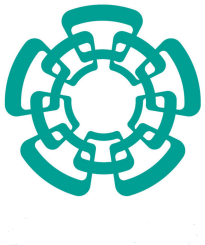


Centro de Investigación y de Estudios Avanzados
del Instituto Politécnico Nacional

Unidad Cinvestav Tamaulipas

Método de orquestación para servicios de fusión de datos definidos por variables espacio-temporales

Tesis que presenta:

José Carlos Morín García

Para obtener el grado de:

Maestro en Ciencias en Ingeniería y Tecnologías Computacionales

Dr. José Luis González Compeán, Co-Director
Dr. Iván López Arévalo, Co-Director

Cd. Victoria, Tamaulipas, México. Agosto, 2022

La tesis presentada por José Carlos Morín García fue aprobada por:

Dr. José Juan García Hernandez

Dr. Hiram Galeana Zapien

Dr. Iván López Arévalo, Co-Director

Dr. José Luis González Compeán, Co-Director

Cd. Victoria, Tamaulipas, México., 8 de Agosto de 2022

...

Agradecimientos

- A mi familia por ....
- A mis revisores, Dr. Hiram Galeana Zapien y Dr. José Juan García Hernández, por su esfuerzo y disponibilidad para revisrar y sugerir modificaciones para mejorar el documento de tesis.
- A mis directores, Dr. X y Y ....
- Al Consejo de Ciencia y Tecnología (CONACYT) por la beca otorgada para la realización de mis estudios de Maestría.
- Al Proyecto 41756 "Plataforma tecnológica para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud y construcción de un repositorio nacional de servicios de análisis de datos de salud" del FORDECYT-PRONACES-CONACYT; proyecto liderado por Cinvestav Tamaulipas.
- Los conjuntos de datos reales utilizados en este trabajo pertenecen y fueron proporcionados gentilmente por el Proyecto 48901 "Sistema de monitoreo de la atención a los trastornos mentales y del comportamiento debidos al consumo de sustancias" del FORDECYT-PRONACES-CONACYT; proyecto liderado por el Instituto Nacional de Psiquiatría.

Índice General

Índice de Figuras

Índice de Tablas

Índice de Algoritmos

Publicaciones

Resumen

Método de orquestación para servicios de fusión de datos definidos por variables espacio-temporales

por

José Carlos Morín García
Unidad Cinvestav Tamaulipas
Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, 2022
Dr. José Luis González Compeán, Co-Director
Dr. Iván López Arévalo, Co-Director

En diversos dominios los procesos de toma de decisiones se basan en información obtenida a partir de cúmulos de datos. No obstante, debido al incremento que han mostrado en últimos años estos cúmulos de datos y a la cantidad de ellos en un mismo dominio, recientemente se han incorporado herramientas de fusión de datos (FD) para conformar conjuntos de datos ($datasets$) que integren una o más fuentes de datos. Esto se realiza con el objeto de tener en un mismo lugar todas las variables de un problema, lo cual permita tener un panorama amplio del problema abordado antes de tomar decisiones sobre el mismo. Desde la perspectiva tecnológica, la creación de un servicio de FD no es una tarea fácil porque estos servicios, generalmente, están asociados al proveedor de servicios de FD (lo que genera una dependencia *usuario-proveedor*), los cuales a su vez dependen del método de orquestación de datos (lo que genera una dependencia *fusión-orquestación*).

En este trabajo de investigación se propone un método de orquestación para fusión de datos, el cual es independiente de la infraestructura de cómputo donde ejecuta y es definido por esquemas basados en variables espacio-temporales para crear servicios dinámicos de FD. Este método incluye un esquema de despliegue y acoplamiento de servicios agnósticos de la infraestructura de cómputo en la nube que evitan la dependencia *usuario-proveedor*. Mediante un modelo de construcción basado en cubos de datos espacio-temporales permite desacoplar la fusión de datos de la orquestación, lo cual habilita a los usuarios para crear servicios dinámicos de FD, a su vez evitando la dependencia

fusión-orquestación. Para la creación de estos servicios se propone un esquema declarativo para definir el modelo de construcción, el cual describe la manera de cómo se conectan los componentes para la FD. A partir del método propuesto de desarrolló un prototipo funcional que se ha probado en entornos reales (climatología y medicina), además que se ha probado con datos sintéticos. Los resultados obtenidos hasta el momento son prometedores.

1

Introducción

En este capítulo se describen los antecedentes que dieron origen al trabajo de tesis. Se contextualiza lo que sucede actualmente en relación a la necesidad de fusuonar datos y cómo se ha abordado de manera general. A partir de ello se formula la pregunta de investigación que se desea resolver mediante la hipótesis de investigación que se enuncia. Para cumplir con la hipótesis se presentan los objetivos que se formularon. Finalmente se describe, de manera general, la metodología de investigación seguida.

1.1 Antecedentes y motivación

Este trabajo de tesis se contextualiza en el área de Sistemas Distribuidos, dentro de Cómputo en la Nube. Específicamente en el proceso de *orquestación para servicios de fusión de datos*, abordando el problema de *vendor lock-in*.

