UNIVERSIDAD PRIVADA ANTENOR ORREGO

FACULTAD DE INGENIERÍA

PROGRAMA DE ESTUDIO DE INGENIERÍA DE COMPUTACIÓN Y SISTEMAS

TESIS PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE COMPUTACIÓN Y SISTEMAS

SOLUCIÓN DE BIG DATA PARA EL ANALISIS DE LOS DATOS ABIERTOS DE MINSA Y CENARES PARA EL MONITOREO Y CONTROL DE LA EMERGENCIA SANITARIA COVID-19 BAJO EL ECOSISTEMA DE APACHE HADOOP Y MICROSOFT AZURE

Área de Investigación:
Gestión de Datos e Información

Autores:
Viteri Gonzales, Alan Percy

Beltrán García, José Antonio

Jurado Evaluador:

Presidente: Castillo Robles, Edward Fernando

Secretario: Meléndez Revilla, Karla Vanessa

Vocal: Abanto Cabrera, Heber Gerson

Asesor:
Ullón Ramírez, Agustin Eduardo

Código Orcid: https://orcid.org/0000-0003-1198-1855

TRUJILLO – PERÚ

2022

Fecha de sustentación: 2022/07/08

“SOLUCIÓN DE BIG DATA PARA EL ANALISIS DE LOS DATOS ABIERTOS DE MINSA Y CENARES PARA EL MONITOREO Y CONTROL DE LA EMERGENCIA SANITARIA COVID-19 BAJO EL ECOSISTEMA DE APACHE HADOOP Y MICROSOFT AZURE”

Elaborado por:

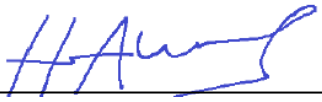
Br. Viteri Gonzales, Alan Percy

Br. Beltrán García, José Antonio

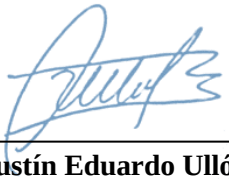
Aprobada por:

Ms. Castillo Robles, Edward Fernando
Presidente
CIP: 192352

Ms. Karla Vanessa Meléndez Revilla
Secretario
CIP: 120097



Ms. Abanto Cabrera, Heber Gerson
Vocal
CIP: 48234



Ms. Agustín Eduardo Ullón Ramírez
Asesor
CIP: 137602

PRESENTACIÓN

De acuerdo a los requisitos del reglamento de grados y Títulos de la Universidad, ponemos a vuestra disposición el presente Trabajo de Tesis: "**SOLUCIÓN DE BIG DATA PARA EL ANALISIS DE LOS DATOS ABIERTOS DE MINSA Y CENARES PARA EL MONITOREO Y CONTROL DE LA EMERGENCIA SANITARIA COVID-19 BAJO EL ECOSISTEMA DE APACHE HADOOP Y MICROSOFT AZURE"**

Los Autores.

DEDICATORIA

Esta dedicatoria va en primer lugar a Dios, la virgen de la puerta que ha incrementado mi fe, mis padres por el esfuerzo para poder cumplir mis objetivos, mi mayor empuje que son mis hijos y a todos los que en esta última etapa me apoyaron y me han motivado a luchar por mi crecimiento.

Viteri Gonzales, Alan Percy

Dedico mi tesis a Dios, a mis padres y hermanos; A Dios, porque ha estado conmigo en cada paso del camino, cuidándome y dándome fuerzas para continuar, mis padres y hermanos, que han velado por mi felicidad y educación toda mi vida. Sin ellos no hubiera logrado, es para mí una gran satisfacción poder dedicarles a ellos, que me lo he ganado con esfuerzo, esmero y trabajo. Por eso soy quien soy ahora. Los amo con mi vida

Beltrán García, José Antonio

AGRADECIMIENTO

A nuestros docentes que nos han aportado con sus conocimientos en nuestra formación profesional, por sus consejos y enseñanzas.

Se agradece también al Ms. Agustín Ullón Ramírez, por su asesoramiento en la presente Tesis.

Muchas Gracias.

Los autores.

RESUMEN

“SOLUCIÓN DE BIG DATA PARA EL ANALISIS DE LOS DATOS ABIERTOS DE MINSA Y CENARES PARA EL MONITOREO Y CONTROL DE LA EMERGENCIA SANITARIA COVID-19 BAJO EL ECOSISTEMA DE APACHE HADOOP Y MICROSOFT AZURE”

Por:

Br. Viteri Gonzales, Alan Percy

Br. Beltrán García, José Antonio

El Big Data ha jugado un papel importante en la respuesta al COVID-19. La primera alarma sobre este nuevo virus se dio el 31 de diciembre de 2019 gracias al rastreo con Big Data e Inteligencia Artificial - de la empresa BlueDot. Desde que el Covid-19 se propagó, en China se intentó decrementar o detectar el número de personas contagiadas a través de la recolección de datos de los contagiados, luego generaron un sin número de aplicaciones para informar a las personas sobre los casos y la gravedad. Las soluciones de Big Data y su uso correcto pueden ser una herramienta de gran utilidad, para la detección y así descender la curva de contagios frente al COVID-19.

El CENARES como un Organismo Desconcentrado del Ministerio de Salud, responsable de la gestión estratégica del abastecimiento de los recursos de salud, estableciendo prioridades de acuerdo a los requerimientos nacionales y desarrollando los mecanismos necesarios de intervención en salud, definidos en el plan del Ministerio de Salud - MINSA.

El problema con estas instituciones es que actualmente no se tiene identificado las variables que se deben de tener en cuenta para realizar una proyección más certera de la cantidad de vacunas e implementos que se deben de comprar y distribuir, según la región, provincia y/o Distrito.

Con el trabajo se pretende construir una solución basada en datos recolectados de diferentes fuentes (MINSA-CENARES), que permita analizar, comprender y monitorizar la información para optimizar la compra y distribución de vacunas e implementos contra el COVID-19 bajo el ecosistema de Apache Hadoop y Microsoft Azure.

Palabra Claves: Big Data, Hadoop, Dashboard

ABSTRACT

“BIG DATA SOLUTION FOR THE ANALYSIS OF THE OPEN DATA OF MINSA AND CENARES FOR THE MONITORING AND CONTROL OF THE COVID-19 SANITARY EMERGENCY UNDER THE APACHE HADOOP AND MICROSOFT AZURE ECOSYSTEM”

By:

Br. Viteri Gonzales, Alan Percy

Br. Beltrán García, José Antonio

Big Data has played an important role in the response to COVID-19. The first alarm about this new virus was given on December 31, 2019 thanks to the tracking with Big Data and Artificial Intelligence - from the company BlueDot. Since the Covid-19 spread, in China an attempt was made to decrease or detect the number of infected people through the collection of data from those infected, then they generated a number of applications to inform people about the cases and the severity . Big Data solutions and their correct use can be a very useful tool for detection and thus lower the contagion curve against COVID-19.

The National Center for the Supply of Strategic Health Resources (CENARES), as a Decentralized Organization of the Ministry of Health, is in charge of managing the supply of strategic health resources, prioritized through national requests, developing the necessary mechanisms for the care of the health interventions defined by the programs of the Ministry of Health - MINSA.

The problem with these institutions is that currently the variables that must be taken into account to make a more accurate projection of the amount of vaccines and implements that must be purchased and distributed, according to the region, province and / or have not been identified. District.

The work aims to build a solution based on data collected from different sources (MINSA-CENARES), which will allow to analyze, understand and monitor the information to optimize the purchase and distribution of vaccines and implements against COVID-19 under the Apache ecosystem. Hadoop and Microsoft Azure.

Keywords: Big Data, Hadoop, Dashboard

ÍNDICE DE CONTENIDO

INDICE DE FIGURAS

INDICE DE TABLAS

1. INTRODUCCION

1.1. Planteamiento del problema

Según la OMS “El Big Data ha jugado un papel importante en la respuesta de China al COVID-19. La primera alarma sobre este nuevo virus se dio el 31 de diciembre de 2019 gracias al rastreo con Big Data e Inteligencia Artificial - de la empresa BlueDot - que alertó sobre un caso de neumonía inusua que estaba ocurriendo en Wuhan, China”.

Desde que el Covid-19 se propagó, en China se intentó decrementar o detectar el número de personas contagiadas a través de la recolección de datos de los contagiados, entre estos datos sus nombres, movimientos entre lugares de la ciudad o país, además de síntomas, para finalmente procesarlos en un entorno de Big Data y así tener un soporte en tomar mejores decisiones.

Luego generaron un sin número de aplicaciones para informar a las personas sobre los casos y la gravedad. Las soluciones de Big Data y su uso correcto pueden ser una herramienta de gran utilidad, para la detección y así descender la curva de contagios frente al COVID-19.

La pandemia del coronavirus, tanto en otros países como en el Perú, las cifras siguen en aumento, por eso que el uso de herramientas como el big data se presenta como una solución viable en la recolección de datos y presentación de información para luchar contra el coronavirus.

El big data nos va permitir recolectar y analizar datos e información en gran cantidad o volumen, utilizando para ello ecosistemas como Hadoop que nos permitirán analizar grandes cúmulos de información soportado en la nube de Microsoft Azure. Sus características de big data como: el volumen, velocidad, variedad, veracidad y valor, nos permitirán plasmar una estratégica basándonos en el cumplimiento de los objetivos del proyecto.

El estudio del COVID-19 empleando Big Data puede valerse de la analítica retrospectiva y descriptiva avanzadas basadas en la inteligencia de negocios; ya que esta permite focalizar el estudio mediante indicadores y tendencias en el tiempo, lo que incluye predicciones a futuro.

La actual situación generada por la emergencia sanitaria del coronavirus, permite evidenciar que el uso del big data en el sector salud es esencial para acelerar la obtención de información y conocimiento.

La misión del Ministerio de Salud es proteger la dignidad de la persona, promover la salud, prevenir enfermedades y garantizar un servicio integral de salud para todos los ciudadanos del país; proponer e implementar lineamientos de política de salud en consulta con todos los sectores públicos y actores sociales. El ser humano está en el centro de nuestra misión, y nos comprometemos a respetar la vida y los derechos fundamentales de todos los peruanos, desde antes de su nacimiento, respetando el desarrollo natural de sus vidas y contribuyendo a la gran tarea nacional de desarrollar a todos. nuestros ciudadanos Los trabajadores de la salud son agentes de cambio, promoviendo continuamente el mayor bienestar de las personas. (MINSA, 2020)

Asimismo, representantes de la Presidencia del Consejo de Ministros, del Seguro Social de Salud (Essalud), de la Universidad Peruana Cayetano Heredia, y de la Universidad Nacional Mayor de San Marcos. De otro lado, se establece que la secretaría técnica de esta comisión estará a cargo del Ministerio de Salud y asume las labores de vocería

El problema con estas instituciones es que actualmente no se tiene identificado las variables que se deben de tener en cuenta para realizar una proyección más certera de la cantidad de vacunas que se deben de comprar y distribuir, según la región, provincia y/o Distrito.

Con el trabajo se pretende construir una solución basada en datos recolectados de diferentes fuentes (MINSA-CENARES), que permita analizar, comprender y monitorizar la información para optimizar la compra y distribución de vacunas contra el COVID-19.

1.2. Delimitación del problema

El siguiente proyecto se realizará basada en el análisis de datos abiertos recolectados de diferentes fuentes de MINSA y CENARES para el monitoreo y control de la emergencia sanitaria Covid-19 utilizando Big data en un ecosistema Apache Hadoop y MS Azure, para luego visualizar la información en PowerBI.

1.3. Características problemáticas

- ✓ Falta de una mejor visualización de información de zonas focalizadas con mayor número de contagios.
- ✓ Un inadecuado análisis de los resultados en función a la cantidad de personas fallecidas y contagiadas con COVID-19.
- ✓ Información estática de la incidencia y la prevalencia de casos COVID-19.
- ✓ Altos tiempos en ejecución de consultas en la información de las compras y distribución de equipo médico contra el COVID-19.
- ✓ Esta problemática se debe a que los sistemas con los que cuenta las instituciones de MINSA y CENARES no fueron diseñados ni desarrollados con el fin de brindar síntesis, análisis, consolidación, búsquedas de datos.

1.4. Definición del problema

Falta de un análisis de información a las personas y/o entidades encargadas de tomar decisiones en base al comportamiento de casos positivos, fallecidos, compras y distribución de equipo médico contra el COVID-19.

1.5.Formulación del problema

¿Cómo se puede mejorar el análisis de información para las personas y/o entidades encargadas en tomar decisiones basado en el comportamiento de casos positivos y fallecidos contra el COVID-19?

1.6.Formulación de la hipótesis

Una Solución de Big data permitirá mejorar análisis de la información de los datos abiertos de MINSA y CENARES para el monitoreo y control de la emergencia sanitaria COVID-19 bajo el ecosistema de Apache Hadoop y Microsoft Azure.

1.7.Objetivos del estudio

El **Objetivo general** es:

Implementar una solución de Big data para el análisis de los datos abiertos de MINSA y CENARES para el monitoreo y control de la emergencia sanitaria covid-19 bajo el ecosistema de apache Hadoop y Microsoft Azure.

Los **objetivos específicos** son los siguientes:

1. Analizar los requerimientos de información de acuerdo a las perspectivas y necesidades.
2. Utilizar técnicas y métodos para el análisis de datos basados en la metodología ICAV.
3. Utilizar de HDInsight para la creación de cluster y Data factory para el procesamiento de datos
4. Presentar los reportes desde Microsoft Azure a Power BI.

1.8. Justificación de la investigación

1.8.1. Importancia del trabajo

- ✓ El desarrollar big data permitirá reportes prediseñados brindando información en tiempo real, analizando y tomando decisiones en base al comportamiento de casos positivos y fallecidos de COVID-19.
- ✓ La presente investigación permitirá mejorar la distribución de vacunas con énfasis en las zonas focalizadas con mayor número de contagios.
- ✓ El uso de analítica de datos da ventaja sobre la toma de decisiones y mejora el análisis de los resultados en función a la cantidad de personas fallecidas y contagiadas con COVID-19.

- ✓ Con el Big data realizaremos un seguimiento y monitoreo de las compras y distribución de los implementos médicos por el estado de emergencia COVID-19.

1.8.2. Viabilidad de la investigación

- ✓ Es viable porque los investigadores tienen conocimiento del software y herramientas importantes para el desarrollo del proyecto, esto de acuerdo con el tamaño de información de la institución, así como por su nivel de manejo y aprendizaje de los usuarios de la solución.
- ✓ Es viable porque la información de la empresa se encuentra a la mano de los investigadores como son los datos abiertos de las instituciones gubernamentales.
- ✓ Es viable porque contamos con los conocimientos sobre el tema a implementar, así como con el soporte y experiencia del asesor del proyecto.

2. MARCO TEÓRICO

2.1. ANTECEDENTES

- ✓ **Autor:** Rojas García, José Antonio

  Título de Investigación: "Propuesta de un modelo de negocio basado en big data que facilite la integración de los datos de las personas naturales y de soporte a las políticas de e-government en el Perú, apoyado en una empresa de logística integral" UPC – Lima 2018

  Descripción:

  El objetivo general de este trabajo es "abordar las necesidades de los futuros consumidores de Light Logistics Company y cómo puede servir de soporte a varias políticas que actualmente se están implementando para el desarrollo del gobierno electrónico en el Perú". El resultado es "un beneficio adicional que puede complementar y apoyar algunas políticas públicas peruanas para actualizar los datos cada vez más escasos y así transformar la sociedad actual en una nueva digital".

- ✓ **Autor:** Milton Ivan Cañarte Manrique

  Título de Investigación: "Análisis del uso de big data en las empresas guayaquileñas sobre la base de plataformas basadas en tics en el año 2014". Universidad de Guayaquil

  Descripción:

  EL trabajo tiene como objetivo general "Determinar si existen estrategias de grandes datos en las empresas de la ciudad de Guayaquil en el 2014", permite Analizar el impacto que genera la recolección de datos de productos y de clientes en las empresas y Analizar el estado actual de las TICs para el manejo de grandes datos en las empresas". La investigación propone "la evaluación y análisis del estado actual de las estratégicas de grandes datos en empresas de la ciudad de Guayaquil, puesto que grandes resultados requieren una gran estrategia para la recogida, limpieza, correlación y el análisis de todos estos datos, con esto se logrará obtener una idea clara de la importancia que toma la información a gran

escala en la ciudad y si se está sacando provecho de los datos que se adquiere en cada una de las empresas".

- ✓ **Autores:** Rodriguez Torres, Eduardo y Pereda Morales, Piero Armando

  Título de Investigación: "Implementación de un Dashboard para la toma de decisiones estratégicas en la unidad de negocio de producción de huevo incubable de la empresa Avícola Santa Fe S.A.C. Usando Tecnologías Oracle Business Intelligence", Trujillo 2017

  Descripción:

  En el presente trabajo tiene como objetivo "la implementación de Dashboards (Reportes Estratégicos), que será usados en la Unidad de Negocio de Producción de Huevo Incubable de la Empresa Avícola Santa Fe S.A.C". Para lograr dicho objetivo, se usó la herramienta Oracle Business Intelligence. Para el desarrollo del trabajo "se utilizó la metodología de Ralph Kimball con la herramienta Business Intelligence de Oracle para implementar los Dashboards, que permitirán a las gerencias tener un espacio de trabajo adecuado donde puedan consultar los indicadores a través de estos".

- ✓ **Autora:** Sonia Guama Morales

  Título de Investigación: "Estudio comparativo de métodos existentes para integrar la información estructurada y no estructurada de una industria enfocado en la generación de conocimiento, desde la perspectiva de una solución integral de big data." 2018

  Descripción:

  La investigación del presente estudio se "desarrolló en base a la exploración de diversas fuentes entre los cuales se destacan textos de autores especializados en Big Data y consultas en la web de sitios certificados como fuente de apoyo". Esta investigación ha permitido desarrollar "una guía para la introducción de Big Data en una organización, independientemente de su vertical, para generar conocimiento que les permita innovar, renovar o mejorar la visión de negocio que desean alcanzar y reemplazar una cultura basada en los datos estructurado".

- ✓ **Autor:** Guillermo Magaña Bou

  Título de Investigación: “El big data y la convergencia con los insigths sociales”. México, D.F. 2019

  Descripción:

  El propósito de este trabajo es "comprender cómo las nuevas tendencias publicitarias pretenden lograr una percepción social en las guías de programas. Los objetivos de la investigación programática se limitan a la audiencia y la percepción, por lo que es fundamental comprender la ecología de los medios que se ofrecen en el mundo futuro". En última instancia, combina el poder de los grandes datos para generar información procesable, lo que ayuda a los canales de televisión y a los proveedores de servicios a llegar a segmentos publicitarios con sólidas capacidades de orientación o audiencias objetivo. Tienen tasas de transacción y participación de los usuarios muy altas, por lo que tienen un gran potencial para usar big data. El big data es un factor especialmente importante en este ámbito, ya que gran parte de la información que genera y gestiona es muy sensible”.

