Imagen que contiene Forma

Descripción generada automáticamente

**PROYECTO AGORA**

**DESCRIPCIÓN ENTREGABLES FASE 1**

**DOCUMENTO INICIAL**

**Julio 2024**

**Control de cambios**

|  |  |  |  |
| --- | --- | --- | --- |
| **VERSIÓN** | **FECHA** | **MODIFICACIÓN** | **RESPONSABLE** |
| 1.0 | 30 de julio de 2024 | Creación documento descriptivo entregables Fase 1 | Pedro Fabián Pérez Arteaga |

Documentos descriptivos de fuentes entregadas en la Fase 1

[[](https://livejaverianaedu.sharepoint.com/:w:/s/ProyectoAGORA371/EXScjMLVk71EmraaGALN0qIB-_bKQO5KKm5HsyvvcUlDqg?e=2J9Ug4)1. Documento entendimiento de los datos-RIPS- 2024-1.5v.docx](https://livejaverianaedu.sharepoint.com/:w:/s/ProyectoAGORA371/EXScjMLVk71EmraaGALN0qIB-_bKQO5KKm5HsyvvcUlDqg?e=2J9Ug4)

El presente documento ofrece una visión detallada de los principales hallazgos derivados del proceso de exploración inicial de los datos del sistema RIPS (Registro Individual de Prestación de Servicios de Salud). En esta fase inicial de análisis, se identificó la fuente de datos y se examinaron las variables contenidas en ella y sus características fundamentales.

La fuente de datos denominada RIPS, comprende información recopilada durante el periodo comprendido entre 2009 y 2022. Este periodo permite analizar una variedad de contextos y condiciones que resultan ser importantes para la comprensión y el análisis de las tendencias relacionadas con la prestación de servicios de salud en Colombia.

Durante la fase inicial de exploración, se realizó un análisis exhaustivo de las variables contenidas en el conjunto de datos del sistema RIPS. Se aplicaron técnicas estadísticas descriptivas básicas para examinar la distribución, la variabilidad y las tendencias presentes en los registros disponibles. Este análisis preliminar sirvió como punto de partida para comprender mejor la naturaleza de los datos y así poder identificar posibles áreas de interés y enfoque para análisis posteriores.

En estas secciones, se presenta la metodología de exploración de los datos, como los principales resultados obtenidos en esta fase inicial, incluyendo las características de las variables analizadas, las tendencias y patrones identificados en los datos del sistema RIPS.

[[](https://livejaverianaedu.sharepoint.com/:w:/s/ProyectoAGORA371/EeTg4xru3yZFlawKA3sbIJ4B77QWSBdN20X8Wgc6KrpU1w?e=WXOmQa)2. Documento entendimiento de los datos- ESTADÍSTICAS VITALES- 2024-1.5v.docx](https://livejaverianaedu.sharepoint.com/:w:/s/ProyectoAGORA371/EeTg4xru3yZFlawKA3sbIJ4B77QWSBdN20X8Wgc6KrpU1w?e=WXOmQa)

El presente documento ofrece una visión detallada de los principales hallazgos derivados del proceso de exploración inicial de los datos del sistema llamado ESTADISTICAS VITALES. En esta fase inicial de análisis, se identificó la fuente de datos y se examinaron las variables contenidas en ella y sus características fundamentales.

La fuente de datos denominada ESTADÍSTICAS VITALES, comprende información recopilada durante el periodo comprendido entre 2014 a 2023. Este periodo permite analizar una variedad de contextos y condiciones que resultan ser importantes para la comprensión y el análisis de las tendencias relacionadas.

Durante la fase inicial de exploración, se realizó un análisis exhaustivo de las variables contenidas en el conjunto de datos del sistema ESTADÍSTICAS VITALES. Se aplicaron técnicas estadísticas descriptivas básicas para examinar la distribución, la variabilidad y las tendencias presentes en los registros disponibles. Este análisis preliminar sirvió como punto de partida para comprender mejor la naturaleza de los datos y así poder identificar posibles áreas de interés y enfoque para análisis posteriores.