Este documento de tesis reporta el trabajo realizado para crear un método de orquestación para

fusión de datos que incluya, en una primera fase, un modelo de manejo declarativo para desplegar servicios de fusión de datos sobre múltiples infraestructuras y plataformas de cómputo. En una segunda fase obtener un modelo de fusión de datos basado en variables espacio-temporales definidas por el usuario para desacoplar los procesos de fusión de la orquestación, que se realizará mediante el modelo de la primera fase. El tema central de este trabajo de tesis es la fusión de datos. Ésta puede verse desde distintas perspectivas, inicialmente fue un tema que inició en el dominio de bases de datos, pero que ha venido evolucionando hasta hoy día siendo abordado en este trabajo desde la perspectiva de Big Data. Así, a continuación se aborda el tema desde el concepto de grandes volúmenes de datos hasta llegar al concepto de manejo de datos en Cómputo en la Nube.

1.1.1 Manejo de grandes volúmenes de datos

En la actualidad se están generando, registrando, analizando, compartiendo y consumiendo cantidades ingentes de datos e información [46] [45] [29]. De acuerdo con un estudio de IBM, se estima que diariamente en el mundo se generan más de 2.5 quintillones de bytes [1]. Actualmente un gran porcentaje de los cúmulos de datos (volúmenes de datos que crecen a tazas constantes y que acumulan el volumen de crecimiento al volumen de datos original) [12][2] se encuentran disponibles en ecosistema de la nube [17] y son susceptibles de ser usados para extraer información y conocimiento. Esto resulta crucial para procesos de toma de decisiones en distintos dominios, tales como la observación de la tierra/espacio [47], manejo de territorio [20], estudio/diagnóstico de enfermedades [38], manejo de efectos contaminantes/demografía [58], por nombrar algunos.

Para gestionar una gran cantidad de datos con el fin de analizarlos para extraer información se aplican técnicas de procesamiento de grandes volúmenes de datos, conocidos como técnicas de *Big Data*. Estas técnicas incluyen conjuntos de herramientas, algoritmos de analítica y procedimientos

[1] `https://developer.ibm.com/es/articles/que-es-big-data/`

[2] Estudios recientes indican que el 40 % de los datos son adecuados para producir información mediante procesos de analítica de datos.

de manejo de información para organizar la creación, manipulación y tratamiento de cúmulos de conjuntos de datos [8].

El modelo de procesamiento tradicional utilizado para integrar las herramientas consideradas para el procesamiento de datos es el llamado *Extracción-Transformación-Carga* o (ETL por sus siglas en inglés de *Extraction-Transformation-Load*) [14]. La Figura 1.1 muestra un ejemplo de un procesamiento de un volumen de datos modelado como un esquema *ETL*. De esta forma, los datos son Extraidos desde algún lugar, p.ej. de la nube, Transfomados por una técnica de Big Data, cuyo resultado es entregado (L) a un proceso de toma de decisiones [26].

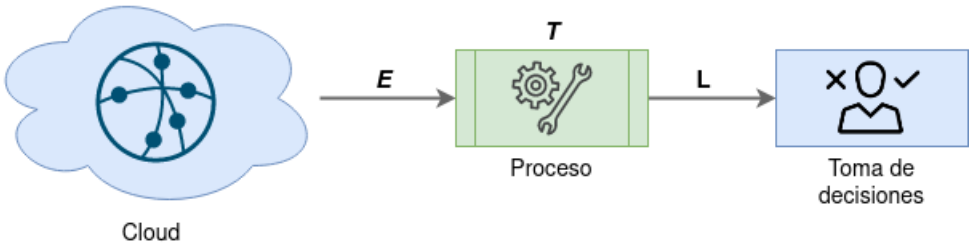


Figura 1.1: *ETL* en Big Data

Los procesos de *extracción* de datos están asociados a herramientas o servicios de extracción de contenidos (crawlers [55] o API's de sistemas de archivos), así como herramientas de preparación de datos (adaptación a formatos XML, JSON, GJson, etc). Los procesos de *transformación* básicamente son algoritmos de preprocesamiento y procesamiento de datos. Los algoritmos de preprocesamiento generalmente incluyen herramientas tales como limpieza, homogenización, eliminación de outliers, mapeo, preparación de datos (series de tiempo, muestreo, reducción de dimensionalidad o enriquecimiento por interpolación/extrapolación de datos faltantes, etc.) [15]. Los algoritmos de procesamiento consideran servicios tales como:

- *Creación de contenidos* (BDC, *Big Data Content*), que concierne al procesamiento en ráfagas (*streaming processing*), el cual es común en escenarios de Internet de las Cosas (IoT) o Industria

4.0 [33] o distribución y entrega de contenidos [24] (como las plataformas de entretenimiento Netflix, Amazon Prime, etc.).

- *Analítica de grandes volúmenes de datos* (BDA, *Big Data Analytics*), los cuales están basados en modelos estadísticos/probabilísticos cuya función es mapear cúmulos de datos a información útil [4]. Comprende el proceso de examinar fuentes de datos para descubrir información, como patrones ocultos, correlaciones, tendencias y preferencias de usuarios, que pueden ayudar a las organizaciones a tomar decisiones informadas [46]. Esta es una forma de análisis avanzado, que involucra aplicaciones complejas con elementos predictivos, algoritmos estadísticos y análisis hipotéticos impulsados por sistemas analíticos.

1.1.2 Manejo de servicios de fusión de datos

Una de las tareas muy importantes en los servicios de *BDA* es la *fusión de datos*, que en sí también se ofrece como servicio en la nube[3]. El servicio de fusión de datos resulta especialmente desafiante, ya que el modelo tradicional de ETL no necesariamente captura su comportamiento [9]. La fusión de datos (*data fusion* -FD-) se puede entender como una herramienta de BDA que une/integra datos provenientes de múltiples fuentes de datos heterogéneas, con el fin de incrementar la calidad de los resultados a obtener en algún algoritmo de análisis que use tales datos (en términos de exactitud o precisión de un modelo) y entregar esta información a los procesos de toma decisiones.