2.2.DEFINICIONES

2.2.1. BIG DATA

Big data “describe grandes cantidades de datos semiestructurados, estructurados y no estructurados que pueden extraerse y utilizarse en proyectos de aprendizaje automático y aplicaciones de análisis de datos avanzado.”. (Iebschool, 2019)

LAS 7V DEL BIG DATA: DATOS TRANSFORMADOS EN VALOR

- ✓ **Volumen**
- ✓ **Velocidad**
- ✓ **Variedad**
- ✓ **Variabilidad**
- ✓ **Veracidad**
- ✓ **Visualización**
- ✓ **Valor**

2.2.2. ANALÍTICA DE DATOS:

La gestión de datos tiene como objetivo último "dotar a las organizaciones de conocimiento y esto no es posible sin la Analítica de datos (Data Analytics). Significa traducir la información en oportunidades para el desarrollo de negocio y mejorar el rendimiento de la organización. En definitiva: se trata de sacar conclusiones de la información. En general de nada sirve tener datos, si luego no hacemos nada con ellos o más concreto: aprendemos de ellos, por lo que hoy tanto los datos, como el análisis, tenemos que comprender que van de la mano. La analítica de datos implica un proceso de limpieza y transformación cuyo objetivo es descubrir cuál es la información que nos ayudará a la mejor toma de decisiones y a extraer conclusiones que mejoren la competitividad de las compañías". (Prometeusgs, 2019)

Es muy importante conocer los diferentes tipos de analítica de datos y estos están divididos en 4 categorías:

a. **Descriptivos ¿Qué está pasando?**
b. **Diagnóstico ¿Por qué está pasando?**
c. **Predictiva ¿Qué es lo más probable que pueda pasar?**
d. **Prescriptivos ¿Qué necesito hacer?**

2.2.3. INTELIGENCIA DE NEGOCIOS:

"Inteligencia de negocios es un término tecnológico que engloba datos, informática y análisis dentro de operaciones de negocios. Es mucho más que algo específico; es un término general que incluye los procesos y métodos para recopilar, almacenar y analizar datos de actividades u operaciones de negocios para optimizar el rendimiento. Todo eso se combina para crear una vista integral de una empresa y ayudar a las personas a tomar decisiones que sean mejores y más útiles". (Tableau, 2020)

"La transformación digital creó una afluencia enorme de información, que no disminuye. Hay datos en todos lados, todo el tiempo. Y, ahora, están profundamente arraigados en los procesos de negocios de organizaciones de todos los tamaños. Todos esperan poder acceder a información nueva y usarla para fundamentar decisiones diarias y satisfacer su curiosidad de negocios sobre cuáles pueden ser los próximos pasos". (Tableau, 2020)

2.2.4. INTEGRACIÓN DE DATOS

"La integración de datos es el proceso que implica combinar datos desde distintas fuentes en una única visión unificada: empezando por la ingesta, la limpieza, el mapeo hasta la transformación en un colector determinado y, por último, convertir los datos en elementos más explotables y valiosos para aquellos que acceden a ellos. Actualmente las empresas llevan a cabo iniciativas de integración de datos para analizar y tomar decisiones a partir de sus datos de forma más eficaz, en especial dada la explosión de datos y de nuevas tecnologías cloud y de big data. La integración de datos es una obligación, puesto que permite a las empresas modernas mejorar la toma de decisiones estratégica y aumentar su ventaja competitiva". (Talend, 2019)

2.2.5. DASHBOARD

"Un dashboard o tablero de operaciones es una herramienta que sirve para visualizar y dar seguimiento a determinados indicadores de desempeño o estado. Condensa en un solo lugar la información crítica de una máquina, una empresa, una estrategia, etc". (Workana, 2020)

"Un dashboard es donde podemos encontrar los principales indicadores clave de desempeño de toda una empresa. Como seguramente ya puedes imaginarte, en la mayoría de los casos un dashboard es una herramienta principalmente de software que se visualiza por medio de una interfaz gráfica. Los ámbitos de la productividad y las ventas son los que más se prestan para la creación de dashboards empresariales. Por ejemplo, una compañía puede tener un dashboard que integre la información de manufactura en la fábrica, además de uno que arroje cifras sobre el estado de las ventas". (Workana, 2020)

2.2.6. APACHE HADOOP

"Apache Hadoop es una estructura para componentes de software diversos basada en Java, que permite fragmentar tareas de cálculo (jobs) en diferentes procesos y distribuirlos en los nodos de un clúster de ordenadores, de forma que puedan trabajar en paralelo. En las arquitecturas Hadoop más grandes pueden usarse incluso varios miles de ordenadores. La ventaja de este concepto es que a cada ordenador del clúster solo se le ha de proporcionar una fracción de los recursos de hardware necesarios. De esta manera, el trabajo con grandes volúmenes de datos no presupone ninguna máquina de última generación, sino que se puede llevar a cabo de forma más rentable con varios servidores estándar". (Ionos, 2019)

Componentes básicos de la arquitectura Hadoop:

El fundamento del ecosistema Hadoop lo constituye el Core Hadoop. Sus componentes en la primera versión son el módulo básico Hadoop Common, el Hadoop Distributed File System (HDFS) y un motorMapReduce. A partir de la versión 2.3 este último fue sustituido por la tecnología de gestión de clústers

YARN, también denominada MapReduce 2.0. Esta técnica excluye el algoritmo MapReduce del sistema de gestión en sí, de forma que a partir de este momento se convierte en un plugin basado en YARN. (Ionos, 2019)

2.2.7. MICROSOFT AZURE

Microsoft Azure "es conjunto en constante expansión de servicios en la nube para ayudar a las organizaciones a satisfacer sus necesidades comerciales. Otorga la libertad de crear, administrar e implementar aplicaciones en una red mundial con sus herramientas y marcos favoritos". (Microsoft, 2019)

Figura N° 01: Magic Quadrant for Analytics and BI Platforms (Gartner, 2019)

2.2.8. AZURE HDINSIGHT

Azure HDInsight "es un servicio de análisis de código abierto de servicio completo alojado para la empresa. HDInsight es un servicio en la nube que hace que sea fácil, rápido y rentable procesar grandes cantidades de datos. HDInsight también es compatible con una amplia gama de escenarios, como extracción, transformación y carga (ETL), almacenamiento de datos, aprendizaje automático e IoT." (Azure HDInsight, 2019)

HDInsight es una distribución Big Data en la nube de Apache Hadoop. Incluye implementaciones de Apache Spark, HBase, Kafka, Storm, Pig, Hive, Sqoop, Oozie o Ambari entre otros productos y servicios. (Bit, 2017)

Figura N° 2: Azure HDInsight (Bit, 2017)

Ofrece la creación de los siguientes tipos de clúster:

- ✓ Apache Hadoop
- ✓ Apache Spark
- ✓ Apache HBase
- ✓ Microsoft R Server
- ✓ Apache Storm
- ✓ Apache Interactive Hive (preview)
- ✓ Apache Kafka (preview)
- ✓ Clústeres Active Directory (preview)
- ✓ Clústeres personalizados con acciones de script con Hue, Giraph, R o Solr. (Bit, 2017)

Su arquitectura es la siguiente:

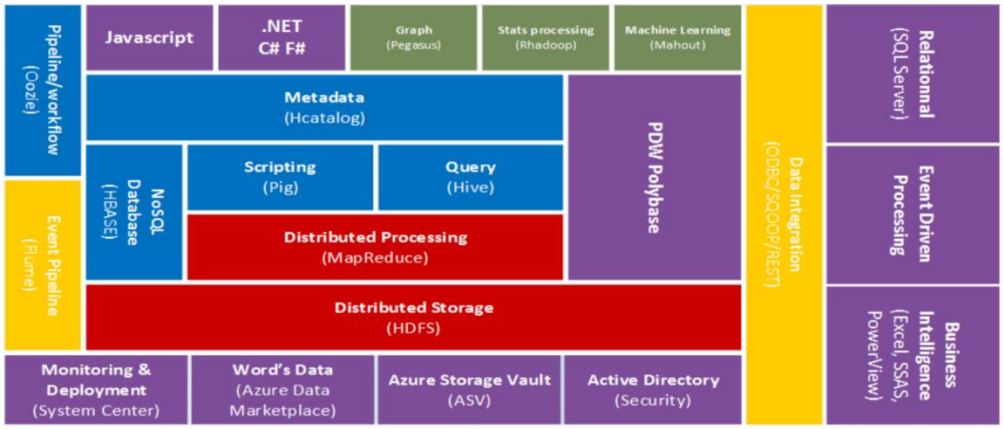


Figura N° 3: Arquitectura HDInsight (Bit, 2017)

2.2.9. POWER BI

Power BI "es un servicio de análisis empresarial diseñado para simplificar la forma en que los ISV y los desarrolladores usan las características de Power BI con análisis integrado. Power BI Embedded simplifica las capacidades de Power BI para ayudar a las aplicaciones a agregar rápidamente imágenes, informes y paneles impresionantes. Similar a las aplicaciones basadas en Microsoft Azure, utiliza servicios como aprendizaje automático e IoT. Al habilitar la exploración de datos de fácil navegación dentro de sus aplicaciones, los ISV permiten a los clientes tomar decisiones rápidas e informadas en contexto". (Microsoft, 2019)

2.3.METODOLOGIA: ICAV

La Metodología ICAV teniendo las siguientes fases:

- **Identificar**. "Al iniciar un proyecto de análisis de datos, los requisitos comerciales deben estar claramente definidos. ¿Quiénes son los usuarios finales que pueden saber qué preguntas necesitan respuestas para tomar mejores decisiones comerciales?". (Big Data SAC, 2019)

- **Consolidar**. “Una vez que se identifican las necesidades comerciales, se deben encontrar las fuentes de información necesarias para responder las preguntas comerciales”. (Big Data SAC, 2019)

- **Analizar**. “Una vez que tenga toda la información recopilada, puede comenzar a analizar e identificar tendencias que nos ayuden a predecir el escenario futuro”. (Big Data SAC, 2019)

- **Visualizar**. “Una vez que se completa el análisis, debe ser posible comunicar esta información de manera gráfica o tabular para la toma de decisiones comerciales.” (Big Data SAC, 2019).

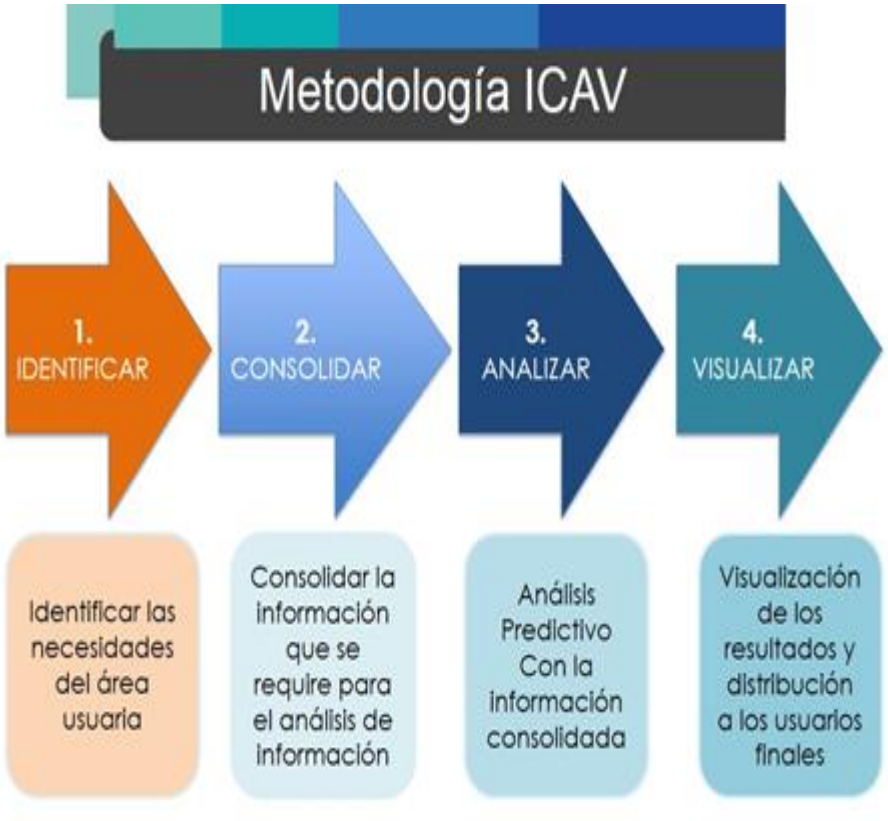


Figura 4: Metodología ICAV Fuente: (Big Data SAC, 2019)

3. MATERIALES Y METODOS

3.1. MATERIAL

3.1.1. Población

Datos Abiertos de MINSA y CENARES.

3.1.2. Muestra

Datos abiertos correspondiente a la emergencia sanitaria en el año 2020-2022 de MINSA y CENARES.

3.1.3. Unidad de análisis

Los datos proporcionados por el portal de transparencia de MINSA y CENARES.

3.2. MÉTODO

3.2.1. Tipo de investigación

Aplicada.

3.2.2. Diseño de Investigación

Diseño Pre-experimental con pre-prueba y post-prueba

Diseño del modelo pre-experimental	**G -> O_1 -> X -> O_2**
G (Grupo a investigar)	Datos Abiertos
X (Tratamiento)	Big Data

Diseño del modelo pre-experimental	**G -> O_1 -> X -> O_2**
O (Observación)	O_1 : pre-test
	O_2: post-test

Tabla 1. Diagrama de investigación

3.2.3. Variables de estudio y Operacionalización

- ✓ Independiente (VI): Solución de Big data
- ✓ Dependiente (VD): Análisis de la información de los datos abiertos de MINSA y CENARES para el monitoreo y control de la emergencia sanitaria COVID-19 bajo el ecosistema de Apache Hadoop y Microsoft Azure.

Tabla 2: Operacionalización de las variables

Variable	Dimensión	Indicador	Unidad de medida	Instrumento de Investigación
VI	Tiempo	Tiempo en obtener registros desde la solución de big data	Minutos	Hoja de captura de tiempos
	Grado de satisfacción de los usuarios	Grado de satisfacción de los sobre los reportes de la solución	% grado satisfacción	Hoja resumen de porcentajes de satisfacción
VD	Oportunidad	Tiempo para analizar información	Minutos	Hoja de captura de datos

3.2.4. Técnicas e instrumentos de recolección de datos

3.2.4.1.Técnicas

- ✓ Observación
- ✓ Análisis Documental

3.2.4.2.Instrumentos

- ✓ Cuestionario
- ✓ Hoja de cálculo.

3.2.5. Técnicas de procesamiento y análisis de datos

3.2.5.1.Procesamiento de datos

A través de tablas y gráficos dinámicos.

3.2.5.2.Análisis de datos

El análisis se desarrolla en cuadros estadísticos y Pruebas de hipótesis nula y alternativa.

4. RESULTADOS: APLICACIÓN DE LA METODOLOGÍA

4.1.PLANIFICACIÓN

La solución se basa en las fases de la metodología:

- ✓ Identificar. usuarios finales y cuáles son sus requerimientos.
- ✓ Consolidar. Identificar que fuentes de información, un repositorio para las agregaciones y transformaciones.
- ✓ Analizar. Utilizar técnicas avanzadas de análisis.
- ✓ Visualizar: Uso de dashboard para visualizar la información

4.2.IDENTIFICAR:

4.2.1. EVALUACIÓN DEL CASO DEL NEGOCIO

El Ministerio de Salud (MINSA) conduce el "Sistema Nacional Coordinado y Descentralizado de Salud basado en Redes Integradas de Salud".

El MINSA se esfuerza por brindar acceso universal a la atención y atención integral en salud individual y grupal de las personas sin importar su condición socioeconómica y ubicación geográfica. Garantizar una atención y una salud pública integrales, solidarias, equitativas, oportunas, gratuitas al nacer, de calidad, accesibles y pertinentes al ciclo de vida de la población, con enfoque de género, derechos a la salud y transculturalidad.

- ✓ **Visión:**

  La salud de todas las personas del país será expresión de un sustantivo desarrollo socio económico del fortalecimiento de la democracia, de los derechos y responsabilidades ciudadanas basadas en la ampliación de fuentes de trabajo estable y formal, con mejoramiento de los ingresos, en la educación en valores orientados hacia la persona y en una cultura de

solidaridad, así como en el establecimiento de mecanismos equitativos de accesibilidad a los servicios de salud mediante un sistema nacional coordinado y descentralizado de salud, y desarrollando una política nacional de salud que recoja e integre los aportes de la medicina tradicional y de las diversas manifestaciones culturales de nuestra población.

- ✓ **Misión:**

  El Ministerio de Salud tiene la misión de proteger la dignidad personal, promoviendo la salud, previniendo las enfermedades y garantizando la atención integral de salud de todos los habitantes del país; proponiendo y conduciendo los lineamientos de políticas sanitarias en concertación con todos los sectores públicos y los actores sociales. La persona es el centro de nuestra misión, a la cual nos dedicamos con respeto a la vida y a los derechos fundamentales de todos los peruanos, desde antes de su nacimiento y respetando el curso natural de su vida, contribuyendo a la gran tarea nacional de lograr el desarrollo de todos nuestros ciudadanos. Los trabajadores del Sector Salud somos agentes de cambio en constante superación para lograr el máximo bienestar de las personas.

- ✓ **Organigrama:**

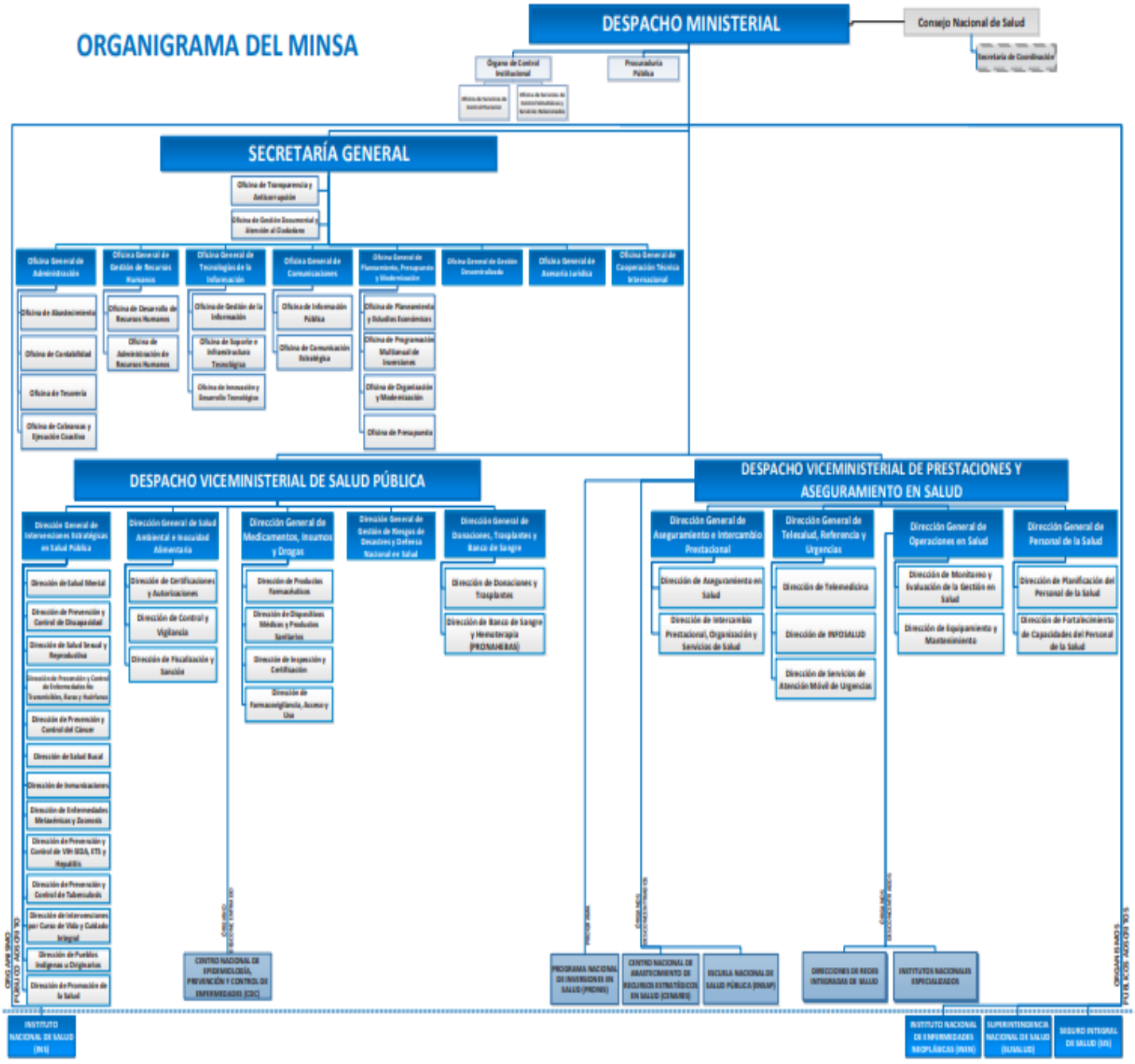


Figura 5: Organigrama de MINSA

- ✓ **<u>Proceso a Desarrollar</u>:**
  PRESTACIONES Y ASEGURAMIENTO EN SALUD.