En estas secciones, se presenta la metodología de exploración de los datos, como los principales resultados obtenidos en esta fase inicial, incluyendo las características de las variables analizadas, las tendencias y patrones identificados en los datos del sistema ESTADÍSTICAS VITALES.

[[](https://livejaverianaedu.sharepoint.com/:w:/s/ProyectoAGORA371/EV0klsTH-s1JvQ1d8VYHmYYBK_KobV9txofnTCfYfqZOfQ?e=JWaCys)3. Documento entendimiento de los datos-SEGCOVID 2024-1.5v.docx](https://livejaverianaedu.sharepoint.com/:w:/s/ProyectoAGORA371/EV0klsTH-s1JvQ1d8VYHmYYBK_KobV9txofnTCfYfqZOfQ?e=JWaCys)

El presente documento ofrece una visión detallada de los principales hallazgos derivados del proceso de exploración inicial de los datos del sistema SEGCOVID. En esta fase inicial de análisis, se identificó la fuente de datos y se examinaron las variables contenidas en ella y sus características fundamentales.

La fuente de datos denominada SEGCOVID, comprende información recopilada durante el periodo comprendido entre 2020 a 2023. Este periodo permite analizar una variedad de contextos y condiciones que resultan ser importantes para la comprensión y el análisis de las tendencias relacionadas.

Durante la fase inicial de exploración, se realizó un análisis exhaustivo de las variables contenidas en el conjunto de datos, cabe mencionar que la información de SEGCOVID se generó a partir de una aplicación en donde se registraba el seguimiento a los casos de COVID-19 . Se aplicaron técnicas estadísticas descriptivas básicas para examinar la distribución, la variabilidad y las tendencias presentes en los registros disponibles. Este análisis preliminar sirvió como punto de partida para comprender mejor la naturaleza de los datos y así poder identificar posibles áreas de interés y enfoque para análisis posteriores.

En estas secciones, se presenta la metodología de exploración de los datos, como los principales resultados obtenidos en esta fase inicial, incluyendo las características de las variables analizadas, las tendencias y patrones identificados en los datos del sistema SEGCOVID.

[4. Documento entendimiento de los datos-SIVIGILA-2024-1.5v.docx](https://livejaverianaedu.sharepoint.com/:w:/s/ProyectoAGORA371/Ec6ptl0Lb4NBr9r0HtCrjakBaWmamQgVqVaLrjTTIUh1qw?e=ghTQeg)

El presente documento ofrece una visión detallada de los principales hallazgos derivados del proceso de exploración inicial de los datos del sistema SIVIGILA. En esta fase inicial de análisis, se identificó la fuente de datos y se examinaron las variables contenidas en ella y sus características fundamentales.

La fuente de datos denominada SIVIGILA, comprende información recopilada durante el periodo comprendido entre 2008 a 2022. Este periodo permite analizar una variedad de contextos y condiciones que resultan ser importantes para la comprensión y el análisis de las tendencias relacionadas.

Durante la fase inicial de exploración, se realizó un análisis exhaustivo de las variables contenidas en el conjunto de datos del sistema SIVIGILA. Se aplicaron técnicas estadísticas descriptivas básicas para examinar la distribución, la variabilidad y las tendencias presentes en los registros disponibles. Este análisis preliminar sirvió como punto de partida para comprender mejor la naturaleza de los datos y así poder identificar posibles áreas de interés y enfoque para análisis posteriores.

En estas secciones, se presenta la metodología de exploración de los datos, como los principales resultados obtenidos en esta fase inicial, incluyendo las características de las variables analizadas, las tendencias y patrones identificados en los datos del sistema SIVIGILA.

[[](https://livejaverianaedu.sharepoint.com/:w:/s/ProyectoAGORA371/EfIcY6b7kBBDh4gHWpig49sBq6vmySRRfmdyK6cFdt2SEQ?e=Ad2bDI)5. Documento entendimiento de los datos-VACUNASCOVID-2024-1.5v.docx](https://livejaverianaedu.sharepoint.com/:w:/s/ProyectoAGORA371/EfIcY6b7kBBDh4gHWpig49sBq6vmySRRfmdyK6cFdt2SEQ?e=Ad2bDI)

El presente documento ofrece una visión detallada de los principales hallazgos derivados del proceso de exploración inicial de los datos del sistema VACUNASCOVID. En esta fase inicial de análisis, se identificó la fuente de datos y se examinaron las variables contenidas en ella y sus características fundamentales.

