



PROFESOR: Dr. Ing. ARADIEL CASTANEDA, HILARIO
AYUDANTE DE CATEDRA: 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, aseguramiento de salud.</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.

- a. **Justificación:** Cada hospital maneja sistemas de información distintos o aislados, sin interoperabilidad ni un repositorio centralizado. Esto dificulta consolidar datos en tiempo real.
- b. **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.

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

3. Infraestructura y equipamiento médico deficiente.

- a. **Justificación:** Las sedes alrededor del país no se encuentran en condiciones óptimas para atender a la demanda presente.
- b. **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.

- a. **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.
- b. **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

- a. **Justificación:** Las campañas no se sustentan en análisis predictivo ni segmentación de la población, y muchas veces se planifican de forma reactiva.
- b. **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 este problema porque representa un punto crítico donde convergen la gestión de datos, la toma de decisiones estratégicas y el



impacto directo en la salud de la población. 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 de EsSalud relacionadas con campañas, atenciones médicas y estadísticas de salud. Esto permite sustentar con evidencia el análisis del problema, medir tendencias y plantear indicadores de mejora.

2. Impacto directo en la salud pública

Una campaña de salud inefectiva se traduce en menor cobertura de población, falta de prevención y aumento en los costos de atención por enfermedades que pudieron evitarse.

3. Relación con otros problema identificados

La falta de integración de datos, los reportes tardíos y la ausencia de indicadores estandarizados son factores que contribuyen a la ineficacia de las campañas de salud. Por ello, abordar este problema implica también atacar causas raíz que fortalecen al sistema en su conjunto.

4. Alta relevancia social y mediática

Las campañas de salud son la cara más visible de EsSalud frente a la ciudadanía. Su éxito o fracaso repercute directamente en la confianza de la población hacia el sistema de salud. Una mejora en la efectividad de las campañas impacta de manera 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 ocupación hospitalaria	Mide el nivel de utilización de camas en hospitales	$(\text{Camas ocupadas} \div \text{Camas disponibles}) \times 100$	%	Diario
Tiempo promedio de espera en consulta	Tiempo promedio desde la solicitud hasta la atención	$\Sigma (\text{hora atención} - \text{hora solicitud}) \div N \text{ pacientes}$	Horas	Diario
Cobertura de campañas preventivas	Porcentaje de población objetivo alcanzada en campañas	$(\text{Población atendida} \div \text{Población objetivo}) \times 100$	%	Mensual
Oportunidad de reportes epidemiológicos	Rapidez en la entrega de reportes desde los establecimientos	$(\text{Reportes entregados a tiempo} \div \text{Total de reportes}) \times$	%	Semanal



		100		
Incidencia de enfermedades priorizadas	Casos nuevos detectados de enfermedades críticas	$(\text{Nuevos casos} \div \text{Población total}) \times 1000$	Casos por 1000 hab.	Mensual
Tasa de errores en registros de campaña de salud	Proporción de inconsistencias en registros de campañas	$(\text{Errores detectados} \div \text{Registros revisados}) \times 100$	%	Mensual
Efectividad de alertas tempranas	Porcentaje de alertas tempranas correctas	$(\text{Alertas correctas} \div \text{Alertas totales}) \times 100$	%	Mensual
Disponibilidad de medicamentos críticos	Stock real frente a lo planificado	$(\text{Stock disponible} \div \text{Stock planificado}) \times 100$	%	Semanal
Ratio de personal por paciente	Relación entre personal médico y pacientes atendidos	$(\text{N médicos} \div \text{N pacientes atendidos})$	Ratio	Diario
Nivel de integración de datos de campañas	Grado de centralización de datos de campañas.	$(\text{Campañas con datos integrados} \div \text{Total de campañas}) \times 100$	%	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:



The screenshot displays the Ambari dashboard interface. At the top, there is a navigation bar with the Ambari logo, the name 'Ambari', and a 'Sandbox' environment selector. To the right of the environment selector, a red alert banner indicates '8 alerts'. Further right are tabs for 'Dashboard', 'Services', 'Hosts', 'Alerts', and 'Admin', followed by a user profile icon labeled 'raj_ops'. On the left side, a vertical sidebar lists various services: HDFS, YARN, MapReduce2, Tez, Hive, HBase, Pig, Sqoop, Oozie, ZooKeeper, Falcon, Storm, Flume, Ambari Infra, Atlas, Kafka, Knox, Ranger, Spark2, Zeppelin, Notebook, Druid, and Slider. The main content area is titled 'Metrics' and includes sub-tabs for 'Heatmaps' and 'Config History'. Below these tabs, there are two buttons: 'Metric Actions' and 'Last 1 hour'. The dashboard is organized into a grid of metric cards. The first row includes 'HDFS Disk Usage' (n/a), 'DataNodes Live' (n/a), 'HDFS Links' (NameNode, Secondary NameNode, 1 DataNodes), 'Memory Usage' (No Data Available), and 'Network Usage' (No Data Available). The second row includes 'CPU Usage' (No Data Available), 'Cluster Load' (No Data Available), 'NameNode Heap' (n/a), 'NameNode RPC' (n/a), and 'NameNode CPU WIO' (n/a). The third row includes 'NameNode Uptime' (n/a), 'HBase Master Heap' (n/a), 'HBase Links' (No Active Master, 1 RegionServers, n/a), 'HBase Ave Load' (n/a), and 'HBase Master Uptime' (n/a). The fourth row includes 'ResourceManager Heap' (n/a), 'ResourceManager Uptime' (n/a), 'YARN Memory' (n/a), 'NodeManagers Live' (n/a), and 'YARN Links' (ResourceManager, 1 NodeManagers). On the left side of the metric grid, there are red alert indicators: two '2' icons next to the first two rows and one '1' icon next to the third row.

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

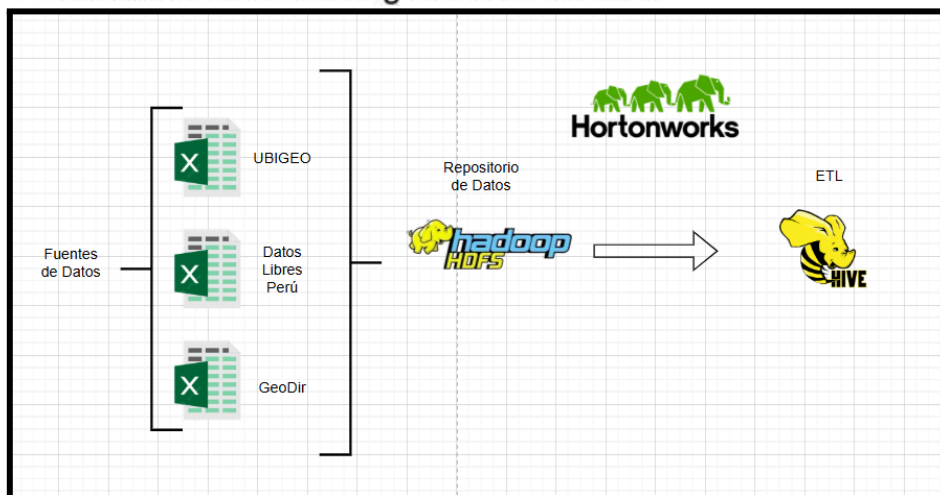
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://sandbox-hdp.hortonworks.com:4040
Spark context available as 'sc' (master = local[*], app id = local-1758581003522).
Spark session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