4.2.2. OBJETIVOS DEL NEGOCIO

- Capacidad resolutiva en equipamiento
- Identificar los principales daños regionales y locales.
- Desarrollar el sistema de inteligencia sanitaria.
- Fortalecer el sistema de vigilancia e información
- Provisión eficiente y sostenible de los medicamentos en los servicios de salud.

4.2.3. EVALUACIÓN DE LA OFICINA GENERAL DE TECNOLOGÍAS DE LA INFORMACIÓN DEL SECTOR SALUD

La Oficina General de Tecnologías de la Información es el órgano de apoyo del Ministerio de Salud, dependiente de la Secretaría General, responsable de implementar el gobierno electrónico; planificar, implementar y gestionar los sistemas de información del Ministerio de salud; administrar la información estadística y científica en salud del Sector Salud; realizar la innovación y el desarrollo tecnológico, así como del soporte de los equipos informáticos del Ministerio de Salud.

También, es responsable de establecer soluciones tecnológicas, sus especificaciones, estándares; diseñar, desarrollar y mejorar las plataformas informáticas en el Sector Salud. Asimismo, establece requerimientos técnicos para la adquisición, aplicación, mantenimiento y uso de soluciones tecnológicas, en el ámbito de competencias del ministerio.

Actualmente está apostando por soluciones basadas en datos e información.

4.2.3.1. Determinación de requerimientos de información

En base a la ubicación geográfica se tiene en cuenta las necesidades de información de los ciudadanos peruanos y de algunas instituciones (MINSA - CENARES) interesadas en el desarrollo del COVID 19 en el Perú. Después de analizar estos requisitos, se determinó que nuestra solución tendría las siguientes características y consideraciones.

1. Cantidad de confirmados y fallecidos
2. Cantidad de pruebas por tipo
3. Cantidad de confirmados y fallecidos mensualmente, semanalmente y diariamente
4. Detalle por departamento, provincia y distrito
5. Detalle de confirmados y fallecidos por sexo
6. Detalle de confirmados y fallecido por momento de vida
7. Indice de letalidad
8. Cantidad de personas vacunadas
9. Detalle de camas UCI en uso
10. Número de personas hospitalizadas

4.2.3.2. Modelo conceptual

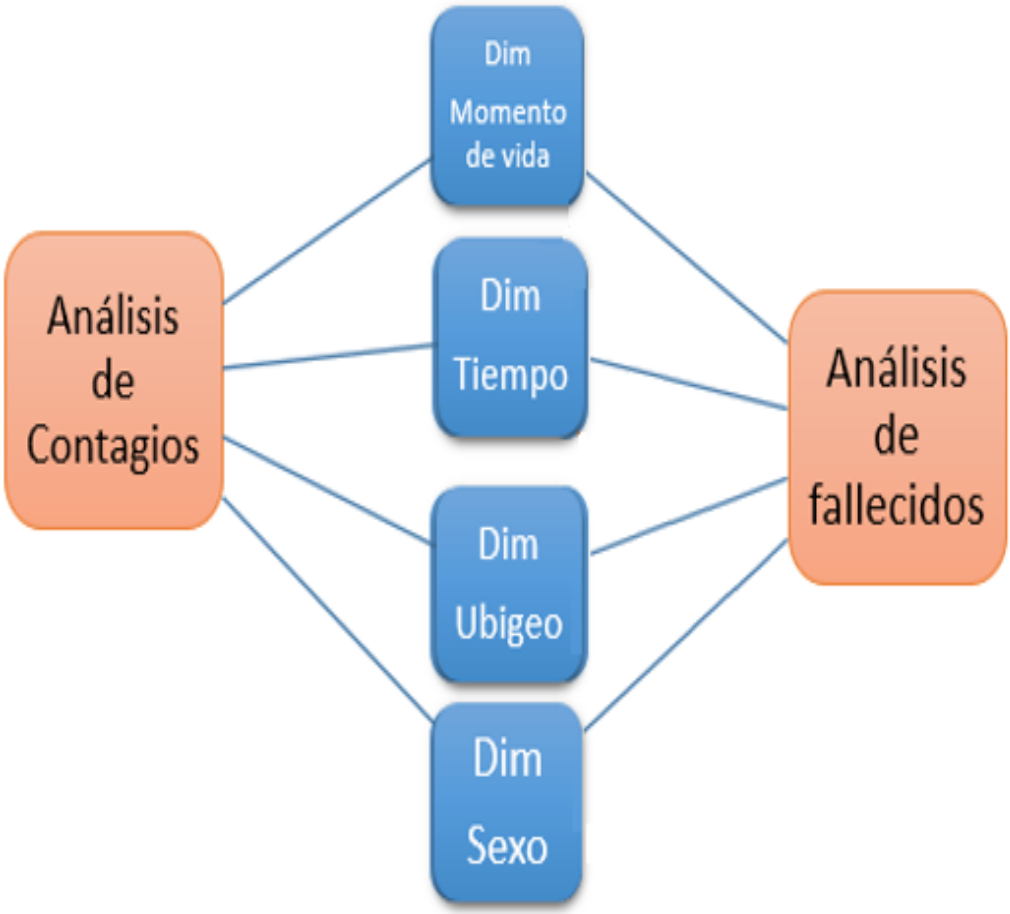


Figura 6: Modelo conceptual

4.3. CONSOLIDAR :

4.3.1. ANÁLISIS DE LOS DATOS

4.3.1.1. Origen de datos:

- ✓ Los datos fueron obtenidos en el Portal de OPEN DATA de la Plataforma Nacional de Datos Abiertos del Gobierno Peruano.

Figura 7: Plataforma Nacional de Datos Abiertos del Gobierno Peruano

Fuente: (Minsa, 2022)

4.3.1.2.Estructura de los Archivos a utilizar

- **Archivo: covid19_vaccine_arrivals_peru.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
cantidad	Cantidad de vacunas	300000
fecha_de_llegada	Fecha en que llegaron las vacunas	2021-02-07

vacuna	Tipo de vacuna	BBIBP-CorV
farmaceutica	Farmacéutica	Sinopharm
covax	Covax Facility	False
last_update	Fecha en que se actualizo la información de las vacunas	2022-03-28

- **Archivo: covid-19-peru-camas-uci.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
fecha	Fecha de la información	2020-04-27
estado	Estado de las camas UCI	en uso
essalud	Número de camas ocupadas en essalud	234
privado	Número de camas ocupadas en el sector privado	153
Minsa	Número de camas ocupadas en Minsa	137
gob_regional	Número de camas ocupadas en el gobierno regional	43
ffaa_pnp	Número de camas ocupadas en las fuerzas armadas y policiales	31
total	Total de camas UCI ocupadas	598

- **Archivo: covid-19-peru-data.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
Country	País	Perú
iso3c	La norma ISO 3166 para código de países	PER
region	Región	NA
date	Fecha de la información	2020-03-06
confirmed	Casos confirmados	1
deaths	Muertes confirmadas	NA

recovered	Pacientes recuperados	NA
total_pcr	Pruebas pcr	NA
total_serological	Pruebas serológicas	NA
total_ag	Pruebas antígenas	NA
total_tests	Total de pruebas realizadas	NA
negative_tests	Casos negativos	154
pcr_test_positive	Casos positivos por pcr	NA
serological_test_positive	Casos positivos por prueba serológica	NA
ag_test_positive	Casos positivos por prueba antígenas	NA
pcr_serological_test_positive	Casos positivos por prueba serológicas	NA

- **Archivo: covid-19-peru-detalle-hospitalizados.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
fecha	Fecha de la información	2021-10-02
hospitalizados	Número de pacientes hospitalizados	3906
sospechosos	Sopechosos de covid	NA
confirmados	Confirmados de covid	NA
essalud	Confirmados en essalud	2755
minsa	Confirmados en Minsa	977
clinica_privada	Confirmados en clinicas	153
ffaa_pnp	Confirmados en las FFAA y PNP	21
uci	Pacientes en UCI	978
ventilacion_mecanica	Pacientes con ventilación mecánica	898
favorable	Pacientes con pronóstico favorable	1015

estacionario	Pacientes estacionario	2422
desfavorable	Pacientes con pronóstico desfavorable	469
altas_medicas	N° de pacientes con altas medicas	87242
comunicado_minsa	Comunicado de minsa	703

- **Archivo: covid-19-peru-fallecimientos.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
fecha_anuncio	Fecha de la información	2020-04-01
fecha_fallecimiento	Fecha del fallecimiento del paciente	2020-03-30
fecha_ingreso	Fecha de ingreso	NA
edad	Edad del paciente	59
sexo	Sexo del paciente	femenino
región	Región del paciente	Lambayeque
viaje	Viaje	NA
fecha_retorno	Fecha de retorno a casa	NA
contacto	Contacto del paciente	NA
contacto_origen	Origen del contacto	NA
fecha_contacto	Fecha del contacto	NA
lugar_fallecimiento	Lugar de Fallecimiento	hospital
insuf_resp	N° de pacientes con insuficienca	NA
neumonia	N° de pacientes con neumonia	1
otros_sintomas	Otros sintomas	NA
factores	Factores	fibrosis pulmonar
misc	Miscelania	NA
comunicado_minsa	Numero de comunicado de Minsa	47

- **Archivo: covid-19-peru-test-results.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
fecha	Fecha de la información	2020-04-08
personas	Número de personas	3614
resultado	Resultado de prueba	positivo
tipo_prueba	Tipo de prueba	moleculares

- **Archivo: vacunas_covid_totales_por_semana.csv**

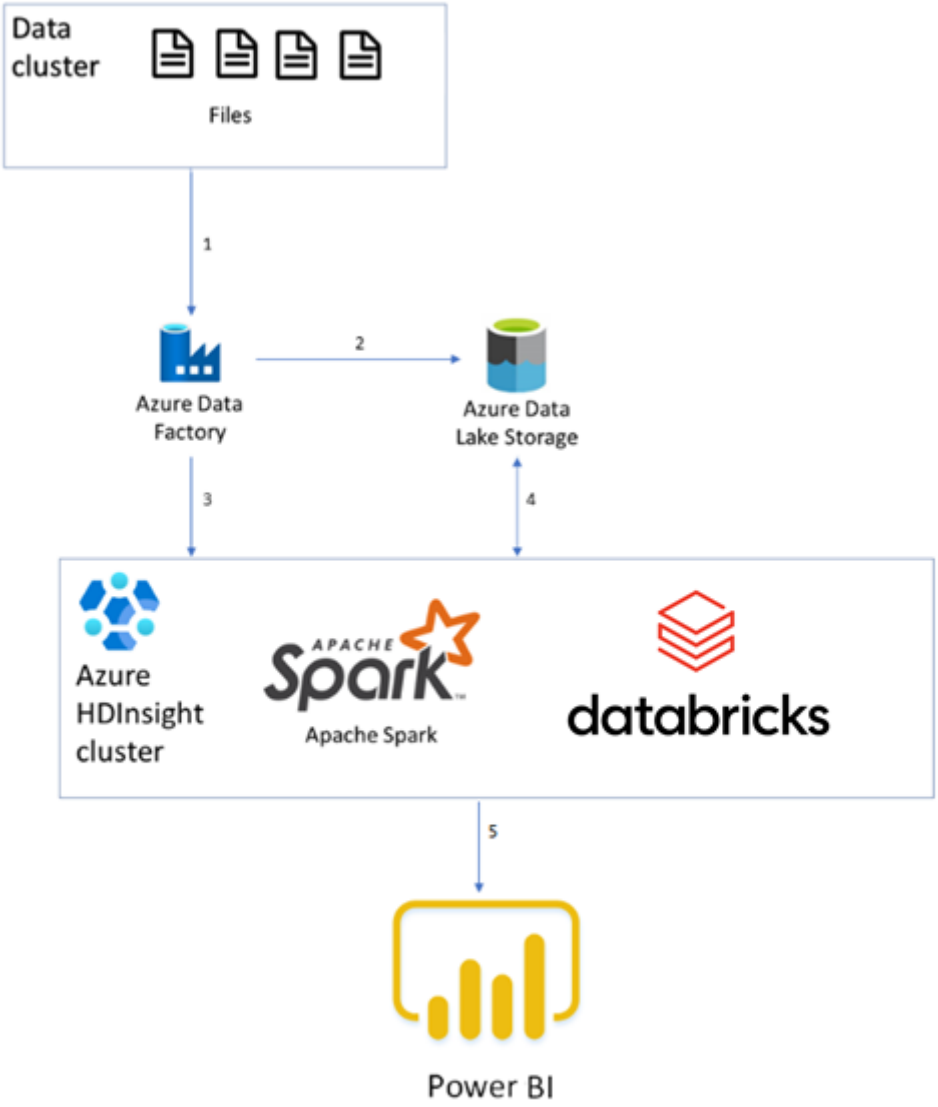
CAMPO	DESCRIPCIÓN	EJEMPLO
location	Ubicación	Perú
epi_year	Año	2021
epi_week	Semana epidemiológica	6
last_day_of_epi_week	Fecha	2021-02-13
complete_epi_week	Semana epi completa	1
vaccine_dose	Dosis de vacuna	1
vaccinations_epi_week	Vacunaciones por semana epidemiológica	105847
total_vaccinations	Total de vacunas	105847
pct_total_population	Porcentaje de población vacunada	0.30404798584834103

- **Archivo: vacunas_covid_resumen.csv**

CAMPO	DESCRIPCIÓN	EJEMPLO
fecha_corte	Fecha de corte de la información	2022-11-12
fecha_vacunacion	Fecha de vacunación	2020-04-27
fabricante	Fabricante de vacunas	SINOPHARM
dosis	Numero de dosis	1
flag_vacunacion_general	Vacunación general	FALSE

n_reg	Número de registro	1

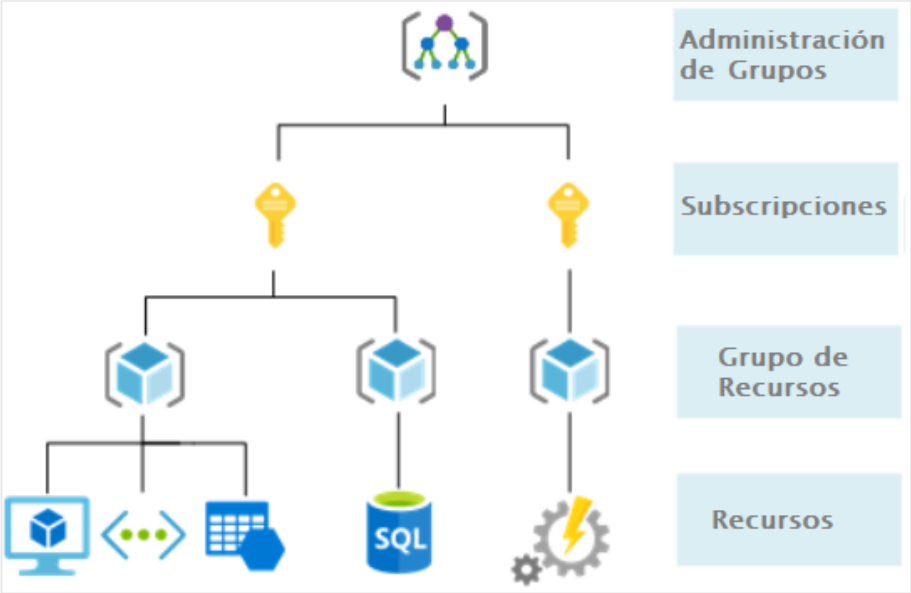
4.3.2. ARQUITECTURA EN MS AZURE A UTILIZAR EN EL TRABAJO



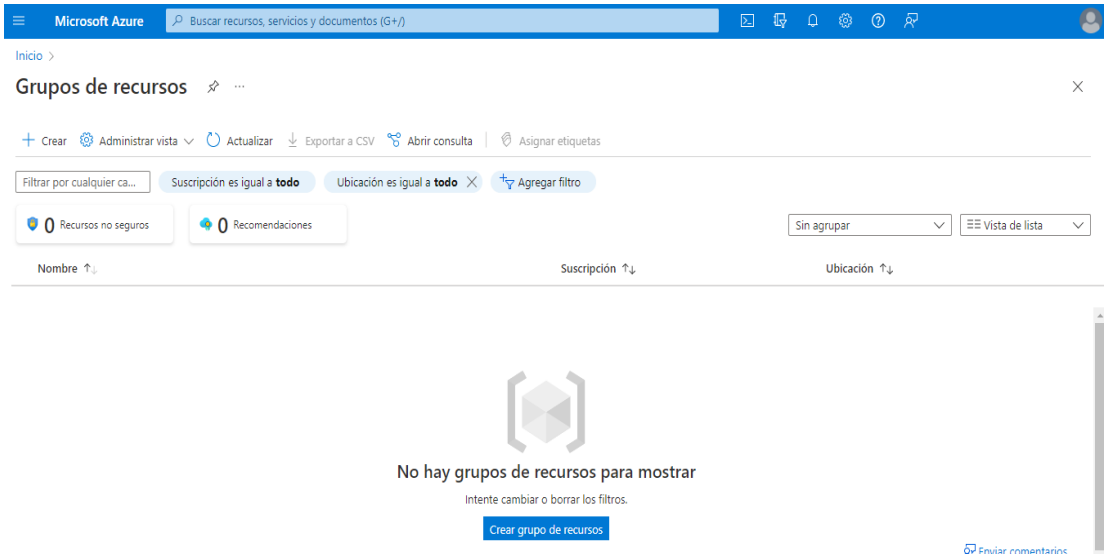
4.3.3. CONFIGURANDO LOS COMPONENTES DE LA ARQUITECTURA DE AZURE:

