Propuesta para la Implementación Data Archiving para Biblioweb SAS

Gloria Isabel Moya González, Liliana Rocío Moya González, Luz Yamile Martínez Alarcón y Romel Alfonso Vega Rincón

Especialización en Big Data, Fundación Universitaria Compensar

Proyecto fin de grado

Carlos Alberto Cañón Romero

Mayo de 2023

Propuesta implementación Data Archiving para Biblioweb SAS

Esta obra está bajo una Licencia Creative Commons Atribución-NoComercial-CompartirIgual 4.0 Internacional.

Gloria Isabel Moya González, Liliana Rocío Moya González, Luz Yamile Martínez Alarcón y Romel Alfonso Vega Rincón

Trabajo de grado presentado como requisito parcial para optar al título de:

Especialista en Big Data

Director:

(Ingeniero de Sistemas, Magister en Gerencia Tecnologías de la Información)

Carlos Alberto Cañón Romero

Línea de Investigación: Big Data

Fundación Universitaria Compensar

Facultad de Ingeniería, Especialización en Big Data

Bogotá, Colombia

2023

Contenido

Resumen 7

Abstract 8

Introducción 9

Antecedentes y Justificación 10

Organigrama 10

Portafolio de Servicios 10

1. Formulación del proyecto 13

1.1 Problema de Investigación 13

1.2 Objetivos 13

1.2.1. Objetivo General 13

1.2.2. Objetivos específicos 13

1.3 Alcances y Limitaciones 14

1.3.1. Alcance 14

1.3.2. Limitaciones 14

2. Marco Teórico 15

2.1 Gestión Documental 15

2.1.1. Procesos de la Gestión Documental 16

2.2 Documento de Archivo 16

2.3 Documento Electrónico 16

2.3.1. Mensaje de Datos 17

2.3.2. Intercambio Electrónico de Datos 17

2.4 Estructura del Documento Electrónico de Archivo 17

2.4.1. Características del Documento Electrónico de Archivo 18

2.4.2. Ciclo de Vida del Documento Electrónico 18
2.4.3. Documento Digitalizado 19
2.5 Ciclo de Vida del Expediente Electrónico 20
2.6 Ciclo De Vida Del Dato 21
2.7 Retención del Dato 24
2.7.1. Política de Retención de Datos 24
2.7.2. Beneficios de Retención de Datos 24
2.8 Data Archiving 25
3. Diseño Metodológico 26
3.1 Fase 1 Entrega de Muestra de Información 26
3.2 Fase 2 Proceso de Limpieza y Transformación de Datos 29
3.3 Fase 3 Estructuración de Políticas para el Manejo de la Información 30
3.4 Fase 3 Requerimientos Tecnológicos para la Implementación 30
3.4.1. Servidores 31
3.4.2. Motores de Bases de Datos 32
3.4.3. Herramientas de Limpieza de Datos 34
3.4.4. Paquete de OFFICE 35
3.4.4. Gastos Indirectos 36
4. Resultados 37
4.1 Fase Uno Exploración 37
4.2 Fase Dos 37
4.3 Fase Tres 37
4.4 Fase Cuatro 37
5. Conclusiones y Recomendaciones 39
5.1 Conclusiones 39

5.2 Recomendaciones 39

Referencias 40

LISTA DE FIGURAS

FIGURA 1 ORGANIGRAMA 10
FIGURA 2 PROCESO DE GESTIÓN DOCUMENTAL 16
FIGURA 3 ESTRUCTURA DEL DOCUMENTO ELECTRÓNICO DE ARCHIVO 17
FIGURA 4 CICLO DE VIDA DEL DOCUMENTO ELECTRÓNICO 19
FIGURA 5 FASES DEL EXPEDIENTE ELECTRÓNICO 20
FIGURA 6 CICLO DE VIDA DEL DATO 21
FIGURA 7 FORMATOS COMÚNMENTE UTILIZADOS 28
FIGURA 8 MODELO ENTIDAD-RELACIÓN 29

LISTA DE TABLAS

TABLA 1 CICLO DE VIDA DEL DATO 21
TABLA2 DATA ARCHIVING VS BACK UP 25
TABLA 3 RECOPILACIÓN MUESTRA INFORMACIÓN SUMINISTRADA POR BIBLIOWEB 26
TABLA 4 SERVIDORES 31
TABLA 5 MOTORES DE BASES DE DATOS 32
TABLA 6 HERRAMIENTAS DE LIMPIEZA DE DATOS 34
TABLA 7 PAQUETE DE OFFICE 35
TABLA 8 GASTOS INDIRECTOS 36

Resumen

La presente propuesta pretende mostrar a la Empresa Biblioweb SAS la implementación del Data Archiving en los procesos que éste maneja específicamente en lo relacionado con la elaboración de las ofertas económicas que presenta para los procesos de contratación, tanto con entidades públicas como privadas. En la actualidad la empresa ha presentado dificultades al momento de realizar sus propuestas ya que no cuenta con la información centralizada, estructurada y categorizada; que necesita para la elaboración de las distintas ofertas según el mercado al que están dirigidas. Por lo anterior se propone la implementación de Data Archiving con el fin de centralizar y categorizar la información que requiere para la elaboración de las ofertas.

Palabras clave: Data Archiving, categorización, información, Archivos

Abstract

This proposal aims to show the company Biblioweb SAS the implementation of Data Archiving in the processes it manages, specifically in relation to the preparation of economic offers presented for contracting processes, both with public and private entities. At present, the company has presented difficulties at the moment of making its proposals since it does not have the centralized, structured and categorized information required for the elaboration of the different offers according to the market to which they are directed. Therefore, the implementation of Data Archiving is proposed in order to centralize and categorize the information required for the elaboration of the required offers.

Keywords: (Data Archiving, categorization, information, Archives).

Introducción

Se estima que el volumen de datos existente en la actualidad está por encima del zettabyte3 y que crecerá de forma exponencial en el futuro (Casas, et al. 2019). Como se evidencia en lo expuesto anteriormente, las organizaciones están generando gran cantidad de información y en la mayoría de los casos, éstas no cuentan con el espacio suficiente para el almacenamiento derivado de su crecimiento exponencial y así como existen procesos de organización, trámite y disposición para la información que se encuentra en soporte físico, también hay procesos que se aplican a la que se genera en entornos electrónicos y/o digitales.

Las empresas pueden aprovechar el poder del big data, la analítica y el Internet de las cosas (IoT) para mejorar su rendimiento y rentabilidad (Marr, 2017) Las organizaciones han tomado conciencia de la importancia que tiene el manejo adecuado de la información, ya que el volumen de datos tiende a aumentar cada vez más, y un mal manejo conllevaría a pérdidas de información, análisis inadecuado y dificultades para acceder a la misma, entre otros.

El presente proyecto tiene como fin proponer el uso de la data archiving a la empresa Biblioweb.

