100$	%	Trimestral
Edad promedio de diagnóstico	Estima la edad promedio de los pacientes al momento del diagnóstico, útil para segmentar campañas.	Σ (Edad de diagnóstico) \div N pacientes	Años	Trimestral
Tiempo promedio entre diagnóstico y control	Mide días transcurridos entre diagnóstico inicial y primer control posterior.	Σ (Fecha control – Fecha diagnóstico) \div N pacientes	Días	Semestral
Distribución por sexo de diagnósticos	Diferencia en proporción de diagnósticos por sexo (Paciente).	(Casos sexo \div Total casos) $\times 100$	%	Trimestral
Índice de diagnósticos en población joven (<40 años)	Casos de hipertensión/diabetes/obesidad en menores de 40 años (Paciente + Diagnóstico).	(Casos <40 años \div Total casos) $\times 100$	%	Trimestral
Disponibilidad de especialistas por paciente	Relación médicos especialistas/pacientes diagnosticados (Médico + Diagnóstico).	N Médicos \div N pacientes atendidos	Ratio	Trimestral
Cobertura de diagnósticos por red hospitalaria	Casos diagnosticados en cada red vs asegurados en esa red (IPRESS + Ubigeo).	(Pacientes diagnosticados \div Población asegurada en red) $\times 100$	%	Trimestral
Variabilidad regional de diagnósticos	Diferencia entre la región con mayor y menor tasa de diagnósticos (Ubigeo + Diagnóstico).	Máx(tasa diagnósticos) – Mín(tasa diagnósticos)	%	Trimestral

2. EVIDENCIA TÉCNICA

2.1. Implementación de Hortonworks

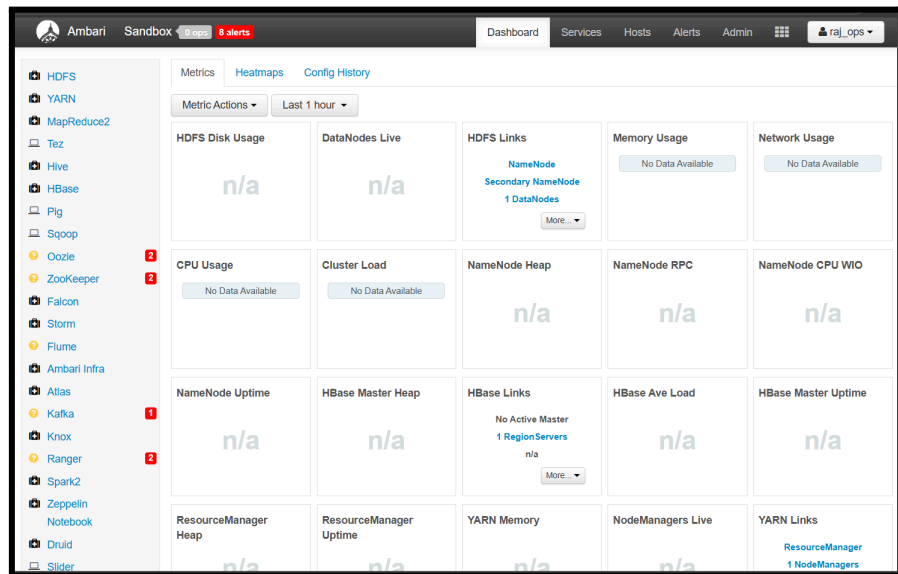
- ✓ Capturas de pantalla de la VM mostrando:
 - Ambari con servicios en ejecución, y modificación de contraseñas.

1. Modificación de Contraseñas



```
localhost:4200
sandbox-hdp login: root
root@sandbox-hdp.hortonworks.com's password:
You are required to change your password immediately (root enforced)
Last failed login: Sun Sep 21 23:45:36 UTC 2025 from 172.18.0.2 on ssh:notty
There were 4 failed login attempts since the last successful login.
Last login: Mon Jun 18 15:28:54 2018 from 172.17.0.2
Changing password for root.
(current) UNIX password:
New password:
BAD PASSWORD: The password is shorter than 7 characters
New password:
Retype new password:
[root@sandbox-hdp ~]#
```