a) Creación y configuración del Grupo de Recursos en Azure

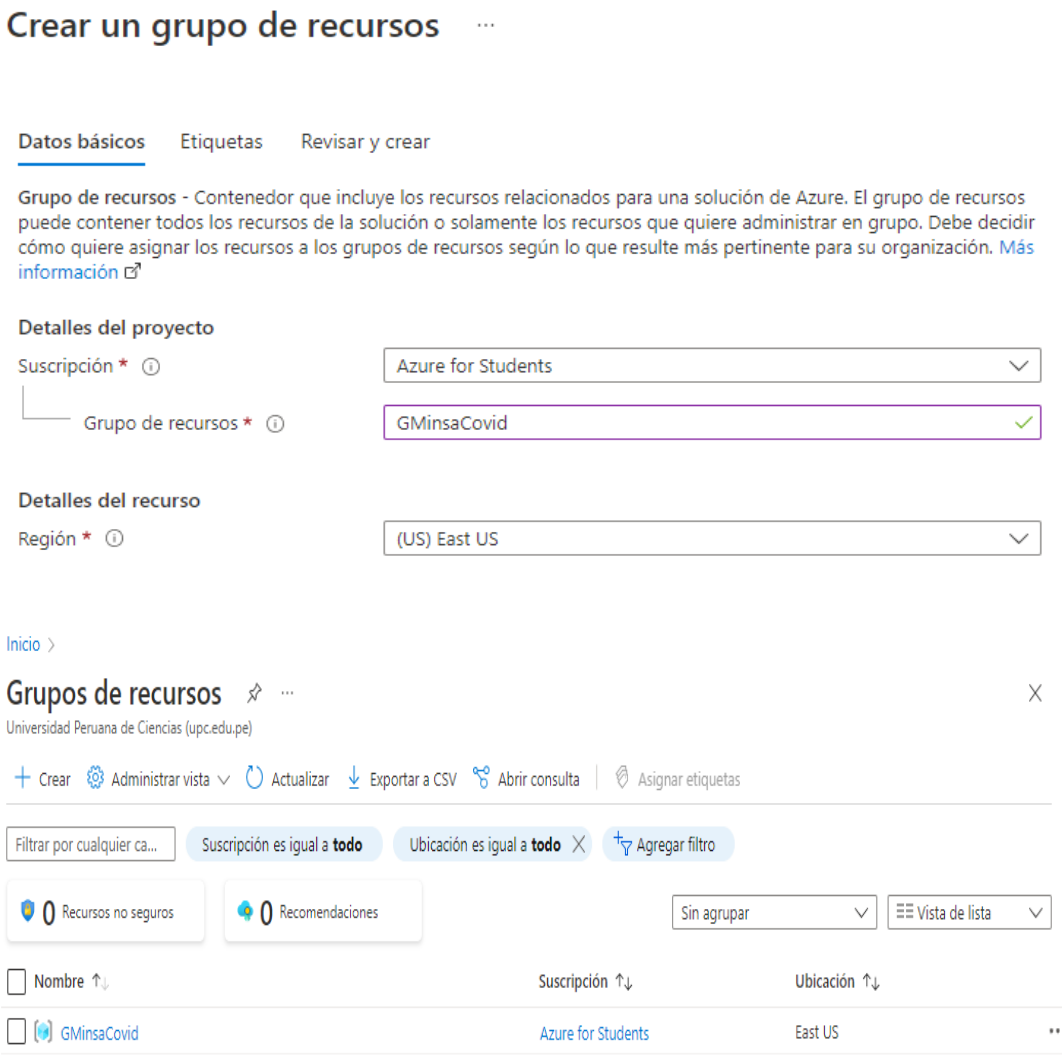
Azure proporciona cuatro niveles de administración: grupo de administración, suscripciones, grupos de recursos y recursos. En el diagrama siguiente se muestra la relación entre estos niveles.



Para crear el grupo de recursos tenemos a Azure Portal como interfaz basada en web diseñada para administrar recursos de Azure.



Creamos y configuramos un Grupo de Recursos para nuestra solución "GMinsaCovid"



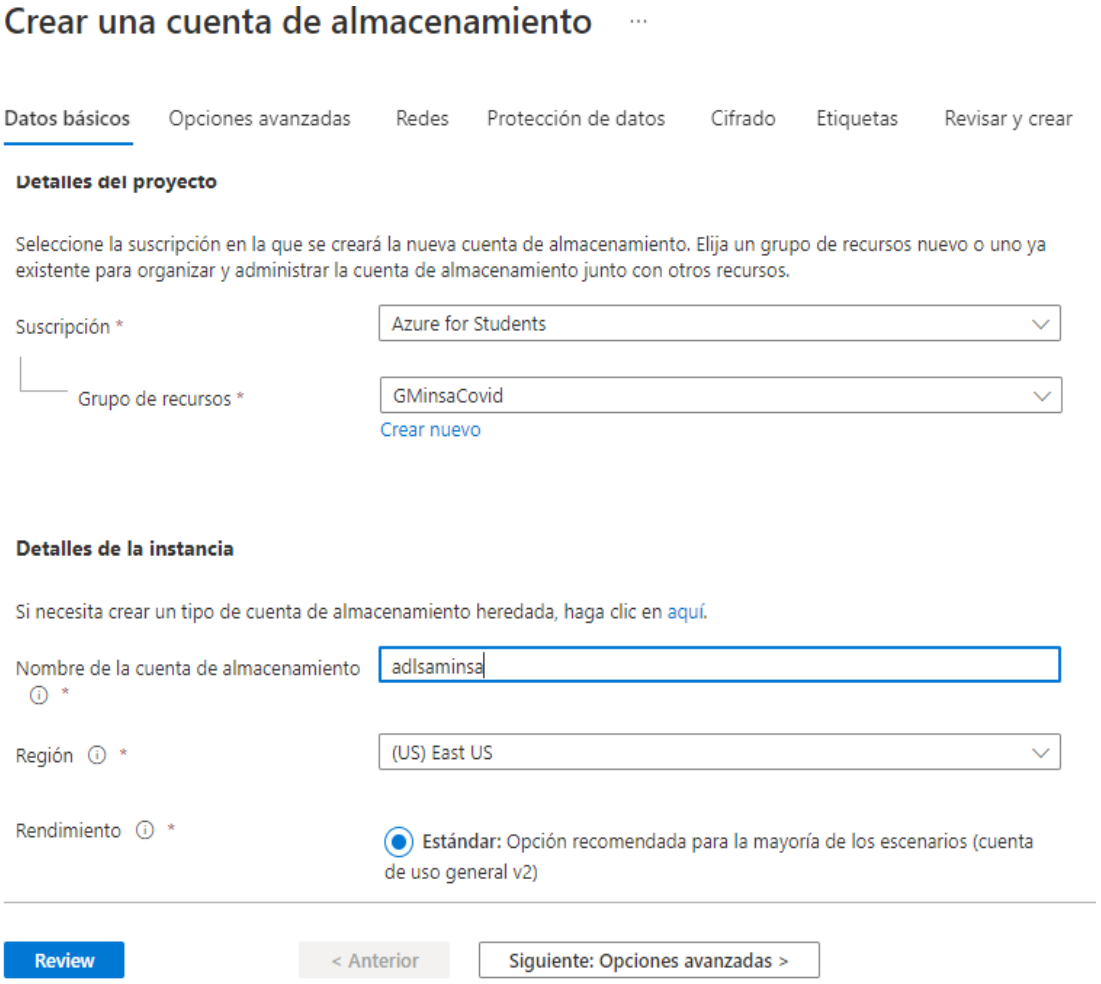
Finalmente, ya tenemos preparado el Grupo de recursos para los servicios a implementar

b) Creación y configuración del Data Lake Storage en Azure

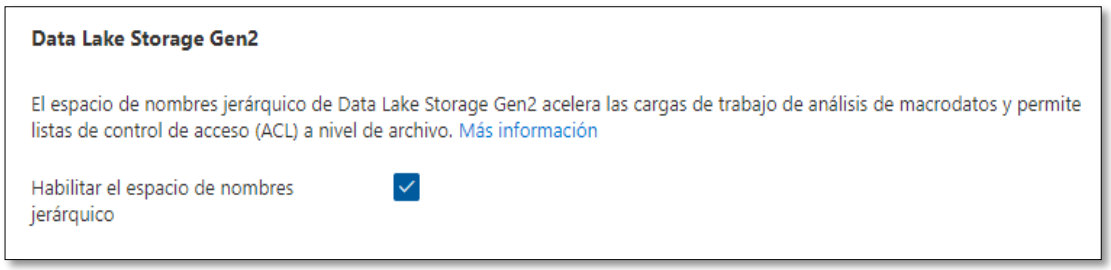
El data lake a crear nos va servir como un repositorio de almacenamiento para una gran cantidad de datos en bruto y que luego lo utilizaremos en la configuración de HDInsight.

La creación del Data Lake pasa por los siguientes pasos:

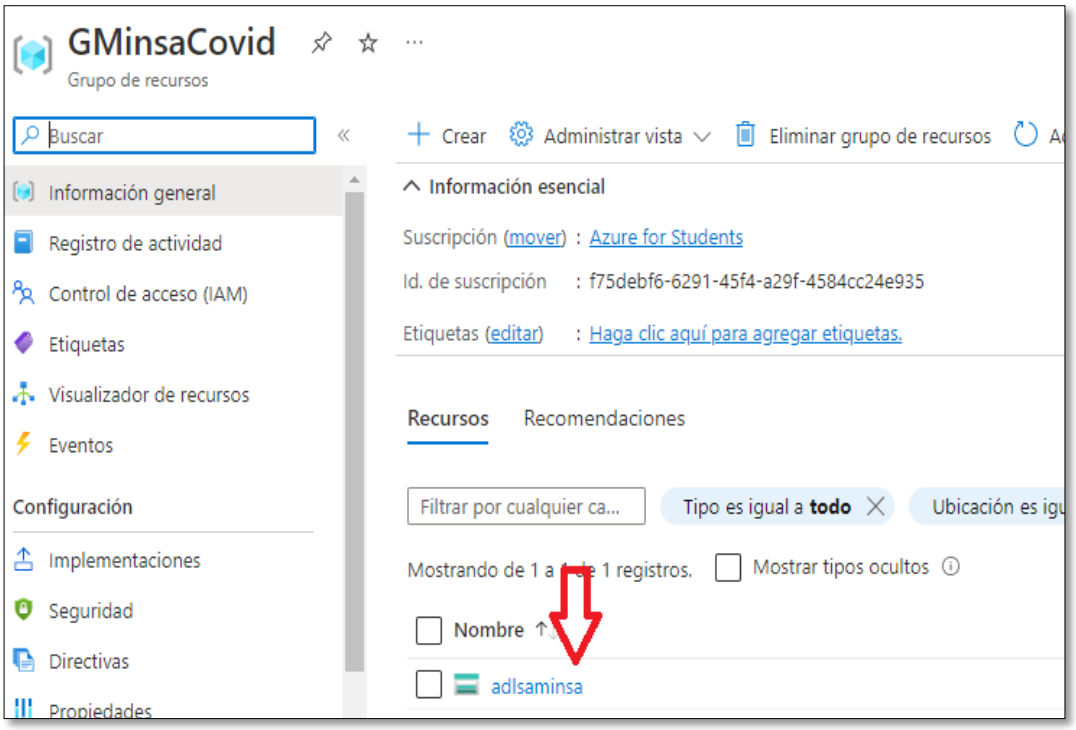
Primero creamos una “cuenta de almacenamiento”: adlsaminsa



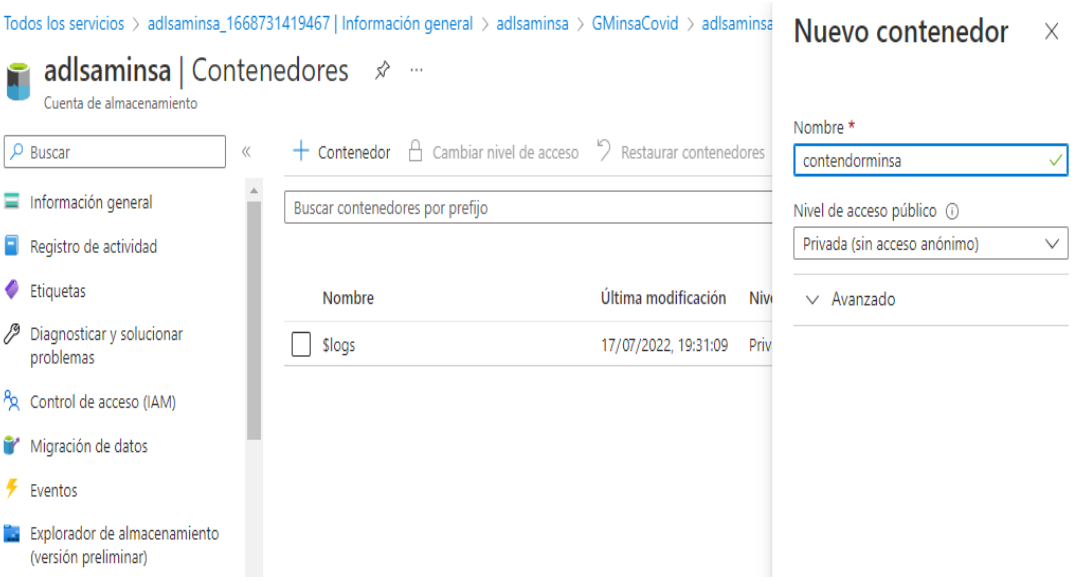
Luego vamos a opciones avanzadas para Habilitar el espacio de nombres jerárquico de esta manera se acelera las cargas de trabajo de análisis de macrodatos y permite listas de control de acceso (ACL) a nivel de archivo.



Luego para finalizar damos clic en Revisar y Crear.

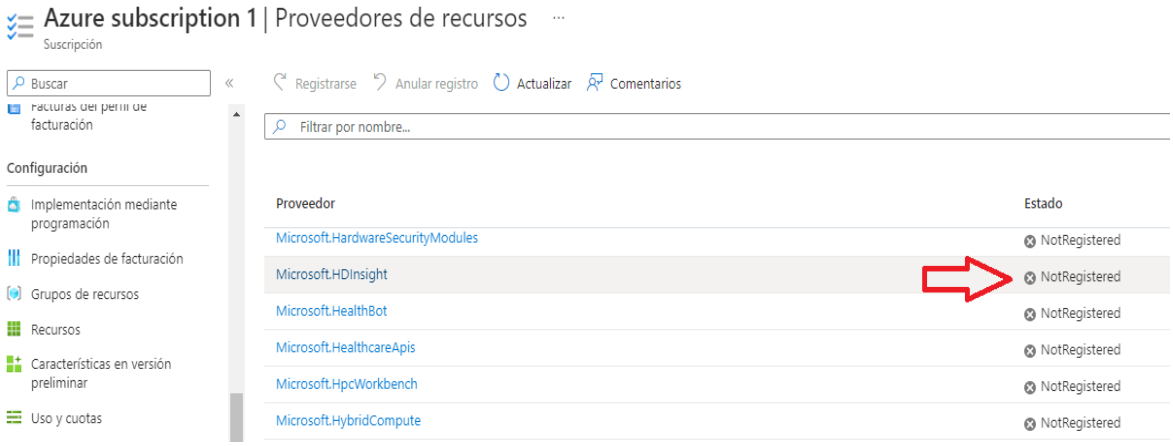


Luego de crear el Data Lake vamos a configurar un contenedor para alojar los datos

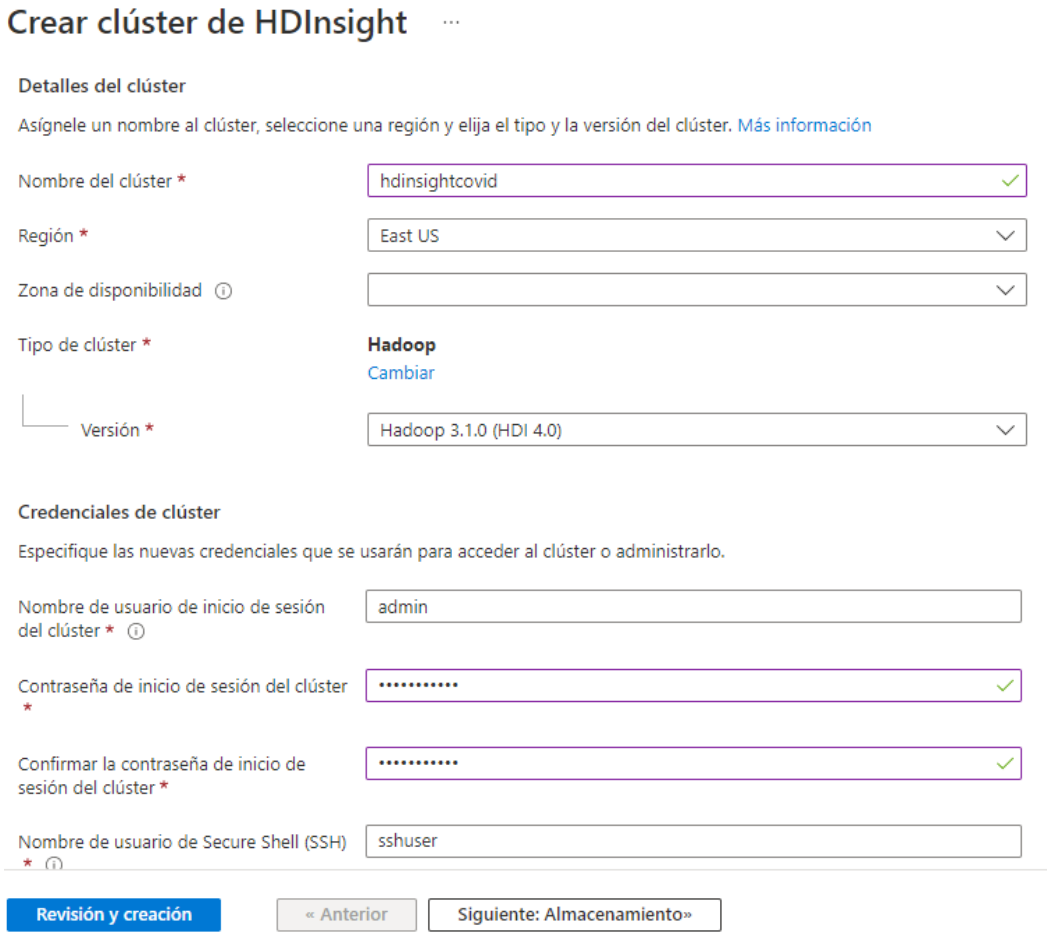


c) Creación y configuración de HDInsight (Hadoop)