El uso de data archiving juega un papel importante en el manejo de la información ya que facilita la accesibilidad, seguridad, agilidad y escalabilidad; permitiendo hallar nuevos modos de monetizar la información archivada, promover la innovación, mejorar el servicio al cliente y optimizar la calidad de las relaciones con los socios, empleados y clientes (PowerData Solutions S.L., 2023).

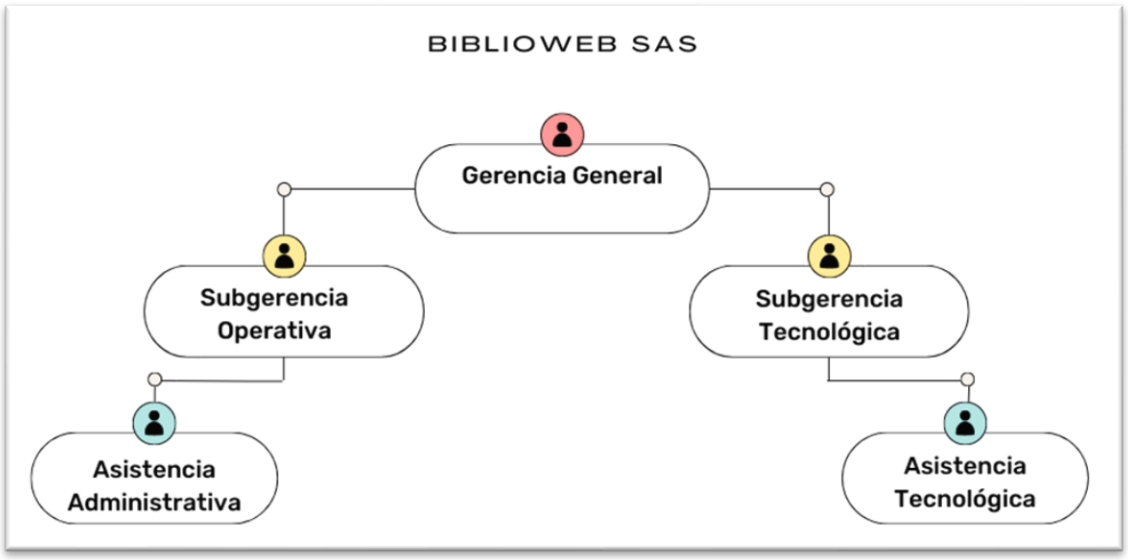
Antecedentes y Justificación

Biblioweb SAS es una compañía que lleva 17 años en el mercado, brindando soluciones e innovación a sistemas de información de Bibliotecas, Archivos, Sistemas Gerenciales entre otros. Los cuales están segmentados en tres grupos (E.G. Delgado, comunicación personal, abril 5 de 2023)

Organigrama

Figura 1

Organigrama



Nota: Elaboración propia con información (E.G. Delgado, comunicación personal, abril 5 de 2023)

Portafolio de Servicios

Bibliotecas

• Administración: A través de convenios institucionales con entidades públicas

• Insumos: Suministro de productos personalizados para gestión de Bibliotecas o centros de documentación

• OJS (OPEN JOURNAL SYSTEM): Sistema libre que permite la gestión, edición y publicación de revistas.

• DSPACE para repositorios digitales

• KOHA para bibliotecas (Hosting y Módulo de Informes Avanzado)

Gestión documental

• Saneamiento y verificación de archivos: Aplicado tanto a depósitos como a estanterías cumpliendo con la normatividad establecida.

• Digitalización de Archivos: Servicio de digitalización y además venta y alquiler de equipos.

• Insumos para Archivos: Suministro de todo tipo de material para la implementación de proyectos de gestión documental (cajas para archivo, carpetas 4 aletas entre otros).

• Consultoría Instrumentos Archivísticos: Elaboración e implementación de éstos acorde con la normatividad.

• Custodia de Archivos: Cuenta con bodegas o depósitos especializados cumpliendo con requerimientos técnicos exigidos.

• Sistemas de Medición y control de condiciones ambientales: Aplicados a instalaciones de almacenamiento de archivo y Bibliotecas. Éstos permiten realizar la medición de condiciones como Temperatura, humedad, calidad del aire, intensidad de luz, radiación UV, material particulado

Según la entrevista realizada al representante legal de Biblioweb SAS Elkin Giovanny Delgado nos expone que para su organización la mayor problemática presentada es no contar con la centralización de los datos (E.G. Delgado, comunicación personal, abril 5 de 2023).

Por lo anterior se propone a Biblioweb SAS la implementación de Data Archiving con el fin de centralizar y categorizar la información que requiere para la elaboración de las ofertas requeridas.

1. Formulación del proyecto

1.1 Problema de Investigación

Biblioweb SAS cuenta actualmente con un volumen de información de 475 Gigas alojadas en diferentes fuentes (PCs, Drives, Correos electrónicos, USB, etc.) la cual no se encuentra organizada, ni estructurada por tanto no permite la accesibilidad de una manera inmediata, lo que genera una mayor disposición de tiempo al momento de elaborar las propuestas técnicas y económicas para presentar en los distintos procesos de contratación (E.G. Delgado, comunicación personal, abril 5 de 2023).

Objetivos

1.2.1. Objetivo General

Generar una propuesta de implementación del proceso Data Archiving que permita la categorización de los datos de Biblioweb SAS en la ciudad de Bogotá.

1.2.2. Objetivos específicos

- Diagnosticar la situación actual de los datos de la empresa Biblioweb SAS.
- Identificar las necesidades y requisitos para implementación del método de Data Archiving y evaluar las tecnologías disponibles con el fin de establecer cuál se requiere implementar en Biblioweb SAS.

1.2 Alcances y Limitaciones

1.3.1. Alcance

La propuesta está diseñada para ser aplicada en la sede principal de Biblioweb SAS la cual está ubicada en la ciudad de Bogotá para ser utilizada por los colaboradores directos de organización que intervienen en los procesos de cotizaciones, atención a clientes y facturación tanto de proveedores como de clientes.

Esta propuesta de data archiving está enfocada a las 475 gigas de información suministrada por Biblioweb SAS al 30 de abril de 2023 y aplica solamente para archivos digitales.

1.3.2. Limitaciones

No contempla la información generada después del 30 de abril de 2023.

2. Marco Teórico

Una vez identificado el problema que presenta Biblioweb SAS, se hace necesario aproximarse a aquellos conceptos relacionados tanto con el objeto de la organización como con la presente propuesta.