En los servicios de fusión de datos se consideran aspectos tales como:

- La integración o concatenación de las variables de múltiples fuentes de datos. Es decir, juntar unas a la par de otras, variables de múltiples fuentes para tener un conjunto de datos

[3]Un servicio en la nube, ofrecido por un proveedor/vendedor, es una herramienta/aplicación de software que puede consumirse por aplicaciones o usuarios finales. Los hay de distinta naturaleza, desde aplicaciones simples que realizan cálculos numéricos hasta aplicaciones que encadenan la salida de unas aplicaciones con otras.

más completo. Este es un desafío porque debe asegurarse que la unión de esas variables es consistente. Si las fuentes comparten variables la integración debe tomar en cuenta los datos existentes en tales fuentes.

- El enriquecimiento de los valores de las variables del dataset resultante mediante regresiones, interpolaciones, extrapolaciones, o algún método de análisis aplicado a dichas fuentes con nuevos datos.

- La reducción de datos mediante intersecciones o uniones de variables. La unión no necesariamente debe juntar variables de las distintas fuentes, sino que incluso puede disminuir el numero de variables o registros cuando lo que se desea es la intersección de las fuentes de datos. Por ejemplo, eliminar aquellas variables que no se ocuparán o hay pérdida de datos en sus registros, por lo que no tiene sentido que se conserven.

Dentro de la fusión de datos existen varios modelos y técnicas que se pueden emplear, las cuales se pueden dividir de acuerdo con los siguientes criterios:

- Atendiendo la relación entre las entradas y salida propuesta por Durrant-Whyte [16].

- De acuerdo a las entradas/salidas de los tipos de datos y su naturaleza definida por Dasarathy [13].

- Basado en los diferentes niveles de fusión de datos definidos por el Departamento de Defensa de los Estados Unidos [6].

- Dependiendo de la arquitectura en donde se despliegue el servicio de fusión de datos (centralizada, descentralizada y distribuida) [9].

Todos los modelos de fusión de datos incluyen la entrada de i puntos de datos (pre-establecidos) a un servicio de BDA (transformador) y el resultado de la fusión es usado como insumo de otros procesos ya sea de FD, BDA o de resumen de datos (usado por los procesos de toma de decisiones).

1.1.3 Manejo de servicios de fusión de datos en la nube

Actualmente los servicios de fusión de datos son generalmente ejecutados en la nube. En los últimos años han resultado determinantes para que las organizaciones y/o comunidad científica lleven a cabo estudios de fusión de información en dominios tales como medicina[21] (donde se interconecta información ambiental con historiales clínicos para descubrir prevalencia de agentes de enfermedades), clima [19] (donde se interconecta información de temperaturas con contaminantes para descubrir correlaciones), observación de la tierra [51] (donde se interconecta información satelital con información de monitoreo ambiental), por nombrar algunos.

La Figura 1.2 se muestran los diferentes componentes requeridos para crear un servicio de fusión de datos en la nube [50][36][48]. Como se puede observar, se requieren algoritmos de análisis, manejo de datos, herramientas para acceso a la infraestructura (en donde se está desplegado el servicio), y las fuentes de datos que se van a analizar.

Para la unión de los algoritmos de análisis y manejo de datos se genera software en *BDA*, que cuando son desplegados en una infraestructura de cómputo se obtienen *plataformas para BDA*. Tales plataformas pueden trabajar con datos que son entregados por diferentes fuentes. Una vez que se dispongan estas plataformas diseñadas y desplegadas, es posible realizar la fusión de datos o cualquier otro tipo de proceso de *BDA*.

El modelo de procesamiento tradicional ETL considera una única entrada de extracción para la fase de transformación. Por tanto, este modelo de procesamiento no puede, por definición, producir un modelo de fusión de datos. En este sentido, se debe adecuar el modelo ETL para poder materializar un modelo de fusión de datos en nuevos escenarios de manejo de datos.

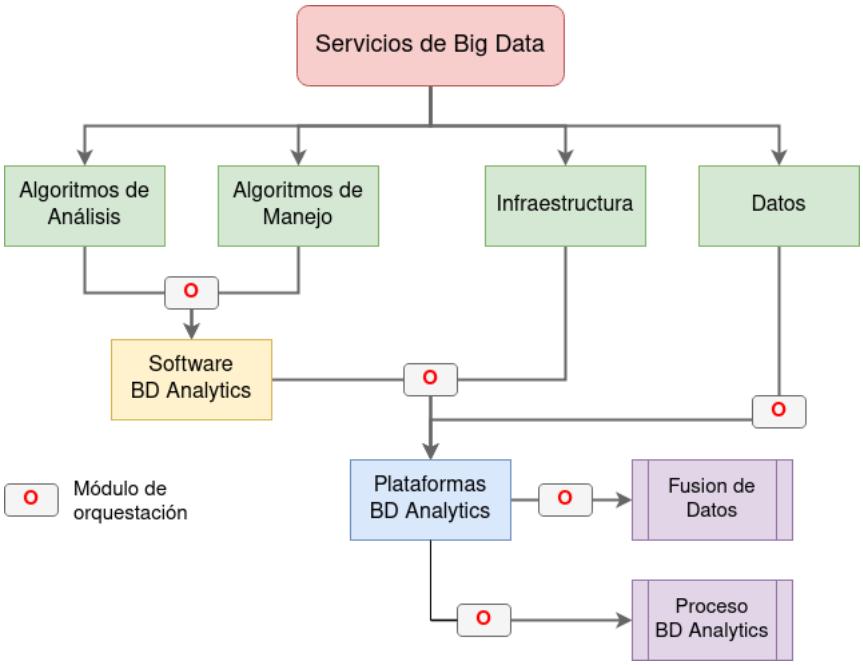


Figura 1.2: Componentes tradicionales de un *BDA* para crear un servicio de fusión de datos.