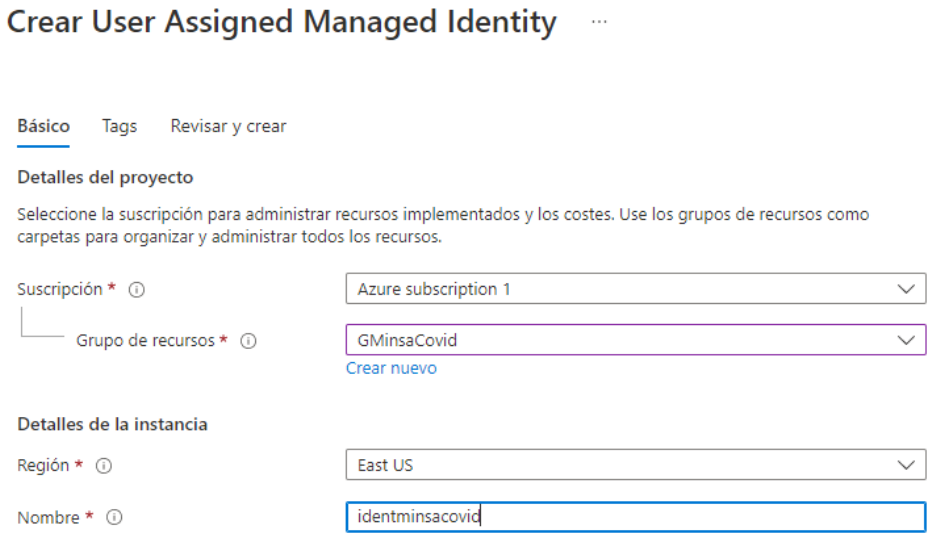
Antes de Utilizar el recurso de HDInsight se debe de registrar la suscripción al tipo de recurso a utilizar



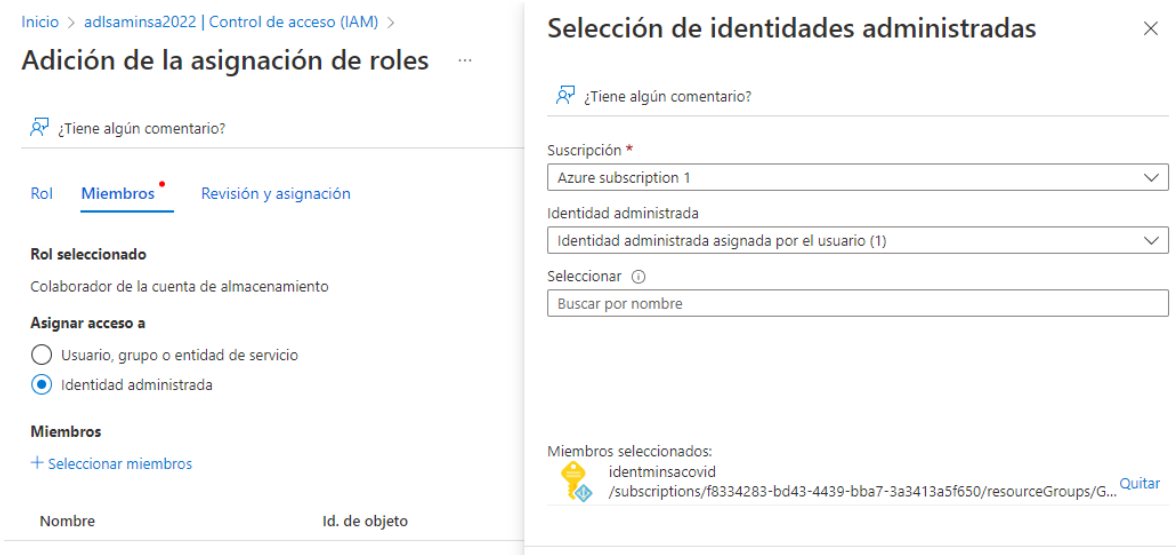
Luego de registrarlo podremos realizar la creación del recurso seleccionando el tipo de cluster, para nuestro caso es Hadoop



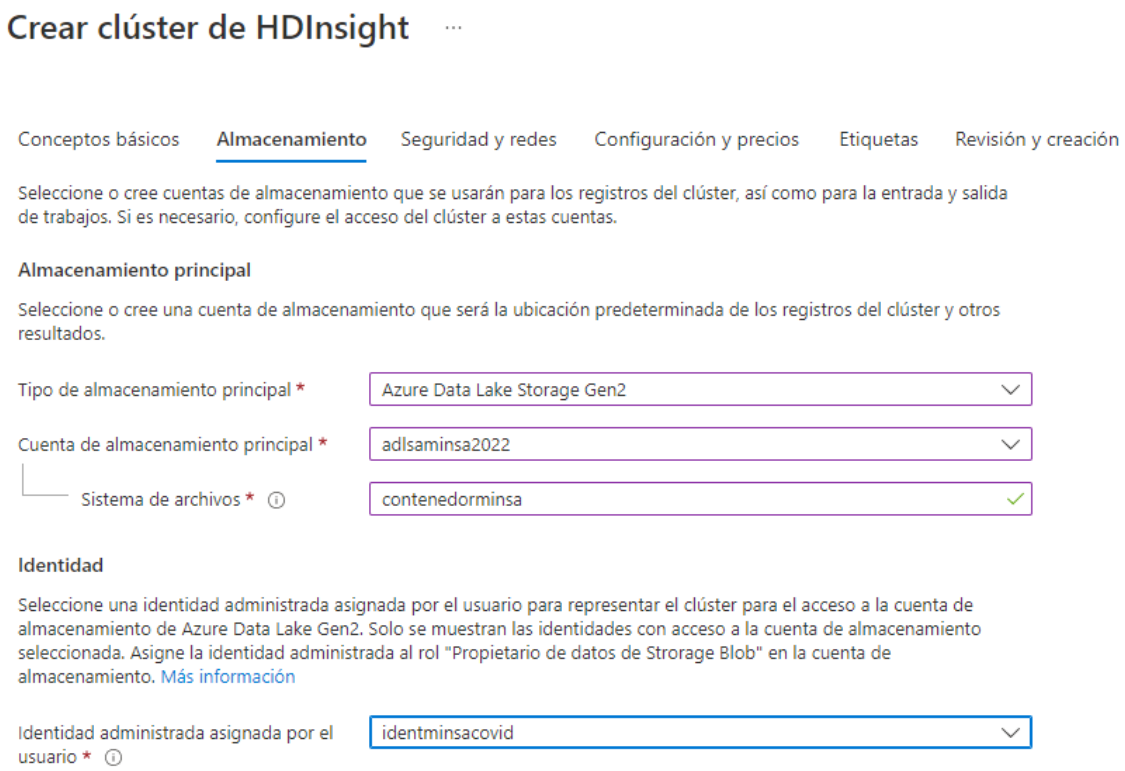
Luego seleccionamos el Data Lake creado anteriormente “adlsaminsa2022” y configuramos una identidad administrada asignada por el usuario para representar el clúster para el acceso a la cuenta de almacenamiento de Azure Data Lake Gen2. Solo se muestran las identidades con acceso a la cuenta de almacenamiento seleccionada. Asignamos la identidad administrada al rol "Propietario de datos de Strorage Blob" en la cuenta de almacenamiento.



Tambien se debe de Agregar en el Data Lake el acceso de control al usuario asignado para el administrador de identidades.



Se selecciona el Almacenamiento para el recurso HDInsight:

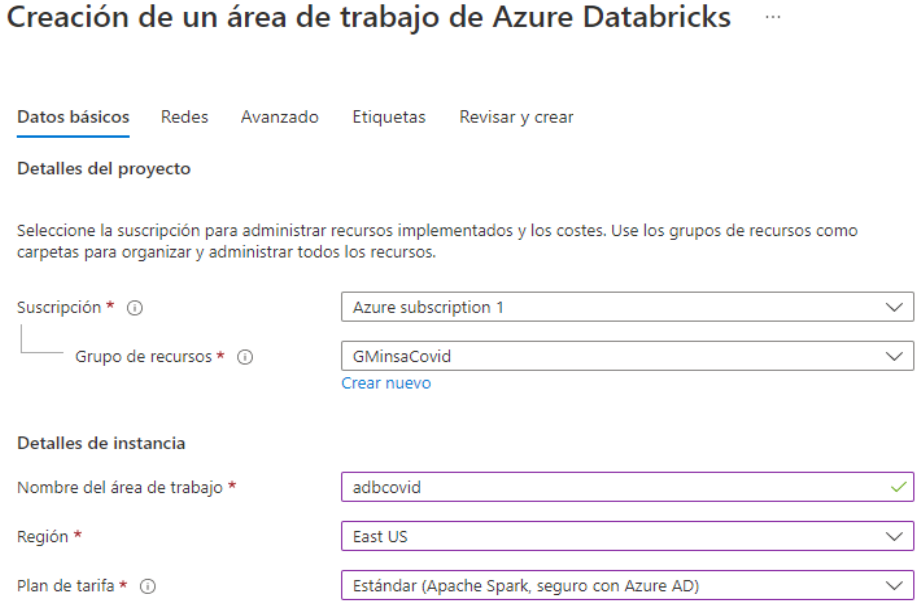


d) Creación y configuración del Servicio Data Factory

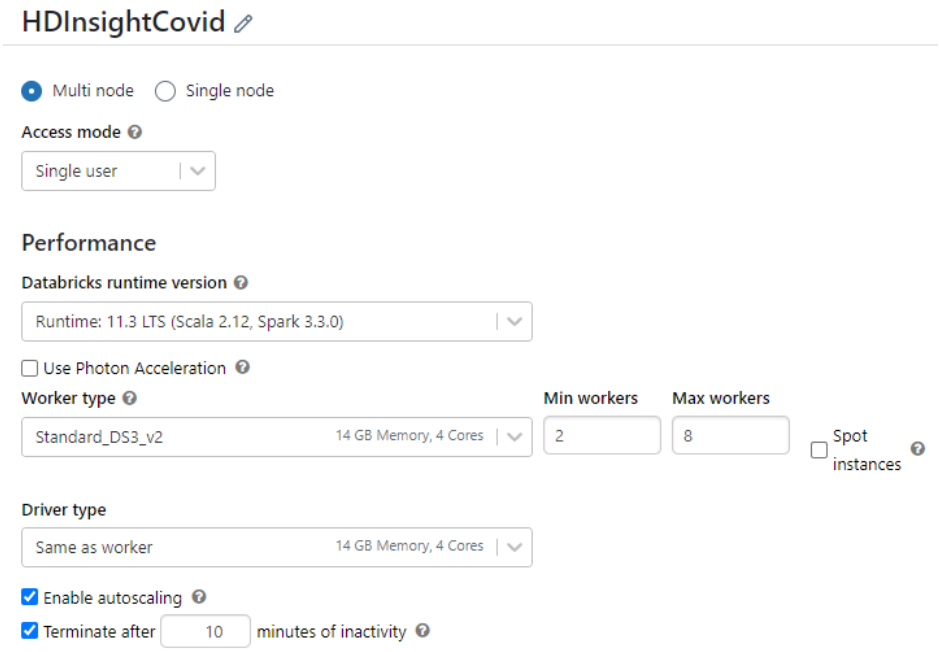
Creamos el servicio de Azure Data Factory para la integración y transformación de datos.

e) Creación y configuración del Servicio Databricks

El recurso de Azure Databricks nos va a proporcionar una plataforma en Azure sin administración basada en las funcionalidades de Apache Spark en un área de trabajo interactiva de exploración y visualización.

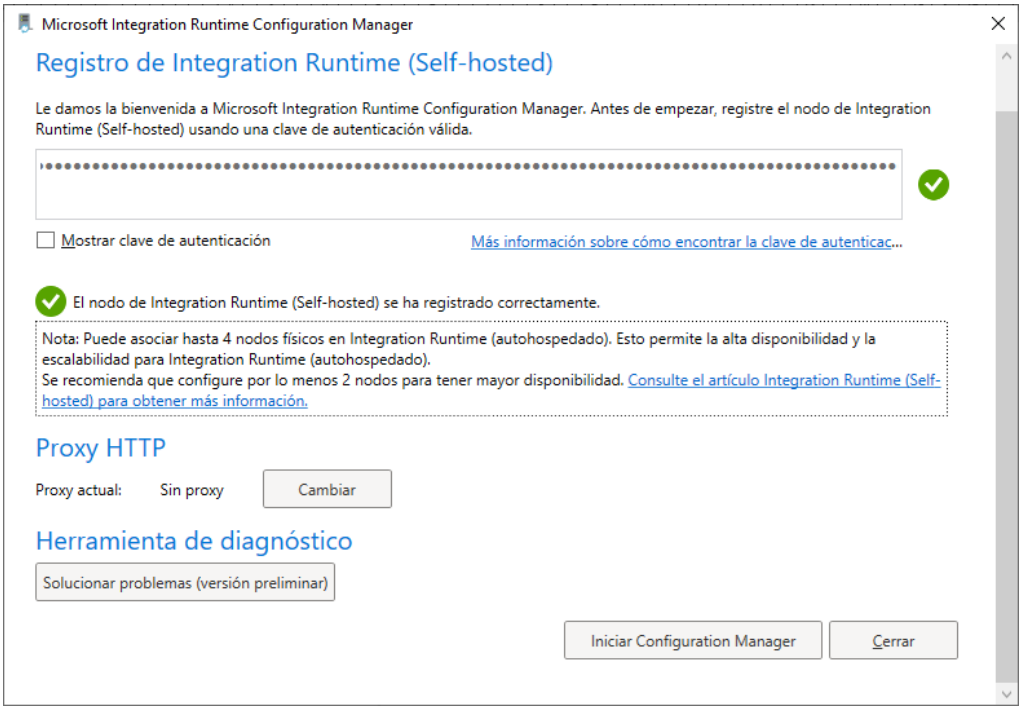


Luego lo configuramos al cluster:

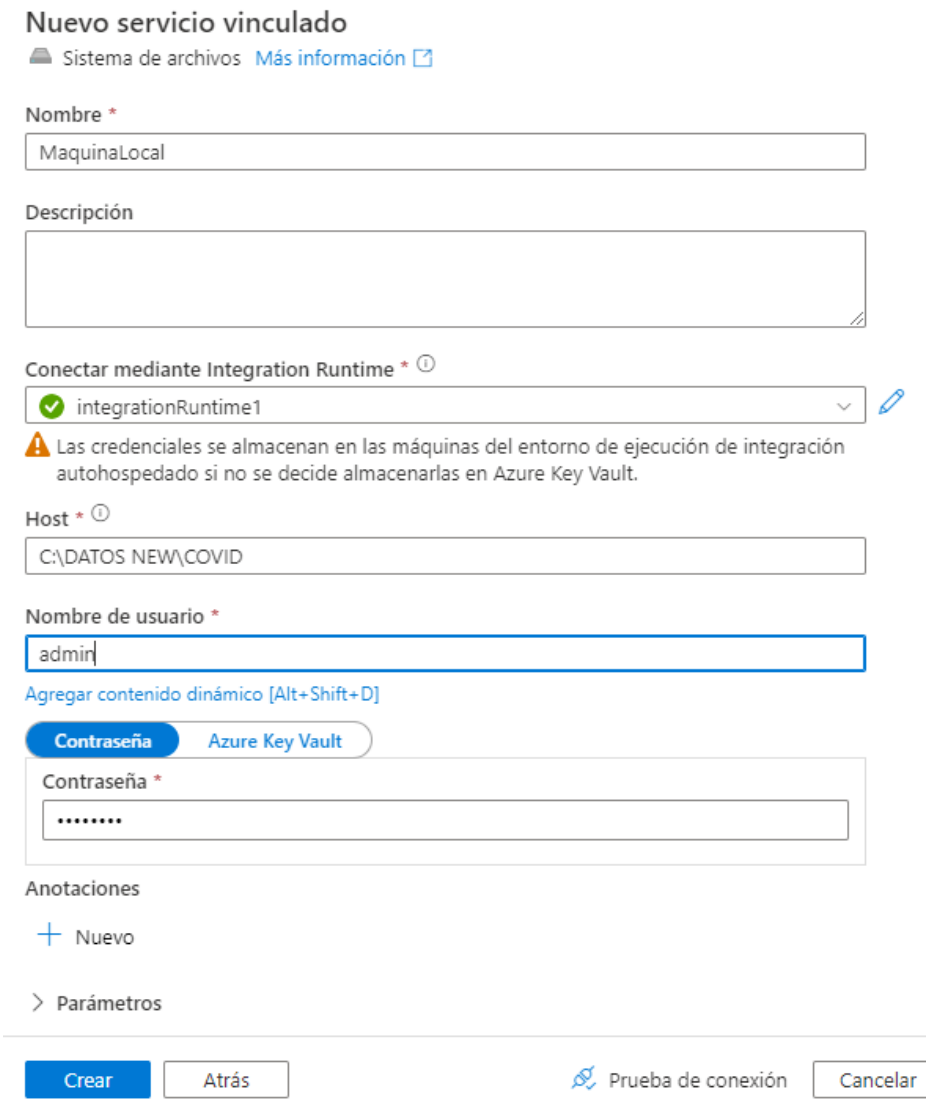


4.3.4. CARGA DE DATOS (INGESTA DE DATOS) EN EL CONTENEDOR DEL DATA LAKE

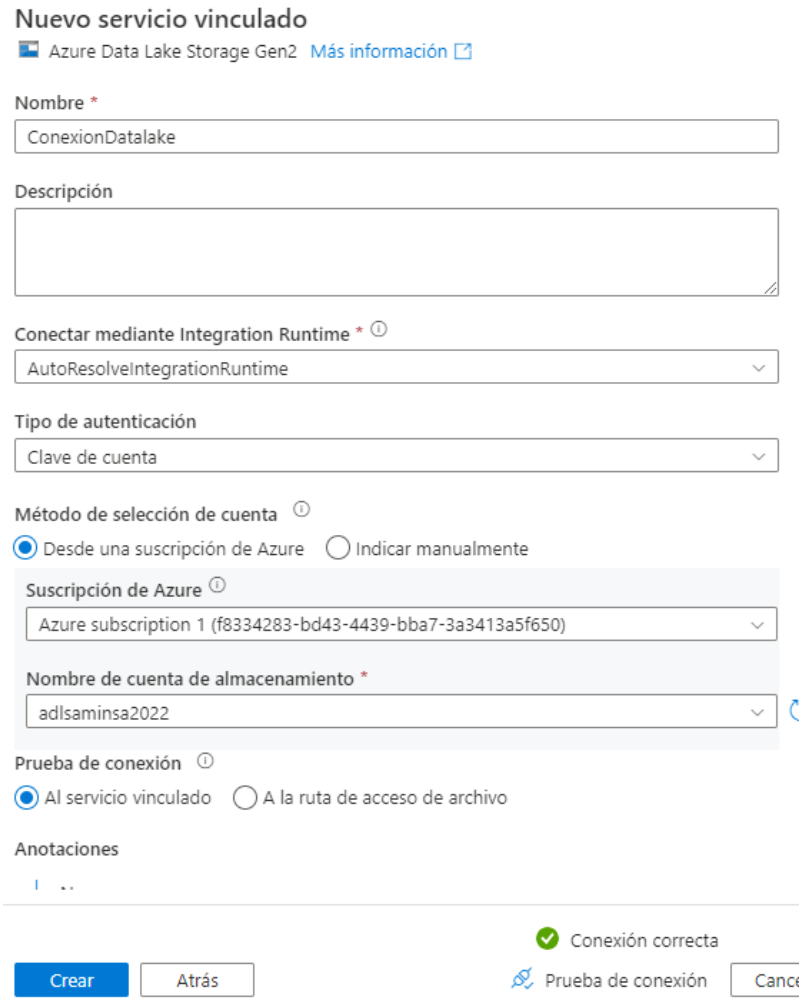
Para la ingesta de datos se realizó la configuración de Microsoft Integration Runtime