Conceptos como Gestión Documental, Documento de Archivo, Documento electrónico, Data Archiving, metadatos entre otros serán definidos a continuación:

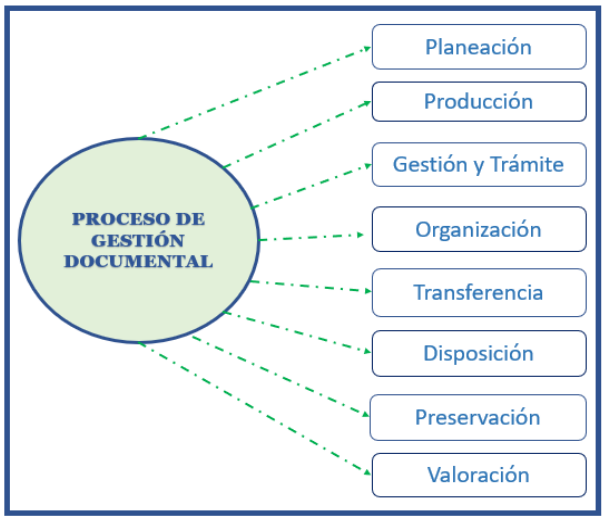
2.1 Gestión Documental

La Gestión Documental se define como “Conjunto de actividades administrativas y técnicas tendientes a la planificación, procesamiento, manejo y organización de la documentación producida y recibida por las entidades, desde su origen hasta su destino final, el objeto de facilitar su utilización y conservación” (Archivo General de la Nación Colombia, 2006, p. 11)

2.1.1. Procesos de la Gestión Documental

La gestión documental y sus diferentes etapas lo podemos evidenciar en la Figura 2.

Figura 2
Proceso de Gestión Documental



***Nota**: Elaboración propia con información tomada de Guía para la Gestión de Documentos y Expedientes Electrónicos (p.56), por Ministerio de Tecnologías de la Información y las Comunicaciones, 2019.*

2.2 Documento de Archivo

Según el Archivo General de la Nación (2006) un documento de archivo es "un registro de información producida o recibida por una entidad pública o privada debido a sus actividades o funciones" (p.8).

2.3 Documento Electrónico

Según el Archivo General de la Nación (2006) se establece como "Información generada, enviada, recibida, almacenada y comunicada por medios electrónicos, ópticos o similares" (p.8).

2.3.1. Mensaje de Datos

Según el Ministerio de Tecnologías de la Información y las Comunicaciones (2019) “Es la información generada, enviada, recibida, almacenada, comunicada por medios electrónicos, ópticos o similares” (p.13).

2.3.2. Intercambio Electrónico de Datos

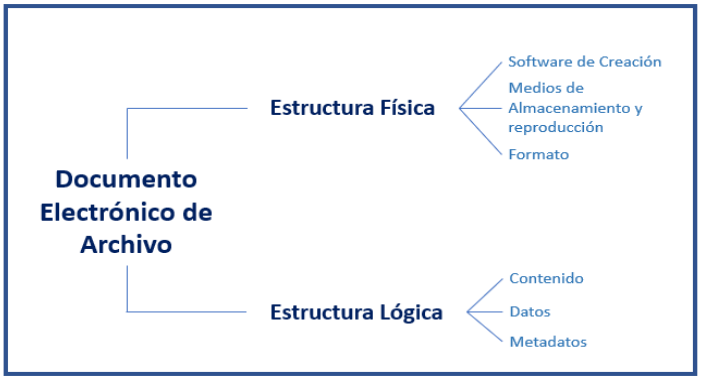
Según el Ministerio de Tecnologías de la Información y las Comunicaciones (2019) “Es la transmisión electrónica de datos de una computadora a otra, la cual está estructurada bajo normas técnicas convenidas al efecto” (p.14).

2.4 Estructura del Documento Electrónico de Archivo

El documento electrónico de archivo está definido por una estructura física y una lógica, en la Figura 3 se puede detallar sus elementos.

Figura 3

Estructura del Documento Electrónico de Archivo



***Nota**: Elaboración propia con información tomada de la Guía para la Gestión de Documentos y Expedientes Electrónicos (p. 17), MINTIC, 2019*

2.4.1. Características del Documento Electrónico de Archivo

Las características del documento electrónico están definidas por:

- Autenticidad: Puede demostrar que es lo que afirma ser, que ha sido creado o enviado por la persona que afirma haberlo creado o enviado.
- Fiabilidad: Refleja de manera exacta y completa la ejecución de actividades u operaciones.
- Integridad: Se encuentra completo y sin alteraciones.
- Disponibilidad: Se puede localizar, recuperar presentar e interpretar.
- Contenido estable: Contenido de un objeto digital que sólo puede sufrir cambios obedeciendo a un conjunto de reglas preestablecidas.
- Forma Documental Fija: Asegura que su contenido permanece completo y sin alteraciones, a lo largo del tiempo, manteniendo la forma original.
- Vínculo archivístico: Relación que vincula a cada documento con el anterior y con el siguiente, de acuerdo con el trámite o asunto, conformando así un expediente.
-

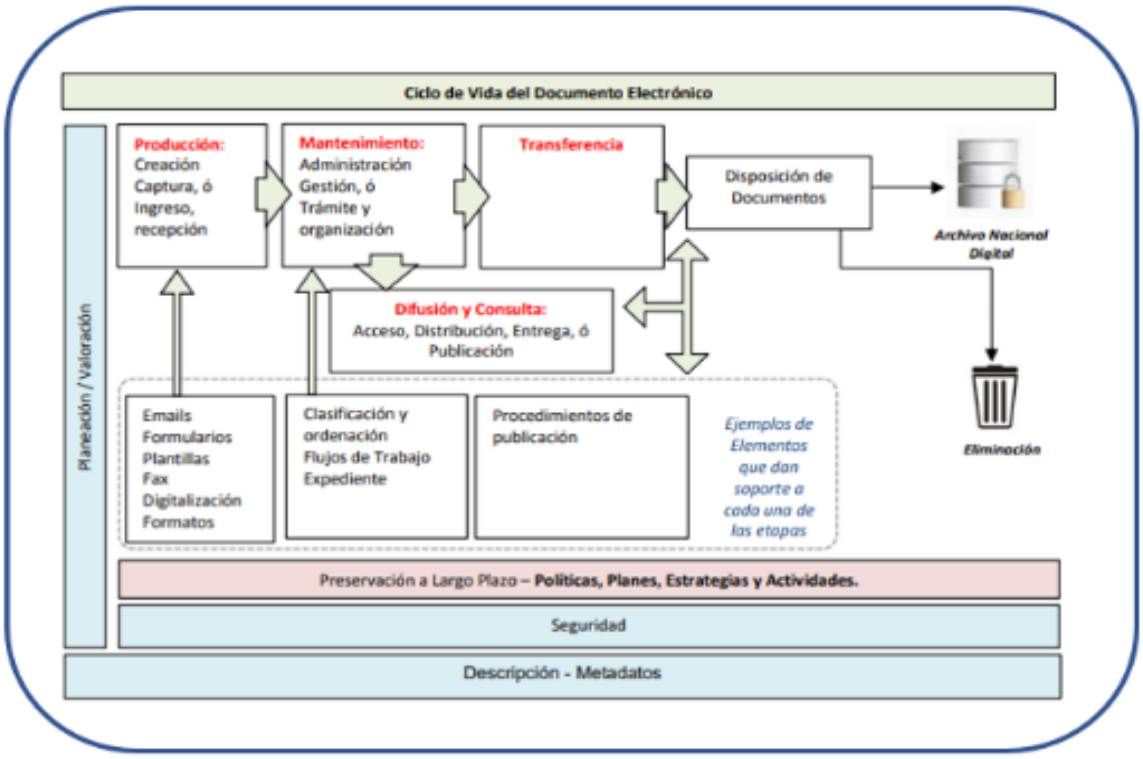
2.4.2. Ciclo de Vida del Documento Electrónico

Según el artículo 2.8.2.5.7 del Decreto 1080 de 2015, para asegurar una adecuada gestión documental se deben tener en cuenta las siguientes etapas: Producción – Mantenimiento – Difusión - Administración (MINTIC, 2019):

En la Figura 4 se detallan las actividades que se generan en cada etapa del ciclo.