Para la construcción de un servicio FD se requiere de un *orquestador* (específicamente *orquestador de datos*), el cual gestione las múltiples entradas y/o posibles potenciales interconexiones recursivas de las salidas a entradas de otros procesos de fusión [50][8]. En el dominio de Cómputo en la Nube, un orquestador es una entidad encargada de automatizar tareas específicas en una infraestructura de nube (elementos físicos y virtuales) para agilizar la ejecución de procesos de interés, dando la apariencia de que los procesos se ejecutan de forma autónoma (configuración, inicialización, encendido/apagado y escalamiento). Su objetivo es garantizar un nivel de servicio adecuado a los usuarios finales sin que se note una degradación de los servicios. Haciendo una analogía, es la misma figura de un *maestro de orquesta musical*. La Figura 1.3 ilustra el funcionamiento de un orquestador, organizando los componentes de la parte superior de acuerdo a su uso en la parte inferior de la figura.

La *orquestación de datos*, por tanto, se puede entender como un proceso automatizado en el que un servicio de software crea las estructuras de software requeridas (integración de módulos ETL) para materializar un modelo de FD [56]. Este servicio crea estructuras de software que se ejecutan

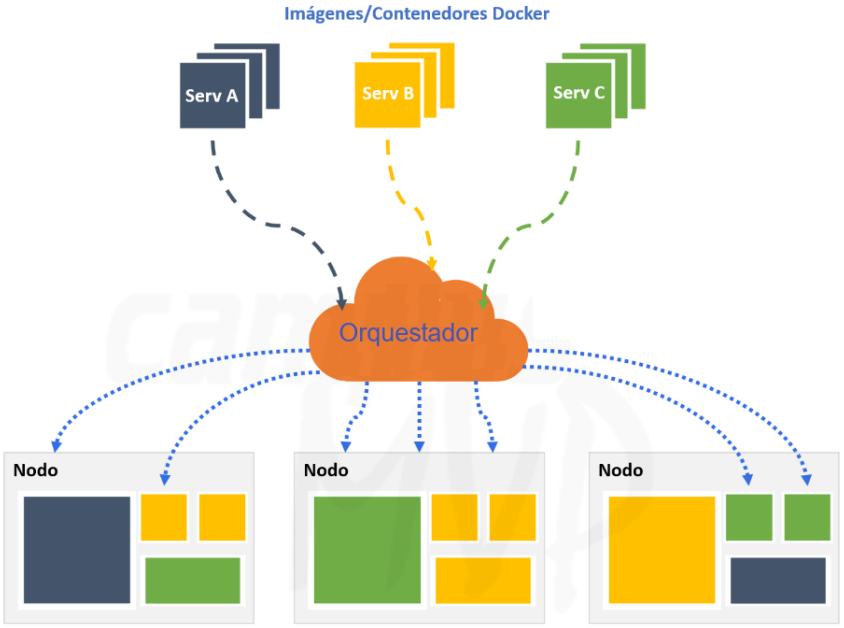


Figura 1.3: Ilustración del funcionamiento de un orquestador.

en la nube de manera encadenada (en serie o paralelo) para manejar la extracción desde múltiples fuentes de datos, entregar los datos extraídos a las múltiples fases de transformación y finalmente consolidar los datos en una sola respuesta para los procesos de toma de decisiones [42]. Los procesos de orquestación son generalmente creados *ad hoc* y esta creación depende de las particularidades de los modelos de fusión (sus múltiples entradas y/o múltiples interconexiones).

1.2 Motivación

En distintos dominios, cada vez más y más, los procesos de toma de decisiones se basan en información obtenida a partir de cúmulos de datos[31]. Recientemente, estos procesos han comenzado a incorporar herramientas para confrontar situaciones donde la información proviene de múltiples fuentes, incluso de distinta naturaleza, con el fin de tener un panorama amplio del problema antes de tomar decisiones enfocadas a resolver un problema dado [18]. Es posible integrar bases de datos de distinta índole, bases de datos con hojas de cálculo, hojas de cálculo con tablas HTML, por

mencionar algunos casos.

Los procesos FD, por tanto, representan una herramienta clave para los procesos de toma de decisiones. El incorporar un proceso FD a un servicio existente (p.ej. un proceso de predicción de umbrales), así como producir servicios FD dinámicos (adaptables a distintos servicios que los necesiten) resulta un desafío tanto tecnológico como de investigación. Crear sistemas de FD dinámicos implica que la FD no esté condicionada ni por el proveedor de servicio (que no se produzca una dependencia *usuario-proveedor*), ni por el método de orquestación de datos (que se produzca una dependencia *fusión-orquestación*); esto último implica que la orquestación de datos no esté definida exclusivamente por el modelo FD. Crear este tipo de servicio no es una tarea fácil porque se deberían afrontar dos retos importantes:

1. Crear servicios agnósticos a la infraestructura[4] que permitan a los usuarios declarar los lugares desde los cuales se obtendrán las fuentes de datos, los lugares donde se ejecutarán los procesos de FD y los lugares en donde se depositará/entregará la información resultante.