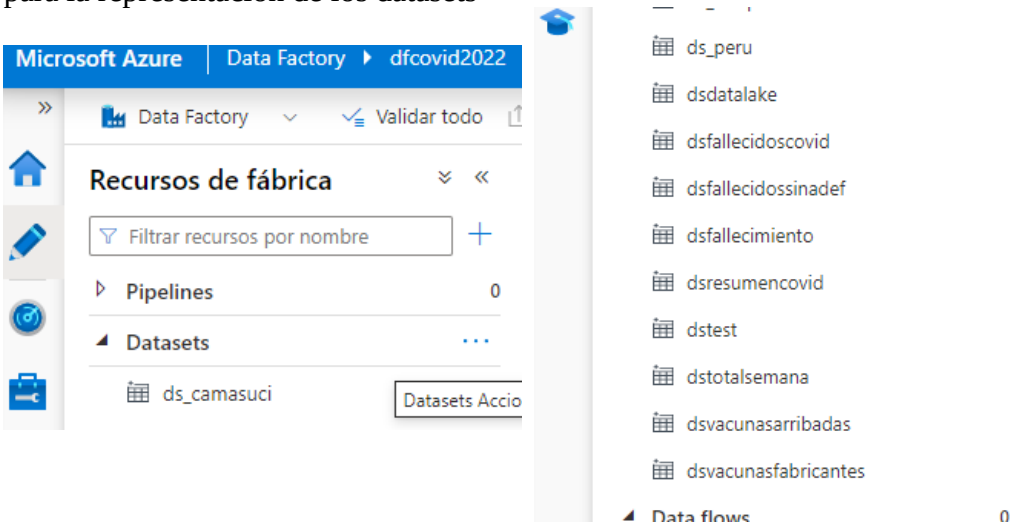
Luego se ha creado nuestros Linked Services a los archivos de orígenes de datos



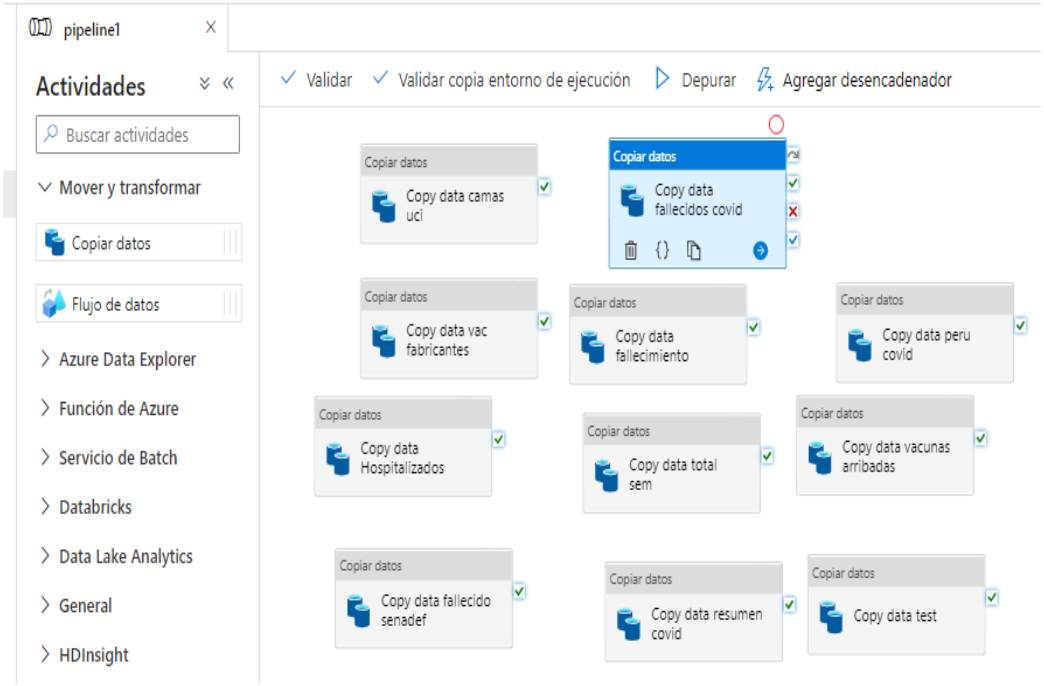
Luego se crea un linked service al Data Lake



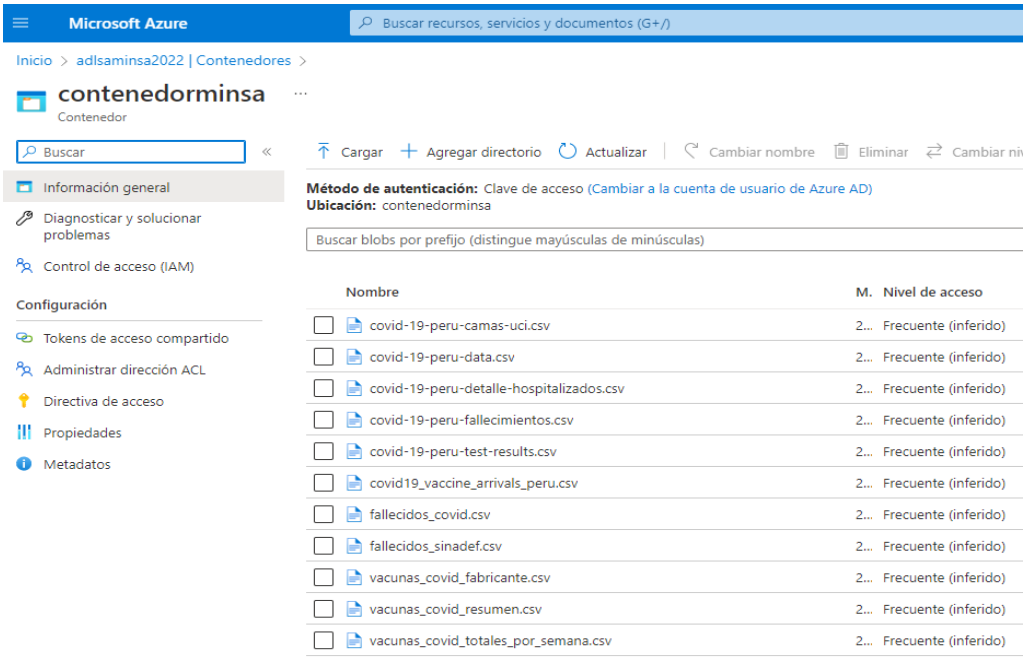
Para realizar la ingesta de los datos también es necesario crear las conexiones para la representación de los datasets



Se crea una canalización par la ingesta de datos de manera directa.



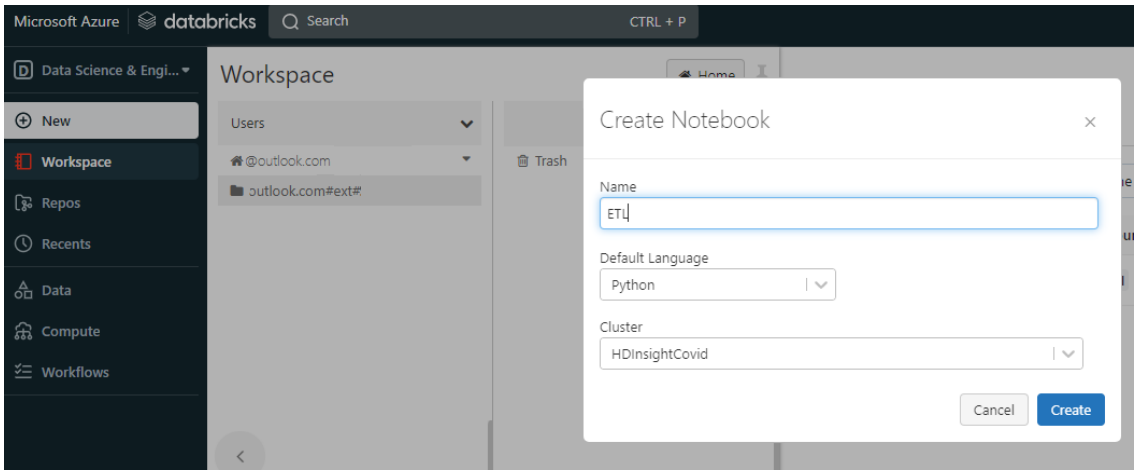
Verificamos la carga de datos. También se verifica que el contenedor haya recibido los archivos.



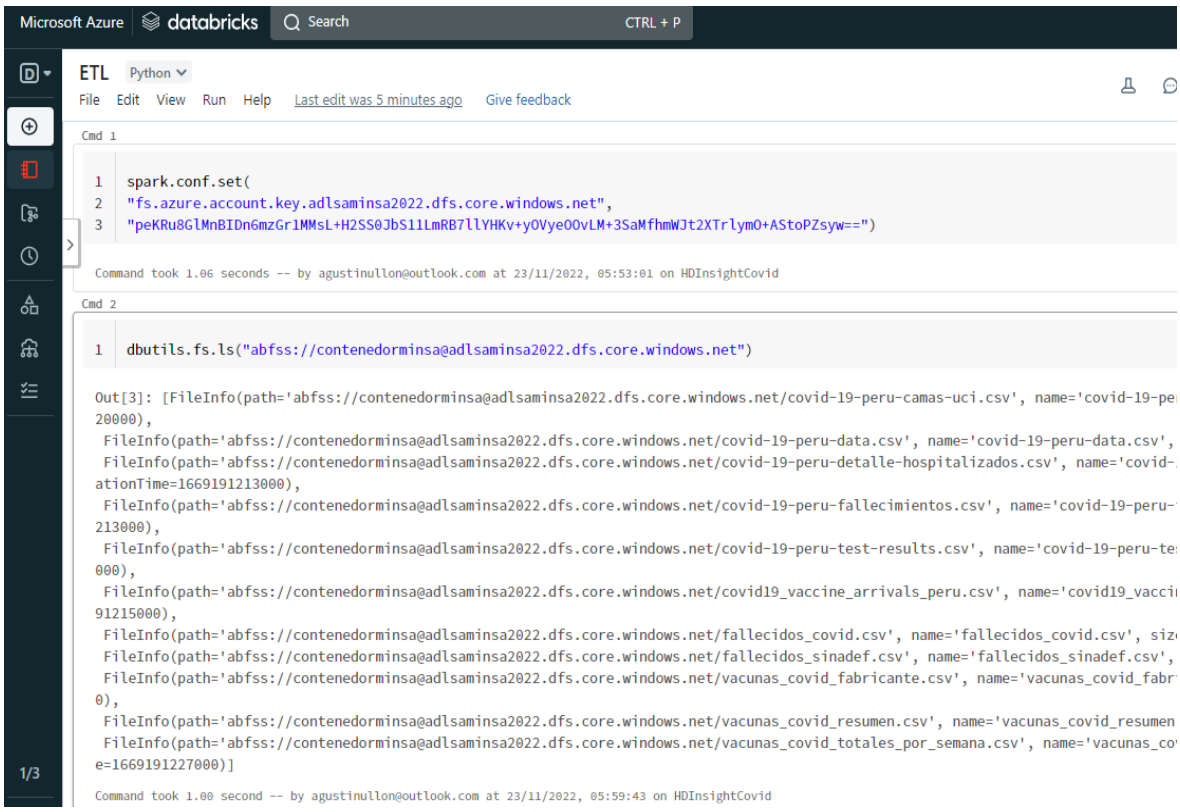
4.4.ANALIZAR

4.4.1. CREACIÓN DEL ETL EN AZURE DATABRICKS

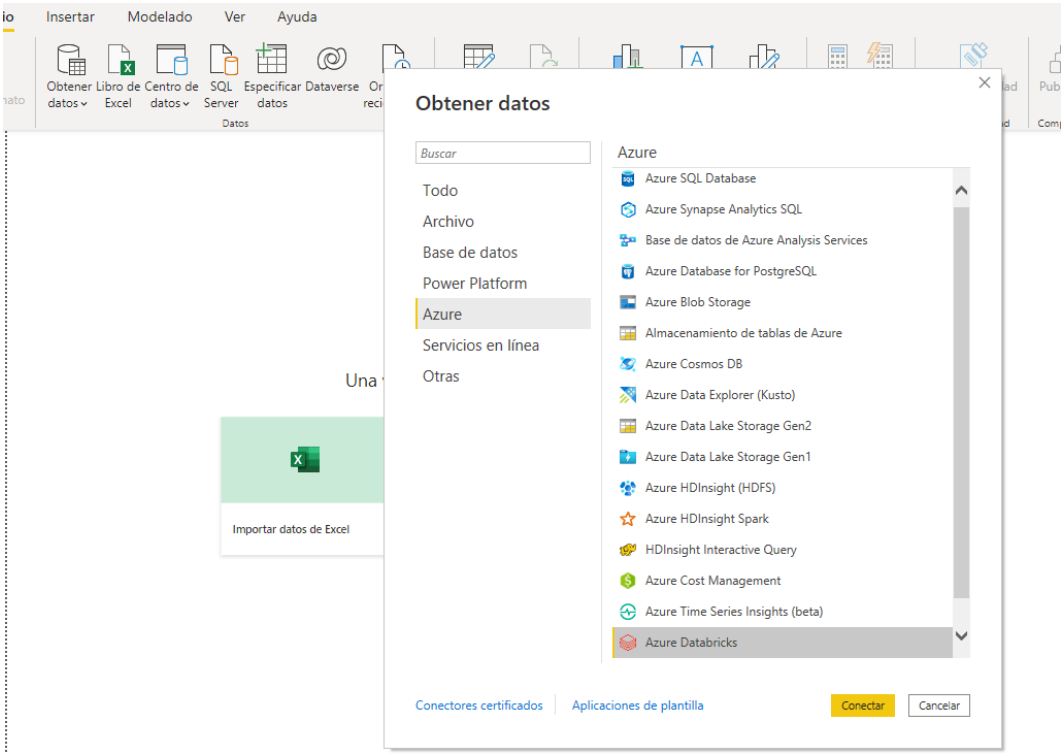
Para el proceso de ETL creamos un Notebook en Python dentro del workspace de Databricks:



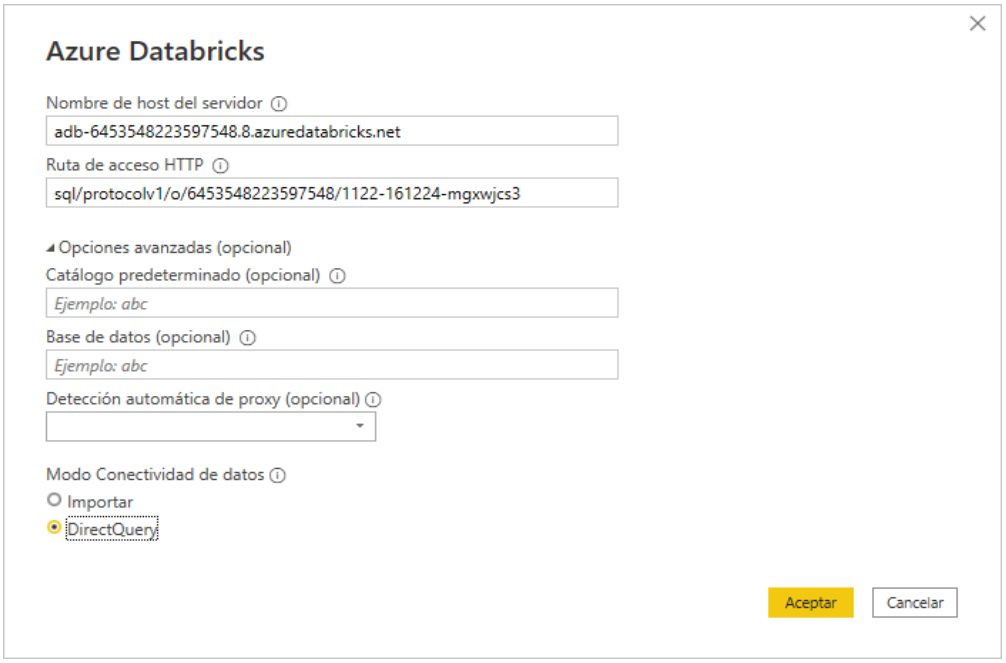
Luego se ejecuta la carga de trabajo en el notebook de Python



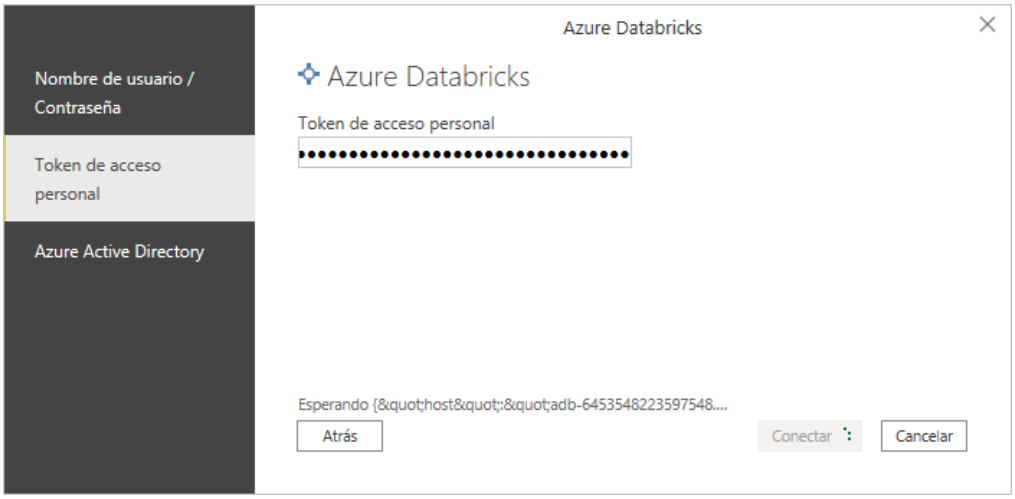
4.4.2. CONEXIÓN A DATABRICKS DESDE POWER BI