Figura 4

Ciclo de Vida del Documento Electrónico



***Nota**: Etapas del ciclo de vida del documento electrónico y los procesos, imagen tomada de Guía para la gestión de documentos y expedientes electrónicos (p.60) por Ministerio de Tecnologías de la Información y las Comunicaciones, 2019.*

2.4.3. Documento Digitalizado

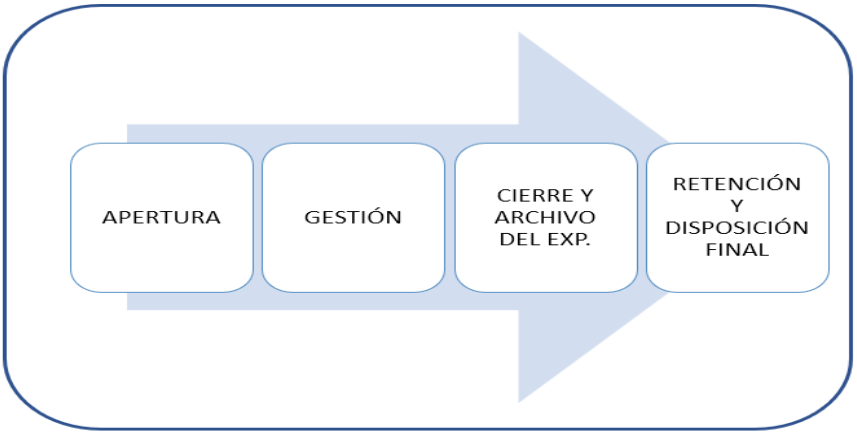
Según el Ministerio de Tecnologías de la Información y las Comunicaciones (2019) se considera que “es una representación digital, obtenida a partir de un documento registrado en un medio o soporte físico, mediante un proceso de digitalización.” (p.13).

2.5 Ciclo de Vida del Expediente Electrónico

Está conformado por las distintas fases por la que atraviesa un expediente, desde la conformación de los documentos electrónicos de archivo que lo integran hasta su disposición final como se puede observar en la Figura 5.

Figura 5

Fases del Expediente Electrónico



***Nota**: Etapas del ciclo de vida del documento electrónico y los procesos, imagen tomada de Guía para la gestión de documentos y expedientes electrónicos (p.76) por Ministerio de Tecnologías de la Información y las Comunicaciones, 2019.*

2.6 Ciclo De Vida Del Dato

Figura 6

Ciclo de Vida del Dato



***Nota**: Elaboración propia con información tomada del Ciclo de vida del dato (p.4) por Ministerio de Tecnologías y las Comunicaciones, 2021.*

Tabla 1

Ciclo de Vida del Dato

Fase Ciclo de Vida	Descripción
Crear y obtener	Los datos pueden provenir de diferentes fuentes, estructuradas, no estructuradas o fuentes secundarias como las redes sociales; asimismo también son diversos los formatos en que pueden venir (Ej. pdf, jpg, docx, xml, txt, json, csv, png, entre muchos otros). Los datos suelen ser creados por las organizaciones usualmente de las siguientes formas: · Adquisición de datos: adquisición de datos ya existentes que se han producido fuera de la organización. · Entrada de datos: entrada manual de nuevos datos por parte del personal dentro de la organización. · Captura de datos: captura de datos generados por dispositivos utilizados en varios procesos de la organización. El modelado y diseño de datos es de suma importancia en esta fase, pues determina que información es la que se determina utilizar.

Procesar	La etapa de procesamiento incluye las siguientes tareas: · Limpieza de datos: en la que un conjunto de datos se limpia y se transforma de su forma sin procesar a algo más accesible y utilizable. Esto también se conoce corrección de datos. · Compresión de datos: en la que los datos se transforman en un formato que se puede almacenar de manera más eficiente. · Cifrado de datos: en el que los datos se traducen a otra forma de código para protegerlos de problemas de privacidad. · Calidad de los datos: en los proceso de obtención y procesamiento de datos, donde principalmente se aplican las políticas y controles, establecidas en el diseño, con el propósito de determinar cuáles datos son utilizables. Además de las tareas de esta etapa mencionadas anteriormente existen muchas otras. incluso el simple hecho de tomar un formulario impreso y digitalizarlo puede considerarse una forma de procesamiento de datos.

Fase Ciclo de Vida	Descripción
Almacenar	El almacenamiento y operaciones incluyen el diseño, la implementación y el soporte de los datos almacenados para maximizar su valor, utilizando tanto formatos como repositorios que busquen el equilibrio entre disponibilidad y coste de almacenamiento, según los distintos escenarios de consulta, por ejemplo, diferenciando datos de alta demanda o disponibilidad de otros que no se consultan frecuentemente. Comúnmente el almacenamiento se realiza mediante la creación de bases de datos o conjuntos de datos. Estos conjuntos de datos pueden almacenarse en la nube, en servidores con premise o utilizando otras formas de almacenamiento físico como discos duros magnéticos o de estado sólido, memorias o cintas magnéticas, entre otros.
Transferir y compartir	A medida que las organizaciones requieren de datos de fuentes secundarias, generados por otros actores del ecosistema de datos, útiles para la toma de decisiones, la planificación, la optimización de operaciones, entre otros, existe una mayor presión para compartir también datos generados. Los datos compartidos pueden ayudar a mejorar las decisiones, ya que las organizaciones pueden obtener una vista más completa de los impactos que sus decisiones han tenido en base a las contribuciones de nuevos conjuntos de datos de una variedad más amplia de fuentes, tanto internas como externas.

	Esta capacidad de compartición de datos debe estar soportada en una capa de interoperabilidad y haciendo uso de un estándar de lenguaje común de intercambio.
Analizar y usar	Durante la fase de uso del ciclo de vida de los datos, los datos se utilizan para respaldar las actividades de la organización. Los datos se pueden ver, procesar, modificar y guardar. Se debe mantener un registro de auditoría para todos los datos críticos, con el propósito de garantizar que todas las modificaciones que se realicen a los datos sean completamente rastreables y auditables. Los datos también pueden estar disponibles para exponer a otros actores del ecosistema de datos que se encuentren fuera de la organización. El análisis de datos se refiere a procesos que intentan obtener información significativa a partir de datos sin procesar. Los analistas y científicos de datos utilizan diferentes herramientas y estrategias para realizar estos análisis. Algunos de los métodos más utilizados incluyen modelado estadístico, algoritmos, inteligencia artificial, minería de datos y aprendizaje automático.