2. Ambari inicial:



○ Comando hdfs dfs -ls /

```
[root@sandbox-hdp ~]# hdfs dfs -ls /
Found 13 items
drwxr-xr-x - maria_dev hdfs 0 2025-09-22 04:19 /Data
drwxr-xr-x - maria_dev hdfs 0 2025-09-16 02:53 /Laboratorio1
drwxrwxrwx - yarn hadoop 0 2018-06-18 15:18 /app-logs
drwxr-xr-x - hdfs hdfs 0 2018-06-18 16:13 /apps
drwxr-xr-x - yarn hadoop 0 2018-06-18 14:52 /ats
drwxr-xr-x - hdfs hdfs 0 2018-06-18 14:52 /hdp
drwx----- - livy hdfs 0 2018-06-18 15:11 /livy2-recovery
drwxr-xr-x - mapred hdfs 0 2018-06-18 14:52 /mapred
drwxrwxrwx - mapred hadoop 0 2018-06-18 14:52 /mr-history
drwxr-xr-x - hdfs hdfs 0 2018-06-18 15:59 /ranger
drwxrwxrwx - spark hadoop 0 2025-09-16 03:15 /spark2-history
drwxrwxrwx - hdfs hdfs 0 2018-06-18 16:06 /tmp
drwxr-xr-x - hdfs hdfs 0 2018-06-18 16:08 /user
```

```
[root@sandbox-hdp ~]# hdfs dfs -ls /Data
Found 9 items
drwxr-xr-x - hive hdfs 0 2025-09-22 04:19 /Data/DF_CExterna_2015_2022
-rw-r--r-- 1 maria_dev hdfs 9948919 2025-09-22 03:27 /Data/DF_CExterna_2015_2022.csv
-rw-r--r-- 1 maria_dev hdfs 30586787 2025-09-22 03:33 /Data/DF_ExLab_CExt_Diabetes.csv
-rw-r--r-- 1 maria_dev hdfs 7406601 2025-09-22 03:34 /Data/DF_ExLab_CExt_EnfermedadRenal.csv
-rw-r--r-- 1 maria_dev hdfs 6152662 2025-09-22 03:35 /Data/DF_ExLab_CExt_Hiperlipidemia.csv
-rw-r--r-- 1 maria_dev hdfs 24308336 2025-09-22 03:30 /Data/DF_ExLab_CExt_Hipertension.csv
-rw-r--r-- 1 maria_dev hdfs 16454244 2025-09-22 03:37 /Data/DF_ExLab_CExt_Obesidad.csv
-rw-r--r-- 1 maria_dev hdfs 611446 2025-09-22 03:39 /Data/Planeamiento_Estrategico_Ubigeo.csv
-rw-r--r-- 1 maria_dev hdfs 123241 2025-09-22 03:39 /Data/geodir-ubigeo-inei.csv
```

○ spark-shell -version

Facultad de Ingeniería Industrial y Sistemas
Escuela Profesional de Ingeniería de Sistemas

```
[root@sandbox-hdp ~]# spark-shell -version
SPARK_MAJOR_VERSION is set to 2, using Spark2

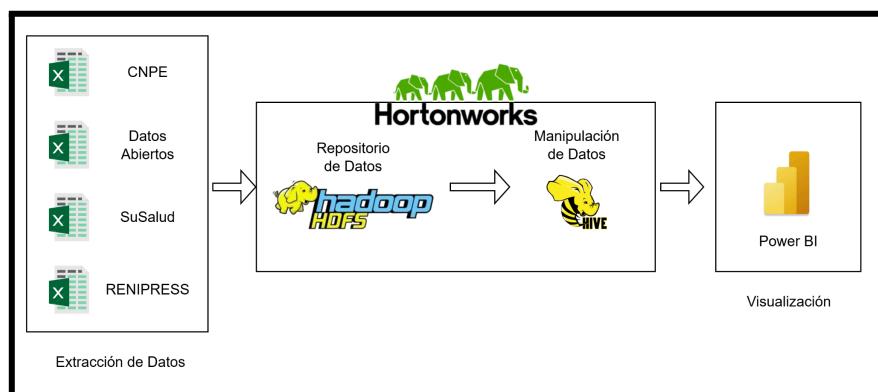
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://sandbox-hdp.hortonworks.com:4040
Spark context available as 'sc' (master = local[*], app id = local-1758581003522).
Spark session available as 'spark'.
Welcome to

      ____
     / ___ \
    /  _ < 
   /  / \  \
  /_____\  \
         \_/_

 version 2.3.0.2.6.5.0-292

Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM, Java 1.8.0_171)
Type in expressions to have them evaluated.
Type :help for more information.
```