2. Crear esquemas de acoplamiento para la orquestación de datos para servicios FD sin depender del modelo fusión, los cuales permitan a los diseñadores de aplicaciones crear servicios FD basados en diferentes modelos de entradas.

1.3 Planteamiento del problema

Actualmente existen servicios FD y BDA que son, generalmente, ejecutados en entornos de nube, pero adolecen de problemas de dependencias, los cuales se expresan en dos direcciones: la dependencia *usuario-proveedor* y la dependencia *fusión-orquestación*, los cuales se describen a continuación.

[4] *Agnóstico a la infraestructura* significa que un servicio/software/aplicación no depende en exclusiva de un tipo de infraestructura de cómputo, sino, por el contrario, puede ejecutar en cualquier infraestructura de cómputo.

- *Dependencia usuario-proveedor*: La dependencia del usuario de servicios de FD con el prestador de dichos servicios (lo que se conoce como *vendor lock-in* [5]) se produce cuando el usuario delega tanto las fuentes de datos, como los servicios de FD o BDA y orquestación a los proveedores de servicios en la nube (ver flecha azul en Figura 1.4) [40]. En tal caso, la factibilidad de ejecución de un FD depende de los recursos y disponibilidad del proveedor. De la misma forma, esta dependencia se extiende a la elección del método FD, los servicios BDA incluidos en el servicio de FD y la recuperación de información. Esta dependencia produce efectos tales como acumulación de datos (que pueden imposibilitar la migración de datos/servicios a otros proveedores o que incrementen los costos de los servicios, sin posibilidad de cambiar de proveedor) y no la disponibilidad del FD o de los servicios del mismo proveedor de servicios en casos de falla (principalmente apagones, no disponibilidad de recursos, problemas de ruteo IP, etc.).

- *Dependencia fusión-orquestación*: Se produce cuando el usuario sólo crea procesos de FD dependiendo de los servicios de BDA que están en el catálogo del proveedor y datos disponibles en la nube del proveedor. Estos servicios están definidos por un esquema de orquestación propio y predefinido con parámetros estáticos (que el usuario no puede modificar). Estos parámetros están definidos por los proveedores a conveniencia de los servicios que ofrecen. Usualmente a los proveedores les conviene que estos parámetros sean estáticos para garantizar una calidad de servicio previamente pactada (ver flecha roja de la Figura 1.4).

La problemática abordada en este trabajo de tesis básicamente se puede resumir en la dependencia de los servicios de FD del proveedor (*vendor lock-in*), ya que en los dos casos antes descritos, ambas dependencias recaen en el proveedor, lo cual sería deseable mitigar, o evitar, en el mejor de los casos.

[5] *Vendor lock-in* es la acepción en inglés que describe la dependencia de un usuario de un servicio con el prestador de dicho servicio[39]. En este trabajo el servicio es FD y el prestador es un proveedor de servicios de FD en la nube.

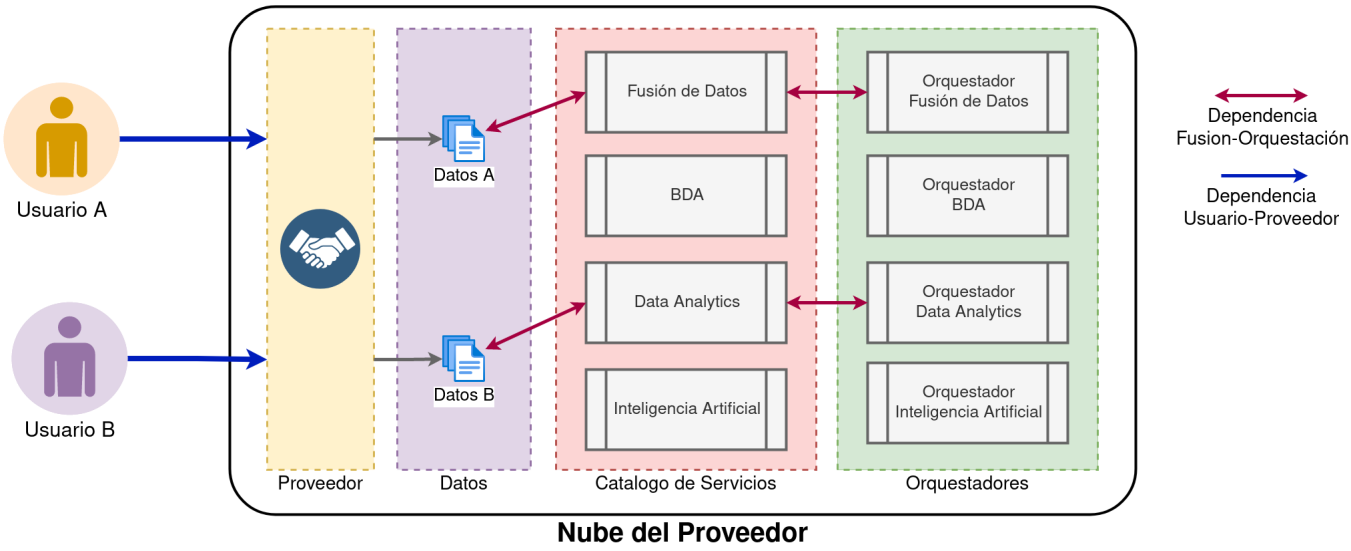


Figura 1.4: Dependencias entre el usuario y los servicios de BD.

Haciendo énfasis en el problema del *vendor lock-in* de los servicios FD se plantea la siguiente pregunta de investigación:

¿Qué método de orquestación para servicios de FD puede reducir situaciones de dependencia usuario-proveedor y fusión-orquestador?