Fase Ciclo de Vida	Descripción
Archivar y preservar	El archivado de datos hace referencia a la copia de datos en un entorno donde se almacenan en caso de que se necesiten nuevamente en un ambiente de producción, así mismo también incluye la eliminación de estos datos de todos los entornos de producción activos. Un archivo de datos es simplemente un lugar donde se almacenan los datos, pero donde no se realiza mantenimiento o uso general. Si es necesario, los datos se pueden restaurar a un entorno en el que se puedan utilizar. Dado que el volumen de datos archivados crece inevitablemente, si bien es posible que desee guardar todos los datos de manera indefinida, los costos de almacenamiento pueden incentivan la destrucción de los datos que ya no se requieren. Por otro lado, es posible que se disponga de datos de uso limitado a una ventana de tiempo o a hasta /durante la ocurrencia de un suceso. La destrucción o depuración de datos es la eliminación de cada copia de un elemento de datos de una organización.

Reutilizar	Reutilizar significa usar datos que originalmente fueron recopilados para otro fin o propósito. La reutilización de datos también se puede llamar análisis secundario.

Nota: *Tabla tomada de Ciclo de vida del dato (p.4-6) por MINTIC, 2021.*

2.7 Retención del Dato

La retención de datos es el período de tiempo durante el cual los datos se mantienen en un sistema de almacenamiento de datos. El archivado de datos se utiliza comúnmente para conservar los datos más allá del período de retención normal para su uso futuro.

2.7.1. Política de Retención de Datos

La política de retención de datos o registros, son todos aquellos protocolos que se tienen establecidos en una organización para archivar la información según las necesidades requeridas. Esta determina como organizar la misma con el fin de establecer los tiempos de acceso o eliminación de esta.

2.7.2. Beneficios de Retención de Datos

Algunos de los beneficios que tiene la retención de datos son:

Organización y almacenamiento adecuado: cada organización genera distintos tipos de información por tanto la adecuada retención de datos ayuda al correcto almacenamiento y establecimiento del tiempo útil para los diferentes tipos de datos.

Costos de almacenamiento reducido: en la actualidad los costos asociados con el almacenamiento de información son muy altos, y la reducción del volumen de la información archivada, también reduce el costo de almacenamiento de estos.

Relevancia de los datos existentes: con la política de retención de datos se logra que la información menos relevante se elimine cuando está ya ha cumplido su ciclo.

2.8 Data Archiving

Recolección de datos antiguos y su traslado a una ubicación segura, con fines de investigación posterior (PowerData Solutions S.L., 2023)

Se diferencia de Back Up o copias de seguridad, el data archiving permite el traslado de los datos con el fin de liberar recursos de almacenamiento. Es útil para organizaciones con espacio limitado de almacenamiento. Al hacer uso de archivado de datos para mensajes de correo electrónico, archivos y registros de bases de datos, las organizaciones pueden liberar espacio sin exponerse a infringir normas o a pérdida de información valiosa.

Tabla2

Data Archiving Vs Back up

Data Archiving	Back up
• Largo Plazo • Datos originales • Almacenamiento orientado al análisis y mejora de la capacidad de respuesta del negocio.	• Corto plazo • Copia de datos originales • Recuperación de desastres informáticos

Nota**:** *Elaboración propia con información tomada del E-book Data Archiving, por PowerData Solutions S.L., 2023.*

Según datos de IDC (International Data Corporación) 2 de cada 3 empresas mantienen seis o más archivos, mediante los que almacenan información muy variada y de formato heterogéneo, lo que complica la recuperación de esta (PowerData Solutions S.L., 2023)

3. Diseño Metodológico

Se conoce la investigación cualitativa como los estudios sobre el quehacer cotidiano de las personas o de grupos pequeños. Este tipo de investigación se interesa por lo que la gente dice, piensa, siente o hace; los patrones culturales que lo caracterizan, así como el proceso y significado de sus relaciones interpersonales y con el medio. A partir de esto, describir o generar teoría a partir de los datos obtenidos (Lerma, 2009).

Para la presente propuesta el instrumento utilizado para la recolección de la información fue la Entrevista la cual nos permitió identificar el problema que presenta la empresa Biblioweb.

Según Lerma la Entrevista es la interacción que hay entre la persona que recolecta la investigación y el entrevistado. Como no se utilizó guía ni formulario se denomina entrevista no estructurada, ya que se realizó de manera virtual en dónde el Representante Legal de Biblioweb nos describe la situación actual de la compañía.

Dentro del desarrollo de la propuesta se tuvieron en cuenta las siguientes fases:

3.1 Fase 1 Entrega de Muestra de Información

La información entregada por Biblioweb SAS estas contenida de la siguiente manera:

Tabla 3
Recopilación muestra información suministrada por Biblioweb

Tipo	Cantidad en Gb
Cuentas Correos Electrónicos	330,252
Equipos Locales	25
Discos Duros	120
Total Información Recibida en GB	475,252

Nota**:** *Elaboración propia con información tomada (E.G. Delgado, comunicación personal, abril 5 de 2023)*

En esta Fase se procede a identificar el tipo de información producida y/o recibida por Biblioweb de forma digital, independientemente del formato. Dentro de los cuales encontramos:

- Correos Electrónicos
- Documentos textuales
- Documentos fotográficos

Basados en la Guía Técnica de Gestión de Documentos y Expedientes Electrónicos de MINTIC y AGN, se procede a realizar la verificación de los formatos en los cuales se está produciendo y/o recibiendo la información, esta tarea requiere de los siguientes pasos:

- Identificar de acuerdo con el tipo de documento electrónico, cuáles son los formatos en los que se está produciendo y/o recibiendo.
- Identificar si los formatos permiten al documento electrónico ser auténtico, íntegro, fiable y usable durante todo el ciclo de vida.

En el caso en que los formatos no sean los adecuados Biblioweb debería:

- Normalizar los formatos de documentos electrónicos
- Utilizar formatos que garanticen la preservación a largo plazo.
- Utilizar formatos abiertos que garanticen la recuperación, lectura e interoperabilidad del documento electrónico a lo largo del tiempo.

- Identificar la información mínima que se requiere para incorporar el documento electrónico producido a las herramientas informáticas de la organización

Para realizar esta identificación se tomó como base el anexo 1 de la Guía Técnica de Gestión de Documentos y Expedientes Electrónicos de MINTIC y AGN.