2.2. Diagrama de Arquitectura Inicial



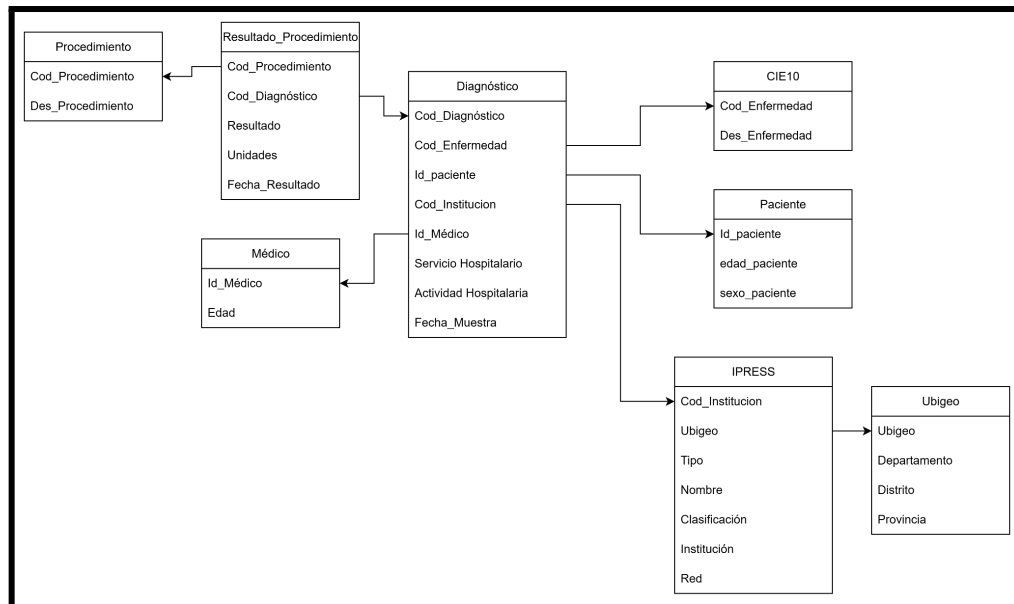
- **Fuente de datos:**

Repositorio	Archivo	Descripción
Centro Nacional de Planeamiento Estratégico	Datos-planeamiento-estrategico.xlsx	Incluye descripciones de cada ubigeo
Datos Abiertos Perú	DF_ExLab_CExt_Diabetes.csv	Contiene todas las consultas externas con diagnósticos de diabetes y sus detalles desde el 2020 hasta el 2024
	DF_ExLab_CExt_Hipertension.csv	Contiene todas las consultas externas con diagnósticos de hipertensión y sus detalles desde el 2020 hasta el 2024
	DF_ExLab_CExt_Obesidad.csv	Contiene todas las consultas externas con diagnósticos de obesidad y sus detalles desde el 2020 hasta el 2024
	Ubigeo.csv	Contiene el código de ubigeo, su distrito, provincia y departamento



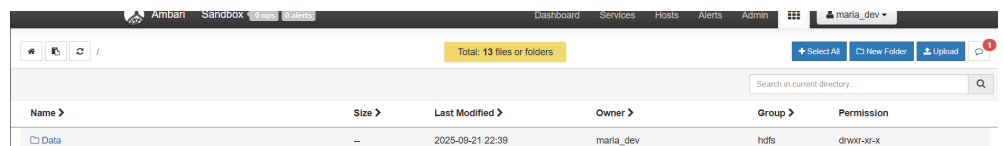
SuSalud	CIE10.csv	Contiene el código de todas la enfermedades y su descripción
RENIPRESS	USLRC20250923.xls	Contiene información de todos los hospitales del Perú