Atendiendo a la pregunta de investigación y tomando en cuenta lo reportado en el estado del arte respecto al tema, es conveniente establecer las siguientes premisas:

Dependencia usuario-proveedor

1. Los usuarios deben crear sus servicios FD asumiendo como límites los recursos de infraestructura ofrecidos por el proveedor.

2. El proveedor de infraestructura es quien determina los factores no-funcionales (seguridad, confiabilidad, eficiencia, costos, etc) que los servicios que ofrece deben cumplir.

Dependencia fusión-orquestador

1. Las variables espaciales y/o temporales son comúnmente utilizadas para realizar procesos de FD. Esto se debe a que este tipo de variables son usadas como metadatos descriptores del contexto de los eventos (e.g. registros de una base de datos o acciones en contenidos). En este sentido, el espacio/tiempo de eventos descritos en una fuente de datos se puede equiparar con el espacio/tiempo de eventos de otra fuente, aunque dichas fuentes sean de naturaleza heterogénea. Razón por la cual, estas variables son comúnmente utilizadas como denominador común en procesos de FD.
2. La orquestación de datos está ligada funcionalmente (por las tareas que se realizan) a la infraestructura, plataforma y/o software empleados.
3. Los métodos de FD dependen funcionalmente de la orquestación definida por el proveedor.

1.4 Hipótesis

Con base en las preguntas de investigación y las premisas establecidas, es posible definir la hipótesis de este trabajo de tesis:

Un método de orquestación, independiente de la infraestructura y determinado por esquemas basados en variables espacio-temporales, puede crear servicios dinámicos de FD que mitiguen los efectos de las dependencias usuario-proveedor y fusión-orquestador.

1.5 Objetivos

Para comprobar el cumplimiento de la hipótesis planteada, se establecieron los siguientes objetivos.

General

Crear un método de orquestación agnóstico para servicios dinámicos de FD sobre fuentes de datos

que contengan variables espacio-temporales.

Específicos

1. Definir un modelo de procesamiento agnóstico (de la infraestructura y/o plataforma) para crear servicios dinámicos de FD sobre un entorno de nube mediante cláusulas declarativas.
2. Crear un esquema de despliegue y acoplamiento de servicios de FD definido por variables espacio-temporales empleando el modelo anterior.

1.6 Metodología de trabajo

Para cumplir con los objetivos planteados, se diseñó una metodología de trabajo compuesta de cuatro etapas encadenadas. En esta sección se describe de manera general cada una de las actividades en cada etapa para el desarrollo de este trabajo de investigación. A continuación se presentan los detalles.

- **Definición del problema**

  Esta primera etapa está compuesta por dos actividades; la primera es la identificación del problema, para ello se plantea conocer todo lo relacionado con el tema. El objetivo principal es analizar el problema, identificar las soluciones actuales (en caso de existir) y conocer la importancia de darle solución. La segunda actividad es el estudio del estado del arte, esto permitió ubicar el avance que se tiene hasta el momento en relación con el tema que se pretende resolver.

- **Definición del método**

  Una vez identificado el tema/problema y la forma de abordarlo, fue necesario definir un método de orquestación para abordar la solución de la problemática expresada en dos dependencias (ver sección), para ello se tomaron en cuenta algunos elementos:

- Definir método de orquestación de datos agnóstico.
- Definir un modelo declarativo que se adapte al método de orquestación de datos.
- Diseñar una arquitectura de sistemas distribuidos que se adapte al modelo de orquestación de datos para servicios de FD.

- **Implementación del método**

  A partir de la etapa anterior se realizó la implementación para comprobar que lo que se está proponiendo resuelve el problema de manera adecuada. Algunos elementos importantes de esta implementación son los siguientes:

  - Implementar un lenguaje declarativo que interprete el modelo declarativo de orquestación de datos.
  - Desarrollar una arquitectura de sistemas distribuidos que mejor se adapte al modelo de orquestación de datos para FD.
  - Diseñar un prototipo adaptado al lenguaje declarativo y la arquitectura de sistemas distribuidos.
  - Generar pruebas sobre el comportamiento del prototipo y solucionar errores encontrado.

  Los elementos anteriores permitieron el desarrollo de una herramienta que nos ayudara a realizar procesos de fusión o si bien desea el usuario de algún proceso de analítica basado en las cláusulas declarativas definidas en el lenguaje.

- **Experimentación**

  En esta etapa se llevaron a cabo una serie de experimentos para conocer si los resultados obtenidos garantizaron que el método propuesto fue el adecuado para el problema identificado. Para ello se tuvo que cumplir con las actividades siguientes:

- Pruebas: Se realizaron pruebas de la implementación y un análisis de los resultados.

  Experimentación realizada:

  1. Estudio de caso 1 con datos meteorológicos.
  2. Estudio de caso 2 con datos médicos.
  3. Estudio de caso 3 con datos sintéticos.

- Ajuste de diseño: Permitió adaptar el diseño propuesto según los resultados obtenidos.