Figura 7

Formatos comúnmente utilizados

ANEXO 1. FORMATOS DE ARCHIVO DE USO COMÚN[55]

En la siguiente tabla, se presentan algunos formatos comúnmente utilizados para la elaboración de documentos electrónicos de archivo:

(D = Difusión) (P = Preservación) Contenido de Texto

FORMATO	CARACTERÍSTICAS	EXTENSIÓN	ESTANDAR	D	P
Texto Plano	Formato que contiene únicamente texto formado solo por caracteres que son legibles, sin ningún tipo de formato tipográfico	.txt	ISO/IEC 646	✓	
		.csv	RFC 4180	✓	
Office OpenXML	Formato de procesador de textos basado en lenguaje XML abierto de Microsoft Office	.docx	ISO/IEC 29500 ECMA-376	✓	
		.xlsx		✓	
		.pptx		✓	
Open Document	Formato de Documento Abierto para Aplicaciones Ofimáticas, basado en XML y diseñado para soportar la edición.	.odt	ISO/IEC 26300 OASIS 1.2	✓	
		.ods		✓	
		.odp		✓	

***Nota**: Formatos de archivo de uso común, imagen tomada de Guía Técnica de Gestión de Documentos y Expedientes Electrónicos (p.82) por Ministerio de Tecnologías de la Información y las Comunicaciones, 2019.*

Se acuerda con el Gerente de Biblioweb el suministro de una muestra de datos de 50 gigas para análisis, obtenida de los equipos locales a la cual se le aplicara la Política de

Protección de Datos Ley 1581 de 17 de octubre 2012 (Congreso de la Republica de Colombia, 2012)

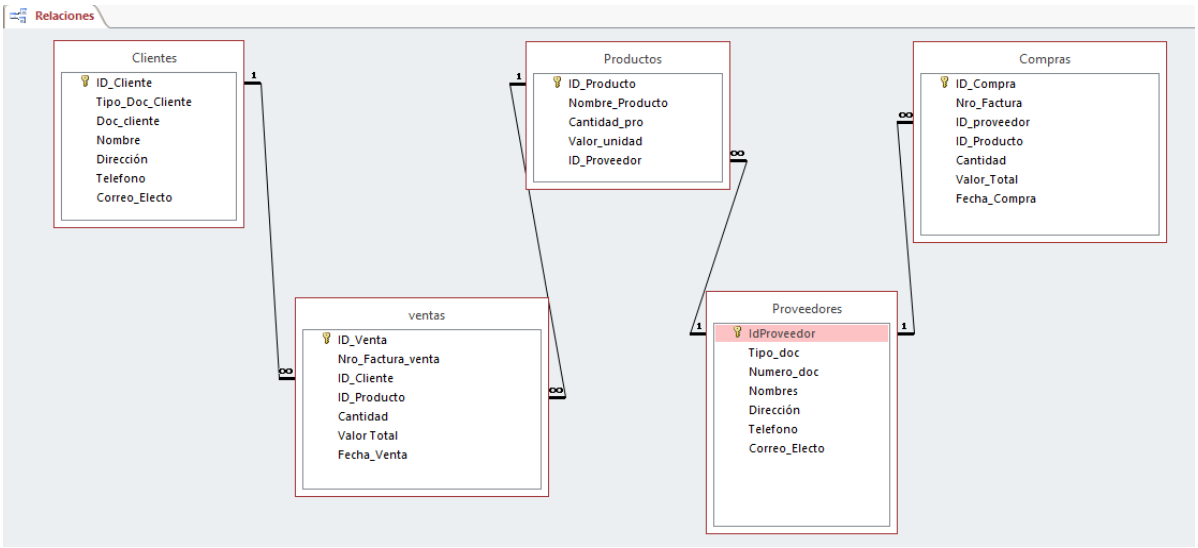
3.2 Fase 2 Proceso de Limpieza y Transformación de Datos

Con los datos entregados por Biblioweb se implementa un modelo de entidad-relación, para establecer una estructura de la información y las relaciones entre los datos para su normalización.

Se debe establecer en conjunto con Biblioweb el diseño de las entidades, los atributos y relaciones para organizar los datos, identificando las reglas y restricciones necesarias para garantizar la integridad de los datos y evitar inconsistencias.

El modelo entidad relación permite identificar relaciones entre entidades, esto es necesario para comprender cómo se relacionan los datos entre sí. Al definir claramente las relaciones entre las entidades, puede mejorar la eficiencia de las consultas y búsquedas configurando índices y otras estructuras de almacenamiento, escalabilidad y mantenibilidad. (Gutiérrez, 2013)

Figura 8
Modelo Entidad-Relación



***Nota**: Modelo de entidad-relación autoría propia realizado en Gestor de Bases de Datos de Microsoft Access.*

3.3 Fase 3 Estructuración de Políticas para el Manejo de la Información

Partiendo de la información relacionada en la Tabla No 1, en donde indica que "el archivado de datos hace referencia a la copia de datos en un entorno donde se almacenan en caso de que se necesiten nuevamente en un ambiente de producción, así mismo también incluye la eliminación de estos datos de todos los entornos de producción activos". Se procede a verificar junto con el Representante Legal de Biblioweb, la información suministrada con el fin de evaluar cuál debe mantenerse en la etapa de consulta y trámite, y cuál debe ser almacenada dada su baja demanda de consulta y cuál debe eliminarse; también establecer cuánto tiempo se deben retener los datos en cada etapa y que personas tiene acceso a esa información.

Es necesario aclarar a Biblioweb que Data Archiving no es lo mismo que Copia de seguridad tal como se indica en la Tabla No 2.

Las copias de seguridad se emplean para la recuperación de desastres, mientras que data archiving se orienta hacia la mejora de la capacidad de consulta. El backup no resulta suficiente para cubrir las necesidades organizacionales de data archiving ya que esta función requiere de: Capacidades de indexación y búsqueda- gestión del ciclo de vida del dato – garantía de que la información pueda ser fácilmente localizada y recuperada – prevención de duplicidades – gestión de metadatos y protección de la información sensible. (PowerData Solutions S.L., 2023)

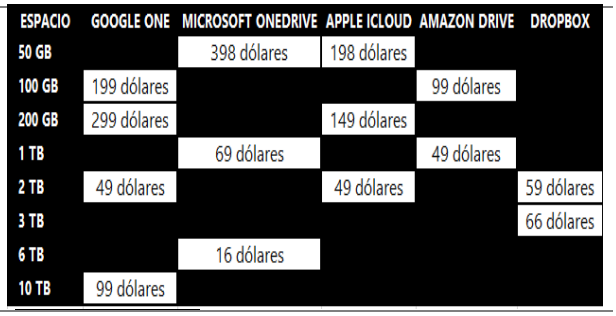
3.4 Fase 3 Requerimientos Tecnológicos para la Implementación

A continuación, se relaciona el presupuesto estimado para dar cumplimiento al diagnóstico de la empresa Biblioweb SAS, el cual tiene como objetivo establecer un plan económico para la ejecución del proyecto y asegurar la viabilidad de este.

Los gastos directos relacionados incluyen la adquisición de herramientas y software para la unificación, limpieza y almacenamiento de datos, el cual es esencial para llevar a cabo la labor de Data Archiving durante el tiempo trascurrido desde la creación de la empresa. Por lo anterior, se relacionan las herramientas y software seleccionados para la ejecución del objetivo principal.