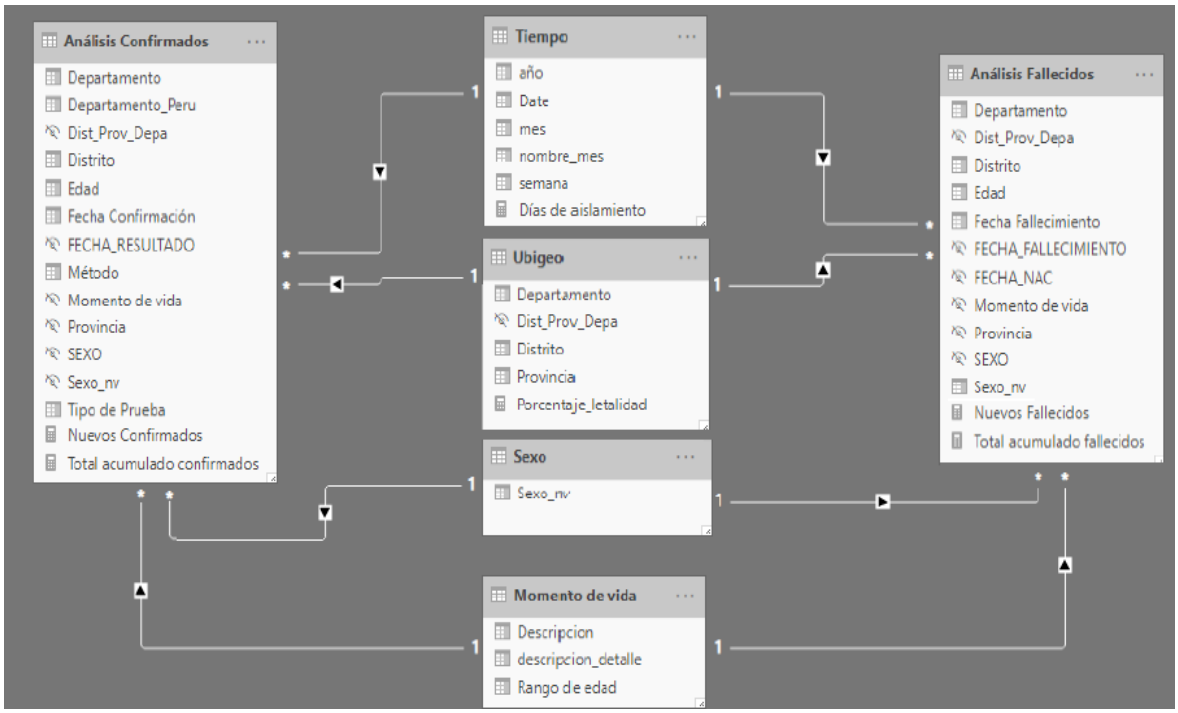
Establecemos la conexión con databricks colocando las credenciales del servicio en Azure



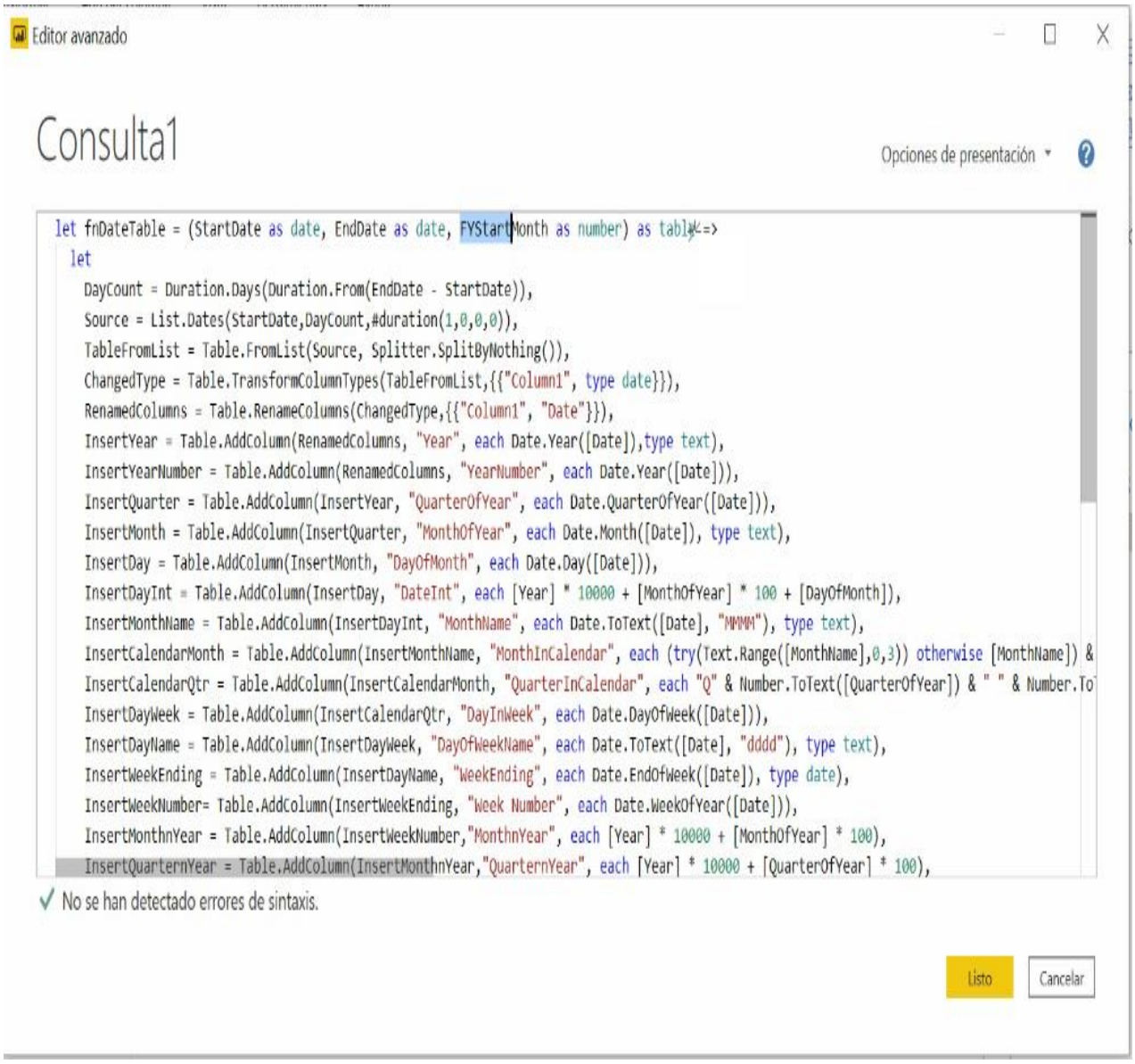
La conexion se hace a través de un Token Personal

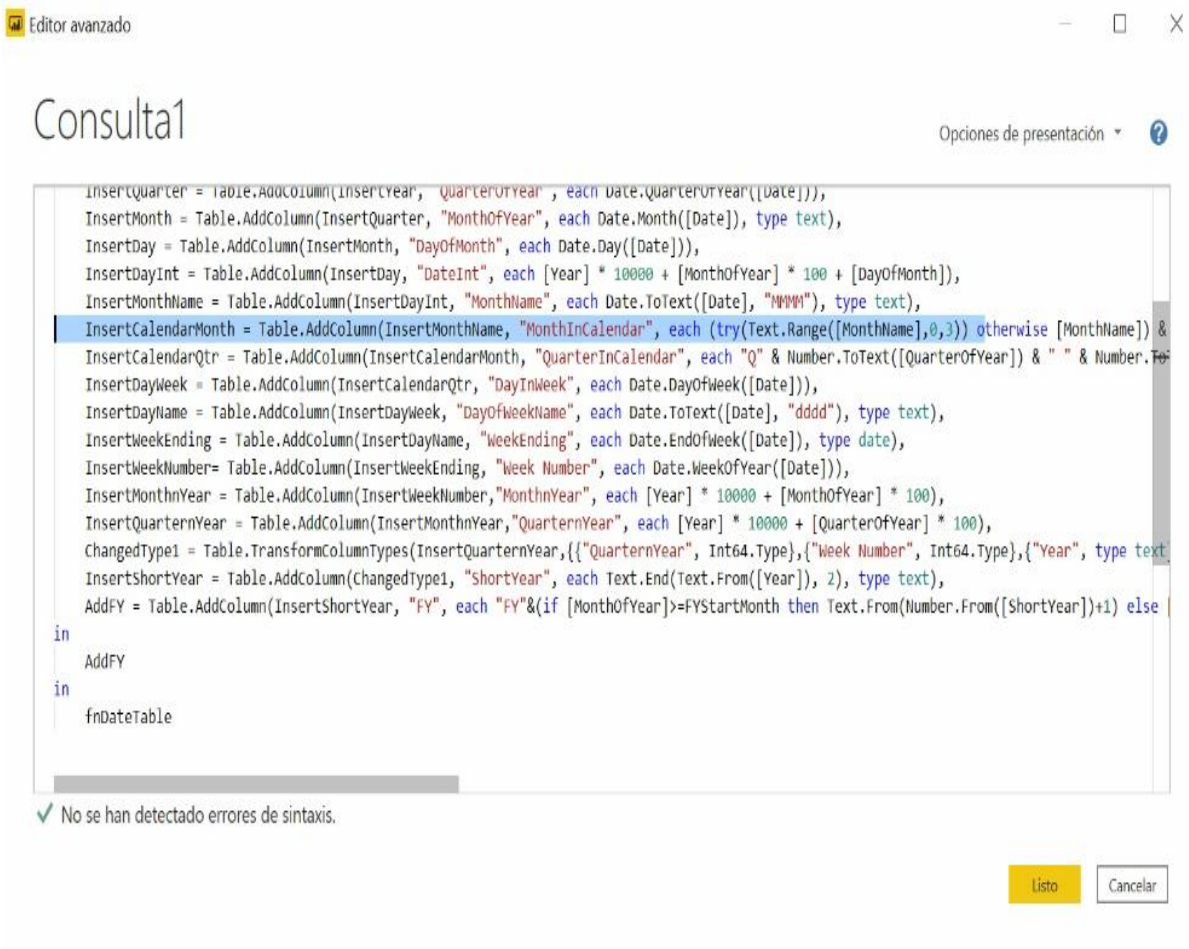


4.4.3. CREACION DEL MODELO DIMENSIONAL



4.4.4. CREACION DE LA DIMENSION TIEMPO





4.5.VISUALIZACION DE LA INFORMACIÓN

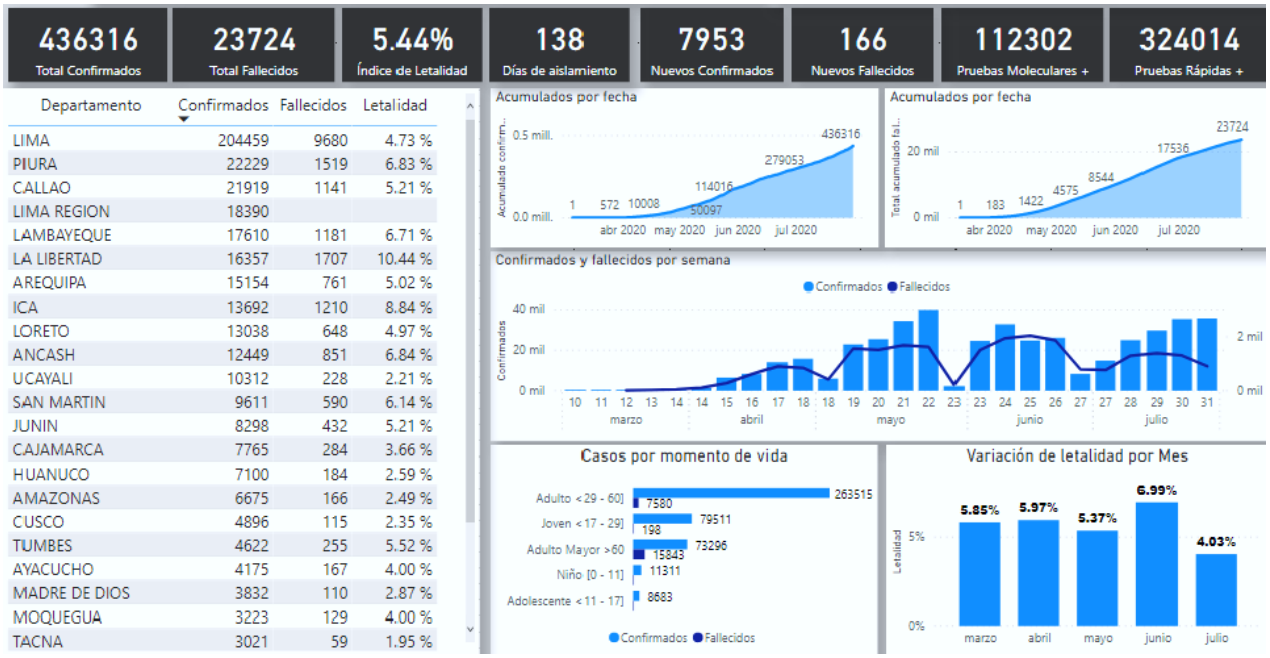
4.5.1. APLICACIÓN A UTILIZAR

Componentes del Proceso	Herramientas
Construcción de Interfaces	Power BI Desktop

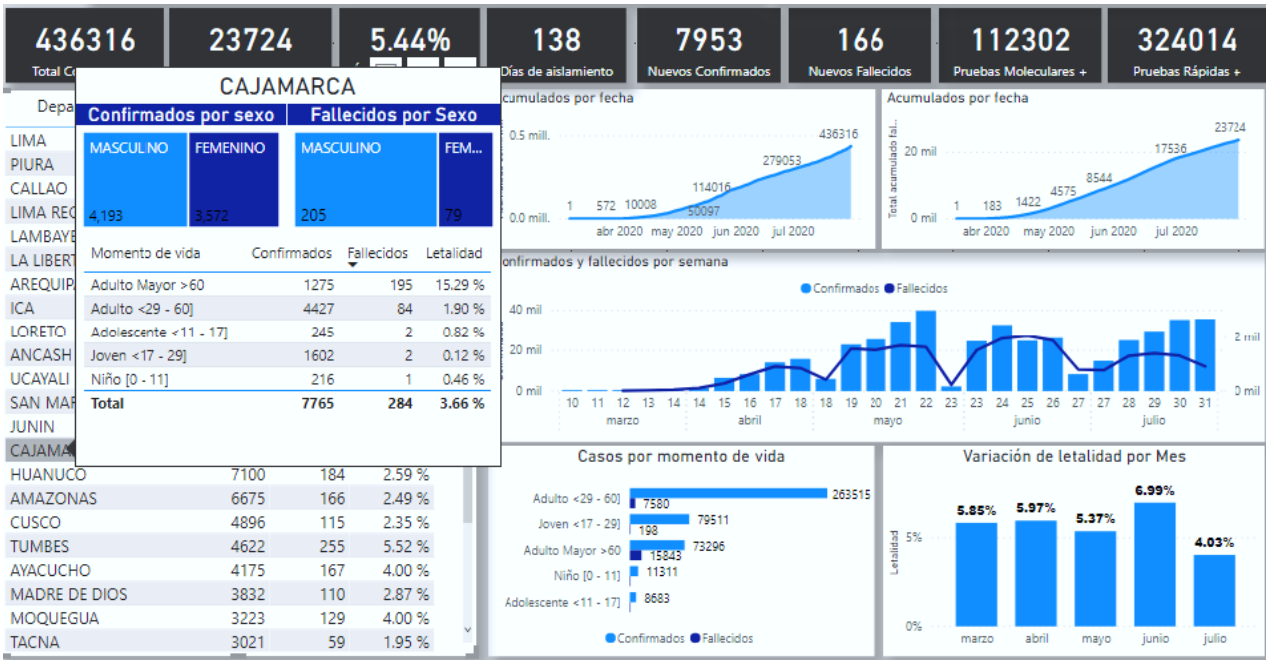
4.5.2. APLICACIÓN DEL USUARIO FINAL

El dashboard implementado se realizó tomando en cuenta los requerimientos de los interesados en la información

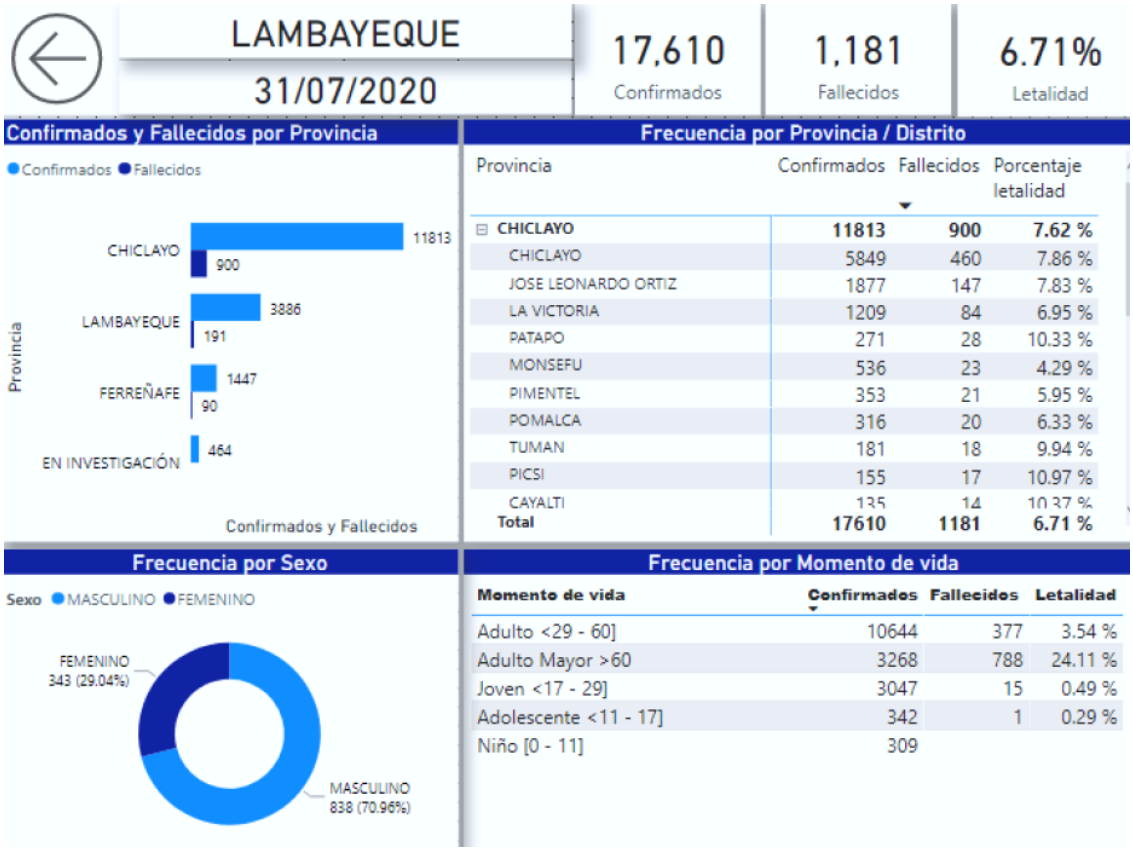
- Dashboard: Resumen de personas con resultado positivo y fallecidos por departamento