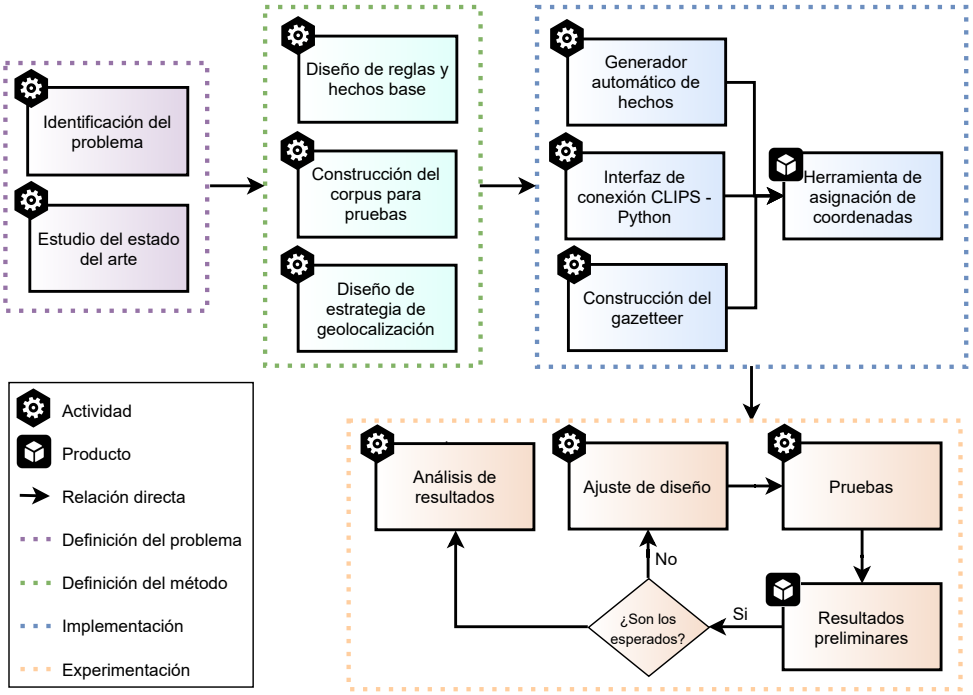


Figura 1.5: Metodología de desarrollo

1.7 Organización de la tesis

El documento está organizado de la siguiente manera. En el capítulo 2 se describen los conceptos básicos necesarios para el desarrollo de este trabajo de investigación, además de los trabajos del estado del arte con respecto a la desambiguación. En el capítulo 3 se presenta el método propuesto,

describiendo los módulos que lo componen. En el capítulo 4 se presenta la evaluación experimental y el análisis de los resultados obtenidos a partir de la implementación del método propuesto. En el capítulo 5 se presentan las conclusiones obtenidas, los inconvenientes que se presentaron durante la realización del trabajo y el trabajo futuro identificado.

2

Estado del arte

El tema central de este proyecto de tesis es la orquestación para crear servicios de fusión de datos. En esta sección se presenta una breve introducción a los conceptos abordados en esta tesis y la comparación entre la fusión de datos con la integración de datos. También se describen algunos trabajos relacionados con lo propuesto en la tesis. La principal intención de esta sección es mostrar el estado del arte y dar a conocer la forma cómo se ha abordado la orquestación de datos para fusión de datos en algunos dominios.

2.1 Conceptos relevantes

En este apartado se muestran los conceptos fundamentales asociados con el desarrollo de esta tesis y cómo han sido involucrados dentro del tema y su relevancia dentro de las investigaciones y desarrollo de servicios.

Orquestación de datos

La orquestación en *Big Data* (BD) se refiere al control centralizado de los procesos que administran datos, sistemas, centros de datos o fuentes masivas de datos (también conocidas como *datalakes*). Las herramientas de orquestación de *BD* permiten a los equipos de Tecnologías de la Información (IT) diseñar y automatizar procesos de un extremo a otro. Estos procesos pueden incorporar datos, archivos y dependencias completas de una organización, sin tener que escribir scripts personalizados a cada situación en donde se empleen [56].

De acuerdo con *Advanced Systems Concepts, Inc.*[1], las plataformas de orquestación de datos permiten a los equipos de IT integrar rápidamente nuevas fuentes de datos existentes utilizando el modelo de procesamiento de datos ETL y procesos de *BD* [McHugh].
Así, la orquestación de datos permite que cada punto de procesamiento de datos del usuario funcione en armonía dentro de un flujo de trabajo (módulos de ETL encadenados) que configure el mismo usuario. Las herramientas de orquestación de datos han tenido aplicación en áreas recientes, como en las redes de telefonía 5G [1] y sistemas de *blockchain* [41].

Fusión de Datos

El Departamento de Defensa de Estados Unidos define la Fusión de Datos (FD) como "un proceso multinivel y multifacético que se encarga de la detección, asociación, correlación, estimación y combinación de datos e información de múltiples fuentes" [7]. Entre las áreas de aplicación de FD se encuentra la combinación de datos espacio-temporales [47], datos médicos [38] y datos de observación terrestre [20]. Esta última es un área fuertemente heterogénea para el desarrollo de sistemas de observación de la Tierra, en donde se requieren múltiples vistas sobre un mismo objeto, común en distintas fuentes de datos. En la fusión de datos se pueden encontrar diversos modelos y/o técnicas que ayudan a realizar este proceso, por ejemplo el modelo propuesto por la Junta de

[1] ActiveBatch, Advanced Systems Concepts, Inc. - https://www.advsyscon.com

Directores de Laboratorios (JDL[2] por sus siglas en inglés) concibe un modelo de procesos multinivel y fases que trabajan en la asociación, correlación y estimación de los datos [6] o bien la propuesta de Dasarathy [13], quien expuso un método de clasificación de fusión de datos de acuerdo a la naturaleza de los datos así como sus entradas y salidas (datos, características y decisiones).