3.4.1. Servidores

Tabla 4
Servidores

SERVIDOR	DEFINICION	CARACTERISTICA	VALOR
Servidores NAS (Network Attached Storage)	Son dispositivos de almacenamiento conectados a una red local que proporcionan almacenamiento centralizado y compartido para una empresa. Los servidores NAS son fáciles de usar y configurar, y ofrecen una buena relación costo-beneficio para empresas de tamaño pequeño o mediano.	Marca: Western Digital Color: Negro Dimensiones del artículo LxWxH 9.13 x 6.69 x 7.56 pulgadas Talla: 40TB	US$ 2599.99
Servidores en la nube (Cloud Storage)	Son servidores alojados en la nube que proporcionan almacenamiento de datos remoto. Los servidores en la nube son una opción conveniente y escalable para empresas que necesitan almacenamiento de datos accesible desde cualquier lugar.		

Nota*: *Elaboración propia con información tomada de la página de web por Amazon, 2023. *https://www.amazon.com/-/es/PR2100-Professional-Network-Attached-Storage/dp/B071GR5VK6?th=1*

Es importante tener en cuenta que la elección del servidor de almacenamiento de datos se tomó de factores como la cantidad de datos que se deben almacenar, los requisitos de rendimiento y la escalabilidad necesaria.

3.4.2. Motores de Bases de Datos

Los motores de bases de datos se relacionan por la calidad de los datos a limpiar y almacenar, que cumplen con las condiciones de datos estructurados y no estructurados.

Tabla 5

Motores de Bases de Datos

MOTOR	DEFINICION	VALOR
MySQL	Es un sistema de gestión de bases de datos relacionales de código abierto muy popular. Es fácil de usar y se puede utilizar en una amplia variedad de aplicaciones, desde pequeñas aplicaciones web hasta grandes sistemas empresariales.	US$ 0.00
Microsoft SQL Server	Es un sistema de gestión de bases de datos relacionales desarrollado por Microsoft. Ofrece un alto rendimiento y escalabilidad, y es compatible con una amplia variedad de aplicaciones empresariales.	SQL Server Enterprise US$ 5434.00 al año SQL Server Standard US$ 1418.00 al año
Oracle Database	Es un sistema de gestión de bases de datos relacionales de alto rendimiento y escalable. Es muy popular en entornos empresariales y ofrece una amplia gama de funciones y herramientas.	DB Enterprise Edition US$ 0.4301 por hora DB Enterprise High Performance US$ 0.8871 por hora DB Enterprise Edition Extreme Performance US$ 1.3441 por hora
Microsoft SQL Server	Es un sistema de gestión de bases de datos relacionales desarrollado por Microsoft. Ofrece un alto rendimiento y escalabilidad, y es compatible con una amplia variedad de aplicaciones empresariales.	SQL Server Enterprise US$ 5434.00 al año SQL Server Standard US$ 1418.00 al año

PostgreSQL	Es un sistema de gestión de bases de datos relacionales de código abierto que se centra en la extensibilidad y la conformidad con los estándares. Ofrece una amplia variedad de características y herramientas avanzadas para administrar grandes conjuntos de datos.	US$ 0.00
MongoDB	Es un sistema de gestión de bases de datos NoSQL de código abierto que se centra en la escalabilidad y la flexibilidad. Es muy popular en aplicaciones web y móviles que requieren un alto rendimiento y una gran capacidad de escalado.	US$ 0.00

Nota**:** *Elaboración propia con información tomada de las páginas web por Capterra, 2023. https://www.capterra.co/software/48791/postgresql y MongoDB, 2023. https://www.mongodb.com/es/what-is-mongodb*

3.4.3. Herramientas de Limpieza de Datos

Se sugiere el uso de OpenRefine ya que proporciona una amplia gama de funciones y operaciones para limpiar y normalizar datos. Puedes detectar y corregir errores de formato, eliminar duplicados, estandarizar valores y realizar otras transformaciones para asegurarte de que tus datos estén limpios y consistentes. estructurados.

Tabla 6

Herramientas de Limpieza de Datos

HERRAMIENTA	DEFINICION	VALOR
OpenRefine	Es una herramienta gratuita y de código abierto que se utiliza para la limpieza y transformación de datos. Permite realizar tareas como la eliminación de duplicados, la corrección de errores de ortografía, la normalización de valores, la extracción de valores de texto y la división de valores. OpenRefine también es compatible con varios formatos de archivo, como CSV, TSV, Excel y JSON.	US$ 0.00
Dataprep de Trifacta	Es una plataforma de preparación de datos que utiliza inteligencia artificial y aprendizaje automático para ayudar a los usuarios a limpiar y transformar sus datos. Ofrece una interfaz visual que permite a los usuarios ver los datos y aplicar transformaciones mediante el arrastre y suelta. Trifacta también permite la integración con varias herramientas de BI y Big Data, como Tableau y Hadoop.	Starter Edition US$ 80.00 al año tarifa por uso Professional Edition US$ 833.33 al año tarifa por uso
Talend Data Preparation	Es una herramienta de preparación de datos que permite a los usuarios limpiar, transformar y enriquecer sus datos. Ofrece una interfaz visual y un conjunto de transformaciones predefinidas que facilitan la limpieza de datos. Talend Data Preparation también es compatible con varios formatos de archivo y bases de datos, como Excel, CSV, MySQL y Oracle.	Los precios son informados mediante asesores dependiendo la necesidad

Nota*: *Elaboración propia con información tomada de la página web de Refine, 2023.
https://escueladedatos.online/openrefine/

3.4.4. Paquete de OFFICE

Se incluye licenciamiento con Microsoft Office ya que ofrece varias herramientas para el manejo de datos, adicional que permite muy buena compatibilidad con las herramientas seleccionadas.

Tabla 7

Paquete de OFFICE

Herramienta	Definición	Valor
Microsoft 365	Es un conjunto completo de aplicaciones y servicios en la nube que permite a los usuarios crear, colaborar y comunicarse de manera efectiva, brindando herramientas de productividad esenciales y servicios de almacenamiento y comunicación integrados en una plataforma unificada.	US$ 12.50 mes

***Nota*:** *Elaboración propia con información tomada de la página web de Microsoft, 2023. https://www.microsoft.com/es-co/microsoft-365/business/compare-all-microsoft-365-business-products*

Se sugiere esta herramienta ya que brinda a los usuarios acceso a aplicaciones web y de escritorio y almacenamiento en la nube, lo que les permite crear, editar y compartir documentos, hojas de cálculo, presentaciones y más desde cualquier dispositivo con conexión a Internet. Los usuarios pueden colaborar con otros en tiempo real, hacer videollamadas, enviar correos electrónicos, administrar calendarios, y utilizar otras herramientas integradas de comunicación y colaboración.

