**.Índice**

* **Tabla de contenido que facilite la navegación.**

**Introducción**

* **Propósito:** Describir el objetivo del proyecto de Big Data.
* **Alcance:** Definir qué datos se manejarán y qué problemas se resolverán.
* **Descripción del sistem**

s Sistema de limpieza y gestiona miento de grandes volúmenes de datos, desarrollaremos un proyecto de "Almacenes de Datos, Catálogo, ETL y Limpieza y Almacenamiento de Datos". Este proyecto tiene como objetivo fundamental garantizar que los datos que utilizamos sean precisos, accesibles y útiles para la toma de decisiones.

Durante este proceso, nos centraremos en implementar un sistema integral que no solo almacene datos, sino que también facilite su limpieza y organización. Con un enfoque sistemático en la extracción, transformación y carga (ETL), pretendo optimizar la calidad de nuestros datos, garantizando que estén listos para ser analizados y utilizados de manera efectiva.

Estoy convencido de que una gestión adecuada de los datos no solo mejorará nuestra eficiencia operativa, sino que también permitirá a la organización tomar decisiones más informadas y estratégicas. Si deseas hacer algún ajuste o agregar algo más a esta introducción, házmelo saber. Cuando estés listo, podemos avanzar al siguiente paso.

**2. Descripción del Proyecto**

**Alcance**

El proyecto se enfoca en la recolección, limpieza y almacenamiento de datos provenientes de diversas fuentes. Se abordarán problemas como datos incompletos, incorrectos y duplicados.

**Motivación**

Los datos erróneos pueden generar decisiones fiscales equivocadas y afectar la eficiencia operativa. Por ejemplo, en el análisis de censos poblacionales, es crucial contar con datos fiables para evitar errores en la planificación gubernamental

**4. Requisitos**

* **Requisitos Funcionales:** Especificaciones sobre qué debe hacer el sistema (por ejemplo, procesamiento en tiempo real, análisis de datos históricos).
* **Req**

**5. Arquitectura del Sistema**

**Infraestructura**

La arquitectura del sistema incluirá herramientas como Apache Spark para el procesamiento de datos y bases de datos NoSQL para el almacenamiento.

**6. Diseño del Sistema**

* **Estructura de Datos:** Esquemas de bases de datos y formatos de datos (JSON, Parquet, etc.).
* **Flujos de**

**Flujo de Datos**

Un diagrama de flujo debe ilustrar cómo los datos se recogen, limpian y almacenan, asegurando un proceso eficiente.

**4. Metodología**

**Recolección de Datos**

Los datos se recopilarán a través de APIs y bases de datos existentes.

**Limpieza de Datos**

El proceso incluirá:

* **Identificación**: Detección de errores mediante auditorías

[2](https://www.purestorage.com/es/knowledge/what-is-data-hygiene.html" \t "_blank)

[4](https://www.questionpro.com/blog/es/higiene-de-datos/" \t "_blank)

.

* **Corrección**: Modificación y eliminación de registros erróneos

[5](https://aws.amazon.com/es/what-is/data-cleansing/" \t "_blank)

.

* **Validación**: Asegurar que los datos cumplan con los estándares establecidos.

**5. Indicadores y Métricas**

**KPIs**

* Tasa de reducción de duplicados.
* Porcentaje de registros corregidos.
* Tiempo promedio para la limpieza de datos.

**6. Resultados Esperados**

Se anticipa que el sistema mejorará la calidad de los datos, lo que a su vez facilitará una toma de decisiones más informada y eficiente. La higiene adecuada de los datos fomentará la seguridad y el cumplimiento normativo

[2](https://www.purestorage.com/es/knowledge/what-is-data-hygiene.html" \t "_blank)

[3](https://www.purestorage.com/la/knowledge/what-is-data-hygiene.html" \t "_blank)

.

**7. Desafíos y Riesgos**

**Identificación de Riesgos**

* Inconsistencias en los datos debido a diferentes fuentes.
* Resistencia al cambio por parte del personal.

**Estrategias de Mitigación**

Implementar capacitación continua sobre la importancia de la limpieza y calidad de los datos.

**8. Consideraciones Éticas**

Es fundamental manejar adecuadamente los datos sensibles, asegurando su eliminación cuando ya no sean necesarios

[2](https://www.purestorage.com/es/knowledge/what-is-data-hygiene.html" \t "_blank)

.

**9. Conclusiones**

El "Sistema de Limpieza y Almacenamiento de Datos" es un paso crucial hacia la mejora continua en la calidad de los datos dentro de la organización, garantizando decisiones más precisas y efectivas.