- **Modelamiento de Datos:**

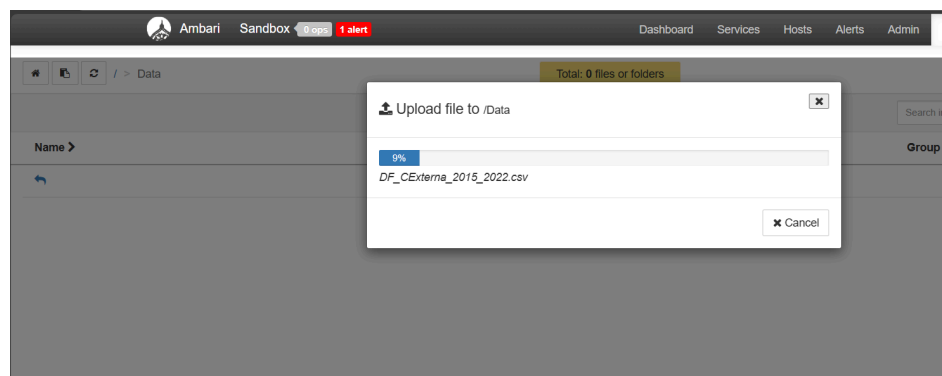


- **Ingesta de datos:**

1. Creando la carpeta de Data:



2. Subiendo los archivos CSV:



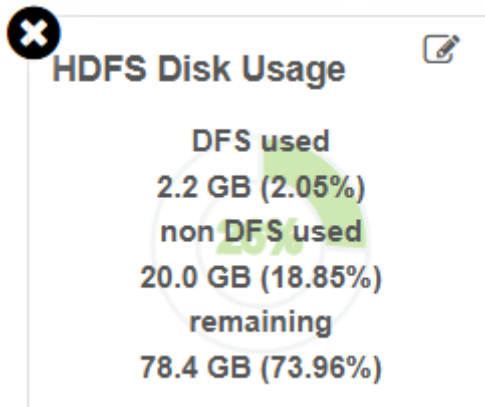
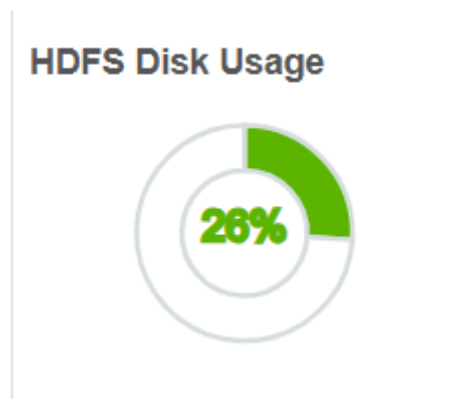
3. Archivos Subidos:



Name	Size	Last Modified	Owner	Group	Permission
DF_Externa_2015_2022.csv	9.5 MB	2025-09-21 22:27	maria_dev	hdfs	-rw-r--r--
DF_ExLab_CExt_Diabetes.csv	29.2 MB	2025-09-21 22:33	maria_dev	hdfs	-rw-r--r--
DF_ExLab_CExt_EnfermedadRenal.csv	7.1 MB	2025-09-21 22:34	maria_dev	hdfs	-rw-r--r--
DF_ExLab_CExt_Hipertension.csv	5.9 MB	2025-09-21 22:35	maria_dev	hdfs	-rw-r--r--
DF_ExLab_CExt_Obesidad.csv	23.2 MB	2025-09-21 22:30	maria_dev	hdfs	-rw-r--r--
DF_ExLab_CExt_Obesidad.csv	15.7 MB	2025-09-21 22:37	maria_dev	hdfs	-rw-r--r--
Planeamiento_Estrategico_Ubigeo.csv	597.1 KB	2025-09-21 22:39	maria_dev	hdfs	-rw-r--r--
geodir-ubigeo-inei.csv	120.4 KB	2025-09-21 22:39	maria_dev	hdfs	-rw-r--r--

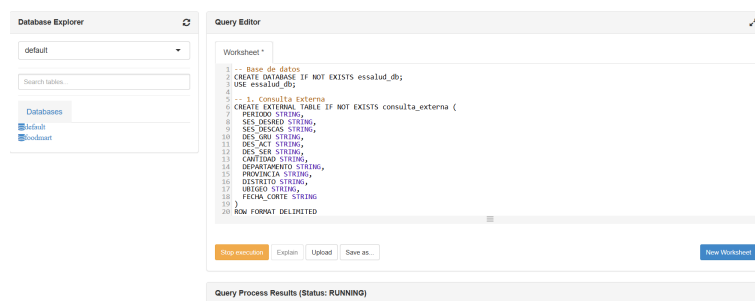
- **HDFS:**

- Función: Es el sistema de almacenamiento distribuido de Hadoop, diseñado para guardar grandes volúmenes de datos dividiéndolos en bloques y replicándolos en distintos nodos, garantizando alta disponibilidad, tolerancia a fallos y acceso eficiente a la información.