3.4.4. Gastos Indirectos

Además de los gastos directos mencionados anteriormente, también existen gastos indirectos que se deben tener en cuenta. Estos gastos incluyen:

Tabla 8

Gastos Indirectos

Herramienta	Definición	Valor
Energía eléctrica y refrigeración	El funcionamiento de los equipos de hardware, como servidores, unidades de almacenamiento y sistemas de copia de seguridad, consume energía y genera calor. Esto puede requerir mayores costos de energía y la implementación de sistemas de enfriamiento apropiados para mantener el equipo en óptimas condiciones.	US$ 0.154
Seguridad y protección de datos	La implementación de un sistema de archivo de datos significa garantizar la seguridad y protección de los datos almacenados. Esto puede requerir una inversión en soluciones de seguridad como firewalls, sistemas de detección de intrusos, cifrado de datos y sistemas de copia de seguridad y recuperación ante desastres.	US$ 200.00 mes
Mantenimiento y actualizaciones	Los sistemas de archivo de datos requieren mantenimiento y actualizaciones periódicas para garantizar un rendimiento y una compatibilidad óptimos con su sistema operativo y sus aplicaciones. Esto puede incluir tarifas de licencia de software, renovaciones de contratos de soporte técnico y tiempo dedicado por el personal técnico al mantenimiento.	US$ 222.00
Ingeniero de Datos	Un ingeniero de datos es un profesional que se especializa en el diseño, desarrollo y mantenimiento de sistemas de gestión de datos. Su función principal es garantizar que los datos se recopilen, procesen, almacenen y utilicen de manera eficiente y segura dentro de una organización.	US$ 922.00

Nota**:** *Elaboración propia con información tomada de las páginas web de Enel, 2023. https://www.enel.com.co/es/personas/tarifas-energia-enel-distribucion.html y Resguarda, 2023. https://www.resguarda.com/ciberseguridad/.*

4. Resultados

4.1 Fase Uno Exploración

De la revisión realizada a la información suministrada se encontró lo siguiente:

- La información contenida en los correos electrónicos corresponde al período comprendido entre el año 2013 a 2023.
- Maneja 5 cuentas de correo electrónico que entre todas contienen 330 Gb de información de las cuales 193 Gb se encuentran almacenados en el Drive y 138 en el correo.
- El resto de información correspondiente a 145 Gb se encuentra contenida en discos duros y equipos de cómputo.
- Aplicación de política de Protección de Datos.

4.2 Fase Dos

- Se establece un modelo de entidad-relación para estructurar la información.
- Diseño de entidades, atributos y restricciones para el modelo.
- Identificación de reglas y restricciones para garantizar la integridad de los datos.

4.3 Fase Tres

- Se establecen las políticas para manejo de datos.
- Establecer las reglas retención de los datos y eliminación de los mismos.

4.4 Fase Cuatro

- Se determina las herramientas tecnológicas necesarias para la aplicación del proyecto tales como servidores, motores de datos, herramientas para limpieza de datos.
- Debido a que Bibliweb ya cuenta con Microsoft 365, la propuesta contempla la continuidad con este paquete de office.

- Se determinan los costos de herramientas en la nube y físicas siendo estas últimas las que se ajustan más a las necesidades actuales de Biblioweb por costos y cantidad de información actual de la empresa

5. Conclusiones y Recomendaciones

5.1 Conclusiones

Al realizar el diagnóstico se evidenció que Biblioweb SAS no contaba con una categorización, clasificación y unificación de la información que reposa en las distintas unidades de almacenamiento.

Se sugiere a Biblioweb la creación de políticas de manejo de la información, teniendo en cuenta cada una de las etapas del ciclo vital, y de igual forma la demanda de consulta de la misma.

Una vez realizado el análisis de la información suministrada y teniendo en cuenta que Biblioweb es una Pyme y además por el volumen de información que administra, se sugiere la adquisición de un servidor físico y el licenciamiento de los distintos software relacionados con limpieza, manipulación y almacenamiento de datos.

5.2 Recomendaciones

Aplicar políticas que permitan la adecuada retención de los datos y almacenamiento de estos para todos los procesos de Biblioweb.

Capacitar al personal de la empresa en el proceso de Data Archiving y cómo acceder a los datos archivados si es necesario.

Establecer un sistema de monitoreo y seguimiento para garantizar que el proceso de Data Archiving esté funcionando.

Referencias

Amazon. (10 de 05 de 2023). *Amazon*. Obtenido de https://www.amazon.com/-/es/PR2100-Professional-Network-Attached-Storage/dp/B071GR5VK6?th=1

Archivo General de la Nación Colombia. (2006). *Acuerdo número 027 de 2006.* Obtenido de normativa.archivogeneral.gov.co/acuerdo-27-de-2006/?pdf=500

Archivo General de la Nación Colombia. (2012). *Política de Porotección de Datos.* Optenido de

Capterra. (10 de 05 de 2023). *Capterra*. Obtenido de https://normativa.archivogeneral.gov.co/ley-estatutaria-1581-de-2012/?pdf=1261https://www.capterra.co/software/48791/postgresql

Casas Roma, J., Julbe López, F. y Nin Guerrero, J. (2019). *Big data: análisis de datos en entornos masivos.* UOC Editorial.

Enel. (10 de 05 de 2023). *Enel Codensa*. Obtenido de https://www.enel.com.co/es/personas/tarifas-energia-enel-distribucion.html

Gutierrez P. (5 de noviembre de 2013). *Fundamentos de las Bases de Datos: Modelo entidad-relación.* Obtenido de https://www.genbeta.com/desarrollo/fundamento-de-las-bases-de-datos-modelo-entidad-relacion

Lerma González, H. D. (2009). *Metodología de la investigación: propuesta, anteproyecto y proyecto (4a ed.).* Ecoe Ediciones. https://elibro-net.ucompensar.basesdedatosezproxy.com/es/lc/ucompensar/titulos/69092

Marr, B. (2017). *Data Strategy: How to Profit from a World of Big Data, Analytics and the Internet of Things.* Kogan Page.

Microsoft. (10 de 05 de 2023). *Microsoft 365*. Obtenido de https://www.microsoft.com/es-co/microsoft-365/business/compare-all-microsoft-365-business-products

Ministerio de Tecnologías de la Información y las Comunicaciones. (2019). *Guía para la gestión de documentos y expedientes electrónicos.* Obtenido de mintic.gov.co/arquitecturati/630/articles-61594_recurso_pdf.pdf

Ministerio de Tecnologías Información y las Comunicaciones. (2021). *Ciclo de Vida del Dato*. Obtenido de www.mintic.gov.co/portal/715/articles-198952_anexo_1_2_ciclo_vida_dato.pdf

MongoDB. (10 de 05 de 2023). *MongoDB.* Obtenido de https://www.mongodb.com/es/what-is-mongodb

Nel Quezada, L. (2010). *Metodología de la Investigación.* Marcombo.

PowerData Solutions S.L. (2023). *Ebook Data Archiving.* Obtenido de https://landings.powerdata.es/ebook-data-archiving-como-optimizar-el-valor-de-los-archivos-de-datos

Refine, O. (10 de 05 de 2023). *Escuela de Datos*. Obtenido de https://escueladedatos.online/openrefine/

Resguarda. (10 de 05 de 2023). *Resguarda Ciberseguridad*. Obtenido de https://www.resguarda.com/ciberseguridad/