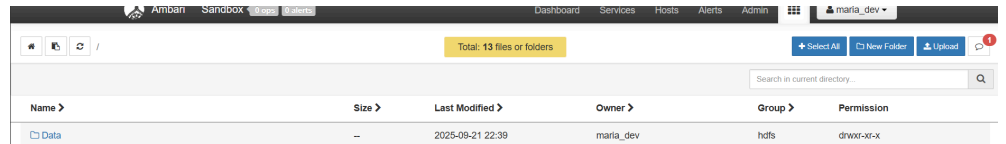
Repositorio	Archivo	Descripción
Centro Nacional de Planeamiento Estratégico	Datos-planeamiento-estrategico.xlsx	Incluye descripciones de cada ubigeo
Datos Abiertos Perú	DF_CExterna_2015_2022.csv	Contiene los datos de todas las consultas externas realizadas entre 2015 y 2022
	DF_ExLab_CExt_Diabetes.csv	Contiene todas las consultas externas con diagnósticos de diabetes y sus detalles
	DF_ExLab_CExt_EnfermedadRenal.csv	Contiene todas las consultas externas con diagnósticos de enfermedades renales y sus detalles
	DF_ExLab_CExt_Hiperlipidemia.csv	Contiene todas las consultas externas con diagnósticos de hiperlipidemia y sus detalles
	DF_ExLab_CExt_Hipertension.csv	Contiene todas las consultas externas con diagnósticos de hipertensión y sus detalles
	DF_ExLab_CExt_Obesidad.csv	Contiene todas las consultas externas con diagnósticos de obesidad y sus detalles
GeoDir	Geodir_Ubigeo_Inei.xlsx	Contiene detalle de todos los ubigeos con sus



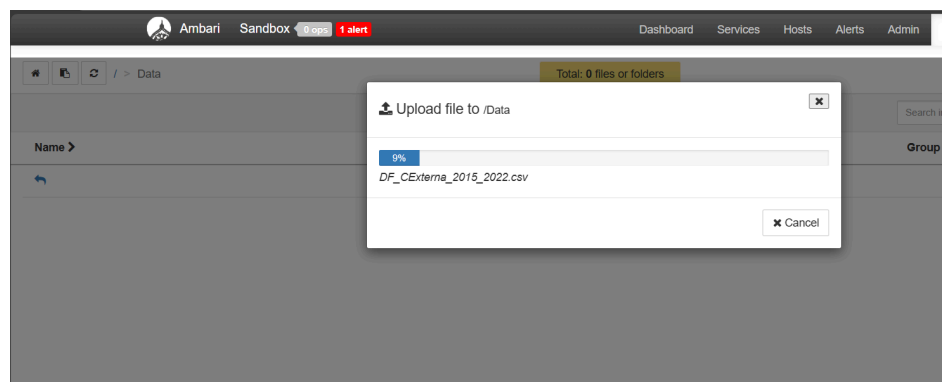
condiciones económicas y demográficas

- **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_CExterna_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_Hiperlipidemia.csv	5.9 MB	2025-09-21 22:35	maria_dev	hdfs	-rw-r--r--
DF_ExLab_CExt_Hipertension.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-inel.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.



HDFS Disk Usage



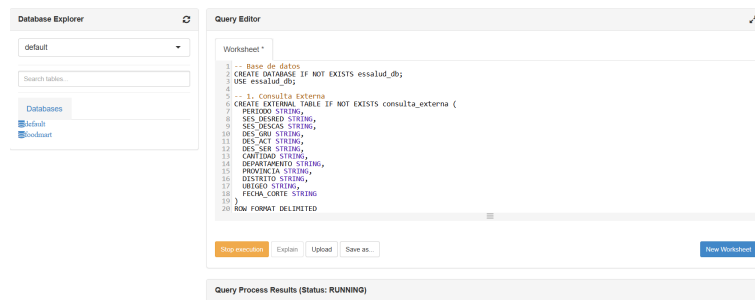
HDFS Disk Usage



DFS used
2.2 GB (2.05%)
non DFS used
20.0 GB (18.85%)
remaining
78.4 GB (73.96%)

- **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_DESCRED 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](#)

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