- **Hive:**

- Función: Es una herramienta de *data warehouse* que permite consultar y analizar datos almacenados en Hadoop mediante un lenguaje similar a SQL (HiveQL). Facilita a los analistas trabajar con datos masivos sin necesidad de programar en Java o MapReduce.





```
Worksheet
1 USE essalud_db;
2
3 CREATE EXTERNAL TABLE IF NOT EXISTS consulta_externa (
4     PERIODO STRING,
5     SES_DESCED STRING,
6     SES_DESCAS STRING,
7     DES_GRU STRING,
8     DES_ACT STRING,
9     DES_SER STRING,
10    CANTIDAD STRING,
11    DEPARTAMENTO STRING,
12    PROVINCIA STRING,
13    DISTRITO STRING,
14    UBIGEO STRING,
15    FECHA_CORTE STRING
16 )
17 ROW FORMAT DELIMITED
18 FIELDS TERMINATED BY ','
19 STORED AS TEXTFILE
20 LOCATION '/Data/DF_CExterna_2015_2022'
```

Execute Explain Upload Save as...

Query Process Results (Status: SUCCEEDED)

3. Referencias

- EsSalud (2025). Planeamiento Estratégico. Extraído de: [6474165-pei-2025-2030-diagramado-vf-13-02-25.pdf](#)
- Gob.pe (2025). Sedes EsSalud. Extraído de: [Sedes - Seguro Social de Salud - Plataforma del Estado Peruano](#)
- SuSalud (2025). CIE 10. Extraído de: <http://datos.susalud.gob.pe/dataset/cie-10>
- RENIPRESS (2025). Establecimientos registrados en Renipress. Extraído de: <http://app20.susalud.gob.pe:8080/registro-renipress-webapp/listadoEstablecimientosRegistrados.htm?action=mostrarBuscar#no-back-button>
- Gob.Pe (2025). Ubigeo del Perú. Extraído de: <https://datosabiertos.gob.pe/dataset/ubigeos-c%C3%B3digos-de-ubicaci%C3%B3n-geogr%C3%A1fica-instituto-nacional-de-estad%C3%ADstica-e-inform%C3%A1tica-inei>

Informe (1PC)

1. Redacción académica y técnica

- Justificar cada punto con bibliografía breve (libros de Big Data, manuales de Hortonworks, artículos indexados sobre BI/Analytics).
- Ejemplo: “La ausencia de indicadores de rendimiento limita la toma de decisiones estratégicas (García et al., 2022)”.

2. Estandarización de tablas y matrices

- Para **problemas del negocio**: numerarlos y vincularlos con las necesidades de información.



- Para **KPI's**: usar un formato homogéneo (nombre, definición, fórmula, unidad, frecuencia, nivel de decisión).

3. Evidencia técnica detallada

- Capturas de pantalla con títulos explicativos debajo.
- Breve explicación de cada servicio: *HDFS (almacenamiento distribuido)*, *Hive (consulta SQL)*, *Spark (procesamiento en memoria)*.



RÚBRICA DE CALIFICACIÓN

Criterio / Valor	4	3	2	1	0	Peso (%)
Contexto de la empresa	Descripción completa y coherente de generalidades, misión, visión y procesos.	Descripción adecuada, con pequeñas omisiones.	Descripción incompleta o poco clara en varios aspectos.	Muy básica, sin alineación estratégica.	No se presenta .	15
Problemas y necesidades	Identificación detallada, bien estructurada y alineada al negocio.	Problemas y necesidades bien descritos, con algunas carencias.	Problemas poco claros o incompletos, sin buena justificación.	Muy básicos o sin relación con el negocio.	No se presenta .	20
KPI's iniciales	Definición completa: nombre, fórmula, frecuencia y unidad claramente establecidos.	KPIs definidos con pequeños vacíos en fichas técnicas.	KPIs incompletos o con inconsistencias.	Muy básicos, sin coherencia con la estrategia.	No se presenta .	20
Evidencia técnica Hortonworks	Instalación y servicios activos comprobados, capturas claras de evidencia + carga en GitHub.	Instalación funcional con detalles menores faltantes + carga en GitHub.	Evidencia incompleta o sin validación clara + carga en GitHub.	Muy básica, sin pruebas suficientes + carga en GitHub.	No se realizó la carga en GitHub.	25
Diagrama de arquitectura	Flujo de datos y componentes claramente representados y justificados.	Diagrama adecuado con pequeñas omisiones.	Diagrama incompleto o con errores de coherencia.	Muy básico, sin reflejar la arquitectura real.	No se presenta .	10
Presentación y redacción	Documento claro, ordenado, bien estructurado y sin faltas ortográficas, cargado en GitHub.	Documento entendible, con fallas menores de forma, cargado en GitHub.	Documento poco claro, con errores frecuentes, cargado en GitHub.	Documento deficiente, desordenado, cargado en GitHub.	Document o no cargado en GitHub.	10