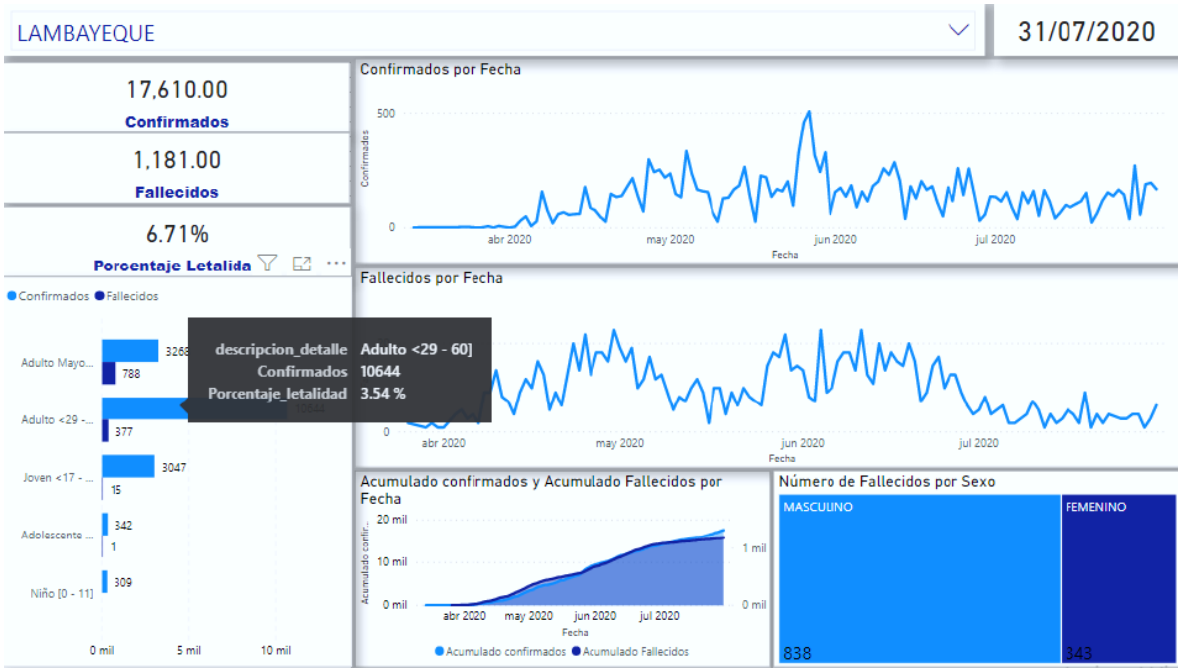
- Dashboard: Cuadro emergente con detalle de sexo de personas contagiadas y fallecida por departamento



- Dashboard: Detalle a nivel de provincia y distrito



- Dashboard: Detalle de comportamiento de los contagios y fallecidos a nivel de provincia, distrito y por edad



5. DISCUSION DE RESULTADOS

5.1.Formulación del Problema

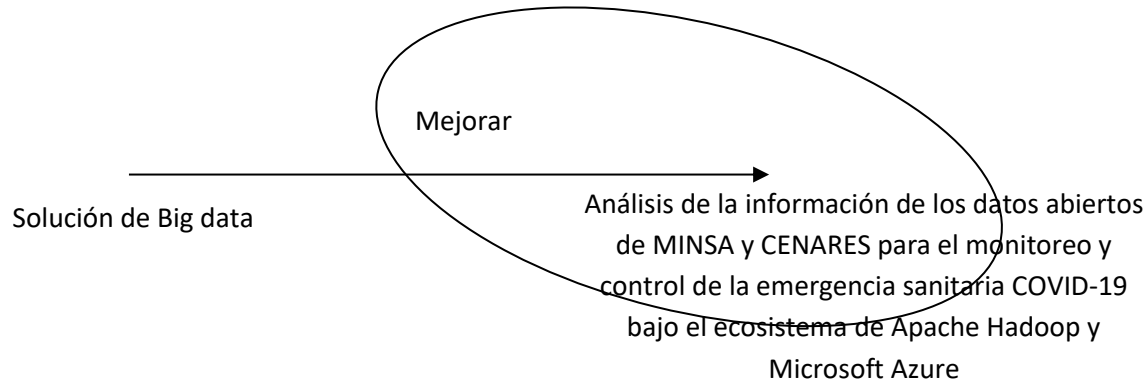
¿Cómo se puede mejorar el análisis de información para las personas y/o entidades encargadas en tomar decisiones basado en el comportamiento de casos positivos y fallecidos contra el COVID-19?

5.2.Hipótesis

"Una Solución de Big data permitirá mejorar análisis de la información de los datos abiertos de MINSA y CENARES para el monitoreo y control de la emergencia sanitaria COVID-19 bajo el ecosistema de Apache Hadoop y Microsoft Azure"

- ✓ Independiente (VI): Big data
- ✓ Dependiente (VD): Análisis de la información de los datos abiertos de MINSA y CENARES para el monitoreo y control de la emergencia sanitaria COVID-19 bajo el ecosistema de Apache Hadoop y Microsoft Azure.

5.3.MANERA PRESENCIAL



5.4.DISEÑO PREEXPERIMENTAL PRE-PRUEBA Y POST-PRUEBA

Para la contrastación de la hipótesis se utilizó el método de diseño lineal llamado adecuadamente también PRE-TEST Y POST – TEST.

Donde:

C1 = Control ANTES de la implementación del sistema del dashboard

X = Implementación del dashboard.

C2 = Control DESPUES de la implementación del sistema del dashboard.

5.4.1. CÁLCULO DE LOS INDICADORES DE LA HIPÓTESIS

Para el cálculo de los indicadores de la hipótesis se realizó un cuestionario evaluando a los usuarios que interactúan con el Dashboard implementado.

- Rango de valoración:

RANGO	VALORACIÓN
1	Desacuerdo
2	Regular
3	Bueno
4	Muy Bueno
5	Excelente

Tabla 9: Grado de valoración

ESCALA DE VALORACION	
Inadecuado	**00-30**
Adecuado	**31- 50**

5.4.2. APLICACIÓN DEL RANGO DE VALORACIÓN

N°	*INDICADORES*	VALORACION					$\overline{X}$
		1	**2**	**3**	**4**	**5**	
1	Los datos del dashboard son suficientes				2		4
2	Los gráficos visualizados son claros				2		4
3	Los gráficos son los requeridos					2	5
4	El Dashboarrd cubren con las expectativas.					2	5
5	EN el dashboard es posible cruces atributos				2		4
6	El acceso al dashboard es lo requerido				2		4
7	Se puede acceder al dashboard desde cualquier dispositivo electrónico					5	5
8	La interfaz es amigable y de fácil entendimiento				1	1	4.5
9	Es acertada la decisión de la implementación del dashboard.				1	1	4.5
10	Se encuentra satisfecho con el dashboard					2	5
$\sum \overline{X}$							**45**

Donde: X= (Valor Valoración* Número de empleados respondieron en nivel valoración)/2

Tabla 10: Evaluación de los indicadores de la hipótesis.

Interpretación: De acuerdo a la escala de valoración la Solución de Big data permite mejorar análisis de la información de MINSA y CENARES para el monitoreo y control de la emergencia sanitaria COVID-19, por ser sumatoria de los promedios 45 y superior a 31.

5.4.3. ANÁLISIS ESTADÍSTICO

Paso 1: Planteamiento

$Ho : O_1 >= O_2$

$H_1 : O_2 >= O_1$

Dónde:

Ho es la hipótesis Nula: "Una Solución de Big data no permite mejorar análisis de la información de los datos abiertos de MINSA y CENARES para el monitoreo y control de la emergencia sanitaria COVID-19 bajo el ecosistema de Apache Hadoop y Microsoft Azure."

H_1 es la hipótesis Alternativa: "Una Solución de Big data permite mejorar análisis de la información de los datos abiertos de MINSA y CENARES para el monitoreo y control de la emergencia sanitaria COVID-19 bajo el ecosistema de Apache Hadoop y Microsoft Azure"

Paso 2: Nivel de significancia.

α = 0.05.

Paso 3: Prueba estadística.

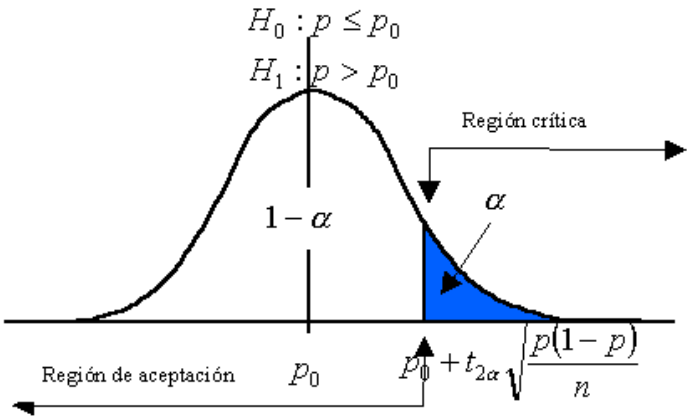
T-student por que el número de encuestados = 2.

Paso 4: Zona de rechazo.

Para todo valor de probabilidad mayor que 0.05, se acepta Ho y se rechaza H1.

Si la t**c** > t**t** se rechaza Ho y se acepta H1.

Dónde: t**c** es la t calculada y t**t** es la t de tabla



Paso 5: Calculo de t**t y** t**c**

Calculo de la t de tabla t**t**

t**t** (95%, 2) = 2,92 → Anexo C.

Calculo de la t calculado t**c**

$$\overline{D} = \frac{\sum D}{n}, \delta = \sqrt{\frac{\sum(Di - \overline{D})^2}{n-1}}, t_c = \frac{\overline{D}}{\frac{\delta}{\sqrt{n}}}$$

Donde:

t_c	:	T calculado.
δ	:	Desviación estándar
n	:	Tamaño de la muestra
$\overline{D}$	:	Valor promedio o media aritmética de las diferencias entre los momentos antes y después.

CÁLCULO DEL VALOR DE "T CALCULADO"

Los valores que los entrevistados dieron a las respuestas del cuestionario fueron aplicados de acuerdo al rango de satisfacción que se muestra en la siguiente tabla:

RANGO	GRADO DE SATISFACION
0 – 2.5	Insatisfecho
2.5 – 5.0	Medianamente Satisfecho
5.0 – 7.5	Satisfecho
7.5 – 10.0	Muy Satisfecho

EVALUACION DEL GRADO DE SATISFACCION:

	INDICADORES	Media Pre U_1	Media Post U_2	$D= (U_2- U_1)$	$(Di - \overline{D})$	$(Di - \overline{D})^2$
1	Detalle de cantidad de confirmados y fallecidos	4.0	9.0	5	-1.3	1.69
2	Detalle de la cantidad de pruebas por tipo	4.0	9.0	5	-1.3	1.69
3	Análisis de la cantidad de confirmados y fallecidos mensualmente, semanalmente y diariamente	5.0	9.0	4	-2.3	5.29
4	Niveles de detalle por departamento, provincia y distrito	4.0	9.0	5	-1.3	1.69
5	Detalle de confirmados y fallecidos por sexo	5.0	9.0	4	-2.3	5.29
6	Detalle de confirmados y fallecido por momento de vida	2.0	10.0	8	1.7	2.89
7	Detalle del índice de letalidad	2.0	9.0	7	0.7	0.49
8	Detalles de personas vacunadas	2.0	8.0	6	-0.3	0.09
9	Detalle de camas UCI en uso	1.0	10.0	9	2.7	7.29
10	Detalle de número de personas hospitalizadas	0.0	10.0	10	3.7	13.69

N = 10 ; $\sum$D= 62 ; $\overline{D}$ = 6.2 ; $\sum(D_i - \overline{D})^2$ = 39.1 ; δ =2.11 ; $\sqrt{n}$ = 3.16

$$t_c = \frac{\overline{D}}{\frac{\delta}{\sqrt{n}}}$$

$t_c = 9.42$

Interpretación: de acuerdo al resultado anterior se acepta la hipótesis alternativa "Una Solución de Big data permite mejorar análisis de la información de los datos abiertos de MINSA y CENARES para el monitoreo y control de la emergencia sanitaria COVID-19 bajo el ecosistema de Apache Hadoop y Microsoft Azure."

5.5. CUADRO DE LA COMPARACIÓN DE TIEMPO DE DEMORA EN LA EJECUCIÓN DE LAS CONSULTAS.

NRO	CONSULTAS	SISTEMA ON PREMISE	SOLUCION DE BIG DATA
1	Detalle de cantidad de confirmados y fallecidos	30 seg	02 seg.
2	Detalle de la cantidad de pruebas por tipo	180 seg	03 seg.
3	Análisis de la cantidad de confirmados y fallecidos mensualmente, semanalmente y diariamente	360 seg	02 seg.

4	Niveles de detalle por departamento, provincia y distrito	100 seg	02 seg.
5	Detalle de confirmados y fallecidos por sexo	600 seg	02 seg
6	Detalle de confirmados y fallecido por momento de vida	20 seg	03 seg
7	Detalle del índice de letalidad	100 seg	02 seg
8	Detalles de personas vacunadas	200 seg	02 seg
9	Detalle de camas UCI en uso	60 seg	02 seg
10	Detalle de número de personas hospitalizadas	100 seg	03 seg
		Fuente: Solución on-premise	Fuente: Solución Big Data y Dashboard

Interpretación: los requerimientos son procesados con menor tiempo utilizando una solución de big data.

6. CONCLUSIONES

- El análisis de la situación de emergencia en la que se ve envuelta la población llegó a determinar la necesidad de información resumida en 10 requerimientos para poder mantener con un apoyo en la toma de decisiones en Minsa sobre contagios, fallecimientos, letalidad y edad de los contagiados de Covid 19.

- El análisis de estos requerimientos llevo a consolidar un modelo estrella constelación basado en 2 tablas de hechos y 4 dimensiones, y por la cantidad de datos que se insertaran se propone una arquitectura basada en la nube de Azure.

- La arquitectura de Big Data en la nube de Microsoft Azure, permitirá una escalabilidad en servicios, pero en esta solución se configuró inicialmente los principales servicios para que de soporte a los requerimientos, entre los cuales fueron, un Data Lake con su contenedor, un Data Factory, permisos y asignación de roles para el acceso al clúster creado en HDInsight (Hadoop).

- El dashboard creado en Power BI permite visualizar de forma dinámica la información apreciando con detalle la solución a cada requerimiento.

7. RECOMENDACIONES

- En cada paso incluido en la metodología, es deseable mantener relaciones con los usuarios del área de ejecución del proyecto, lo que ayuda a determinar lo necesario para el proyecto, especialmente en el diseño de cuadros de mando.

- Se recomienda documentar y probar minuciosamente el uso de herramientas en la nube antes de utilizarlas en "producción".

- Se recomienda la capacitación del usuario final para comprender mejor la solución desarrollada y poder usarla correctamente.

- A futuro, se sugieren otras nuevas tendencias que emergen y entran en el campo del análisis de datos, como los modelos predictivos que utilizan el aprendizaje automático, la IA, que son algunos de los representantes de la gran revolución en soluciones de análisis de datos y el campo de la ciencia de datos.

8. REFERENCIAS BIBLIOGRAFICAS

Azure HDInsight. (22 de 05 de 2019). *Azure HDInsight*. Obtenido de https://docs.microsoft.com/es-es/azure/hdinsight/

Big Data SAC. (05 de 08 de 2019). *Metodologia ICAV*. Obtenido de http://www.bigdata.pe/web/index.php/metodologia

Bit. (19 de 05 de 2017). *Servicios de intelligence y analytics en microsoft azure*. Obtenido de https://www.bit.es/knowledge-center/servicios-de-intelligence-y-analytics-en-microsoft-azure-i/

Gartner. (02 de 2019). *Magic Quadrant for Analytics and Business Intelligence Platforms*. Obtenido de https://www.gartner.com/en/webinars/3900973/the-2019-analytics-and-bi-magic-quadrant-highlights

Iebschool. (15 de 02 de 2019). *Glosario Big Data*. Obtenido de https://www.iebschool.com/blog/glosario-big-data/

Ionos. (13 de 03 de 2019). *Apache Hadoop: sistema de archivos distribuido*. Obtenido de https://www.ionos.es/digitalguide/servidores/know-how/apache-hadoop-el-framework-para-big-data/

Microsoft. (08 de 05 de 2019). *¿Qué es Azure?* Obtenido de https://azure.microsoft.com/es-es/overview/what-is-azure/

MINSA. (10 de 10 de 2020). *minsa.gob.pe*. Obtenido de https://www.minsa.gob.pe/transparencia/index.asp?op=103

Minsa. (01 de 07 de 2022). *Datos Abiertos Covid 19*. Obtenido de https://www.datosabiertos.gob.pe/group/datos-abiertos-de-covid-19

Prometeusgs. (01 de 02 de 2019). *Sin análisis no hay información útil. La importancia del Data Analytics en tu negocio*. Obtenido de https://prometeusgs.com/analisis-de-datos-informacion-util/

Tableau. (15 de 05 de 2020). *¿Qué es la inteligencia de negocios y por qué es importante?* Obtenido de https://www.tableau.com/es-es/learn/articles/business-intelligence

Talend. (29 de 01 de 2019). *¿En qué consiste la integración de datos?* Obtenido de https://es.talend.com/resources/what-is-data-integration/

Tecon. (25 de 04 de 2019). *¿Qué es Microsoft Azure? ¿Cómo funciona?* Obtenido de https://www.tecon.es/que-es-microsoft-azure-como-funciona/

Workana. (24 de 02 de 2020). *¿Qué es un Dashboard?* Obtenido de https://www.workana.com/i/glosario/que-es-un-dashboard/

ANEXOS

ANEXO A

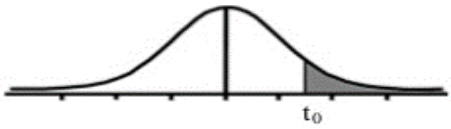
CUESTIONARIO DIRIGIDO: Director Hospital Regional, jefa de enfermería Hospital Regional

PREGUNTAS	VALORES										
	0	1	2	3	4	5	6	7	8	9	10
Detalle de cantidad de confirmados y fallecidos											
Detalle de la cantidad de pruebas por tipo											
Análisis de la cantidad de confirmados y fallecidos mensualmente, semanalmente y diariamente											
Niveles de detalle por departamento, provincia y distrito											
Detalle de confirmados y fallecidos por sexo											
Detalle de confirmados y fallecido por momento de vida											
Detalle del índice de letalidad											
Detalles de personas vacunadas											
Detalle de camas UCI en uso											
Detalle de número de personas hospitalizadas											

Tabla A1.

ANEXO B

Tabla t-Student



Grados de libertad	0.25	0.1	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3137	12.7062	31.8210	63.6559
2	0.8165	1.8856	2.9200	4.3027	6.9645	9.9250
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8408
4	0.7407	1.5332	2.1318	2.7765	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9979	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.6864	1.3232	1.7207	2.0796	2.5176	2.8314
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6848	1.3178	1.7109	2.0639	2.4922	2.7970
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500
31	0.6825	1.3095	1.6955	2.0395	2.4528	2.7440
32	0.6822	1.3086	1.6939	2.0369	2.4487	2.7385
33	0.6820	1.3077	1.6924	2.0345	2.4448	2.7333
34	0.6818	1.3070	1.6909	2.0322	2.4411	2.7284
35	0.6816	1.3062	1.6896	2.0301	2.4377	2.7238
36	0.6814	1.3055	1.6883	2.0281	2.4345	2.7195
37	0.6812	1.3049	1.6871	2.0262	2.4314	2.7154
38	0.6810	1.3042	1.6860	2.0244	2.4286	2.7116
39	0.6808	1.3036	1.6849	2.0227	2.4258	2.7079
40	0.6807	1.3031	1.6839	2.0211	2.4233	2.7045
41	0.6805	1.3025	1.6829	2.0195	2.4208	2.7012
42	0.6804	1.3020	1.6820	2.0181	2.4185	2.6981
43	0.6802	1.3016	1.6811	2.0167	2.4163	2.6951
44	0.6801	1.3011	1.6802	2.0154	2.4141	2.6923
45	0.6800	1.3007	1.6794	2.0141	2.4121	2.6896
46	0.6799	1.3002	1.6787	2.0129	2.4102	2.6870
47	0.6797	1.2998	1.6779	2.0117	2.4083	2.6846
48	0.6796	1.2994	1.6772	2.0106	2.4066	2.6822
49	0.6795	1.2991	1.6766	2.0096	2.4049	2.6800