La información denominada VACUNASCOVID, comprende información recopilada durante el periodo comprendido entre 2021 a 2023. Este periodo permite analizar una variedad de contextos y condiciones que resultan ser importantes para la comprensión y el análisis de las tendencias relacionadas.

Durante la fase inicial de exploración, se realizó un análisis exhaustivo de las variables contenidas en el conjunto de datos de la información suministrada de VACUNASCOVID. Se aplicaron técnicas estadísticas descriptivas básicas para examinar la distribución, la variabilidad y las tendencias presentes en los registros disponibles. Este análisis preliminar sirvió como punto de partida para comprender mejor la naturaleza de los datos y así poder identificar posibles áreas de interés y enfoque para análisis posteriores.

En estas secciones, se presenta la metodología de exploración de los datos, como los principales resultados obtenidos en esta fase inicial, incluyendo las características de las variables analizadas, las tendencias y patrones identificados de VACUNASCOVID.

Documento de instalación de Lago de Datos

[[](https://livejaverianaedu.sharepoint.com/:w:/s/ProyectoAGORA371/EXnbC_8YEm5GuHslH58PaMsB4FRxCEGakRm4TA41SFbnqw?e=S1qFIW)DISEÑO DE ARQUITECTURA Y CONFIGURACIÓN DEL LAGO DE DATOS - PROYECTO AGORA (1).docx](https://livejaverianaedu.sharepoint.com/:w:/s/ProyectoAGORA371/EXnbC_8YEm5GuHslH58PaMsB4FRxCEGakRm4TA41SFbnqw?e=S1qFIW)

En el contexto del proyecto "Desarrollo de Infraestructura de Big Data - Proyecto AGORA", este documento resume las actividades realizadas y los resultados obtenidos en relación con la implementación de un *data lake* para el proyecto.

La naturaleza dinámica de la disponibilidad de recursos informáticos ha llevado a la adopción de dos estrategias distintas, tal como se detalla en la solución previamente aprobada:

* Estrategia 1: Uso de infraestructura dedicada proporcionada por el Departamento de Ingeniería de Sistemas de la Pontificia Universidad Javeriana (PUJ).
* Estrategia 2: Uso de una infraestructura de nube pública proporcionada por la Alianza CAOBA.

Inicialmente, se llevó a cabo la evaluación de dimensiones con los recursos disponibles ofrecidos y las pruebas de instalación en un clúster Big Data utilizando los recursos informáticos ofrecidos por el Departamento de Sistemas y la Facultad de Ingeniería de la PUJ. Este enfoque fue posteriormente descartado debido a las interrupciones por mantenimiento durante los períodos vacacionales y a la virtualización de recursos, los cuales afectaban negativamente el rendimiento.

Posteriormente, en respuesta a la necesidad de llevar a cabo actividades exploratorias con las fuentes de datos Big Data del proyecto, la Alianza CAOBA proporcionó acceso a recursos en la nube pública de Azure. Esto permitió el despliegue parcial de las soluciones requeridas para el proyecto.

En una fase posterior del proyecto, se planteó y ejecutó dentro de la estrategia 1, una variante que consistió en la asignación de una máquina especializada en el centro de datos de Alto Rendimiento Computacional (HPC) ZINE. Este hardware, configurado como un nodo único sin virtualización, se distingue por su arquitectura de tipo *PowerPC*, la cual, aunque es menos común, ofrecía una capacidad sustancialmente mayor en términos de memoria RAM y procesamiento. El desafío residía en diseñar una arquitectura óptima y desarrollar un proceso de instalación que, a pesar de su complejidad, resultó ser más adecuado para satisfacer las exigentes necesidades computacionales del proyecto por la dimensión de sus datos.

A continuación se presenta un capítulo con cada una de las soluciones desplegadas del lago de datos para el proyecto AGORA.