Existen algunos trabajos enfocados en realizar el proceso de fusión de datos para observaciones médicas [32], observaciones dentro de bosques [28] y datos de variables espacio-temporales [54]. En estos escenarios la fusión de datos se utiliza para mejorar la calidad de los datos de entrada y, en consecuencia, el desempeño de los procesos de toma de decisiones. Lo que sucede es que se identifican las relaciones entre las diferentes fuentes de datos. Si las fuentes de datos por procesar son totalmente disjuntas, no es posible realizar el proceso de fusión.

Flujos de trabajos

En algunos dominios, cuando se requiere de la ejecución sincronizada y armoniosa de tareas especializadas (procesos especializados) que producen un resultado común a todas esta tareas, es conveniente utilizar flujos de trabajo (*workflow*). Éstos son utilizados para la automatización de procedimientos en los que los datos, información o tareas se pasan entre los participantes del flujo, de acuerdo con un conjunto definido de reglas para lograr un objetivo común [Hollingsworth].

Los flujos de trabajo pueden ser desplegados en escenarios de nube; por ejemplo Xu *et al.* [57] describieron una estrategia para generar múltiples flujos de trabajo con el fin de asegurar un buen desempeño de los servicios en una nube. Zulfiqar *et al.* [2] presentaron un sistema de manejo de flujos de trabajo en la nube que, bajo distintas condiciones de uso e infraestructura, es tolerante a fallos. Existen además otros trabajos en los cuales se puede ver que los flujos de trabajo ayudan en tareas de control y monitoreo de los diferentes puntos en los que se encuentra un proceso de interés

[2] Joint Directors of Laboratories - Grupo de trabajo de fusión de datos, establecido en 1986.

[23].

Modelo de procesamiento ETL

Extraer, Transformar y Cargar (*ETL -Extract, Transform, Load-*) es un modelo de procesamiento que permite realizar la adquisición de datos a partir de una fuente, su posterior transformación ejecutando alguna operación o modificación a los datos y su transferencia a un repositorio destino [14], como ilustra la Figura 1.1.

Los procesos que basados en este modelo ETL tienen aplicación dentro de Big Data. Por ejemplo Bansal [3] propuso un *framework* semántico utilizando tecnologías semánticas para la integración y publicación de múltiples fuentes basándose en el modelo *ETL*. El modelo ETL también tiene cabida dentro del área médica, por ejemplo el *framework* para la conversión de bases de datos de salud al modelo OMOP[3] [37] [44]. Con base en el proceso ETL básico se pueden construir y extender procesos que sean más grandes y más robustos. Incluso de un proceso *ETL* se puede pasar a otro proceso *ETL* más grande y así sucesivamente para generar flujos de trabajo que conjuguen todos los componentes de los procesos ETL más pequeños.

2.2 Trabajos relacionados

En este apartado se presentan algunos de los trabajos relacionados con el tema estudiado en esta tesis, que si bien no se generalizan explícitamente hacia lo que se propone en este trabajo de investigación, ni afrontan totalmente la problemática descrita, giran en torno al tema abordado. La búsqueda en la literatura se dividió en dos áreas: *(a)* trabajos relacionados con orquestación para fusión de datos y *(b)* herramientas que permiten realizar fusión de datos.

[3]El modelo de datos común de OMOP permite transformar los datos de observación dispares en un formato común - https://www.ohdsi.org/data-standardization/the-common-data-model

Cabe recordar que este trabajo de tesis se focaliza en conjuntos de datos que incluyan variables de tiempo y espacio. El tiempo se refiere a datos que contengan fechas, desde meses y años aislados hasta estampas de tiempo que incluyan detalles en minutos y segundos. El espacio se refiere a variables que denoten una ubicación física (coordenadas geográficas, localidades, espacios territoriales, coordenadas cartesianas, etc.). Con base en ello se exploró la factibilidad de crear un método armonizado de fusión de datos agnóstico para la nube; donde el manejo de distintas fuentes de datos es un desafío en procesos que se realizan para apoyar la toma de decisiones a partir de variables espacio-temporales [54].

2.2.1 Orquestación para fusión de datos

El análisis de los datos meteorológicos es un área propicia para la fusión de datos de distintas fuentes [43]. Esto se debe a la cantidad de fuentes de datos y la cantidad de datos que se generan diariamente. Es importante hacer notar que, aunque existe una gran cantidad de fuentes de datos climatológicas, éstos no siempre se pueden emplear debido a que no comparten campos que permitan su integración. No todas las fuentes de datos de interés son candidatas para realizar el proceso de fusión de datos debido a la poca información que contienen sobre el traslape de variables o carencias en los datos; y por ende, es necesario complementarlas con variables de una o más fuentes de datos externas que tengan información sobre el área analizada y que permitan identificar el traslape.

En esta sección se presentan aquellos trabajos, identificados hasta el momento, en la literatura que están relacionados con modelos de orquestación de datos. Estos trabajos tienen cierta relación con el objeto de este trabajo de tesis y la fusión de datos para análisis de comportamientos climáticos.

A) Fusión de datos espaciales en infraestructuras de datos espaciales utilizando *Linked Data*

Wiemann y Bernard en 2015 [54] describieron enfoques, requisitos y factores limitantes para la fusión

de datos espaciales basada en servicios, con un enfoque particular en la interacción de SDI (*Spatial Data Infrastructures*) y estándares de la Web Semántica. Las SDI establecidas proporcionan un medio para publicar, buscar, acceder y procesar datos espacio-temporales en la Web [5]. En ellas se aprovechan las tecnologías de la Web Semántica para permitir acceso ubicuo a datos interconectados en la Web [McHugh].

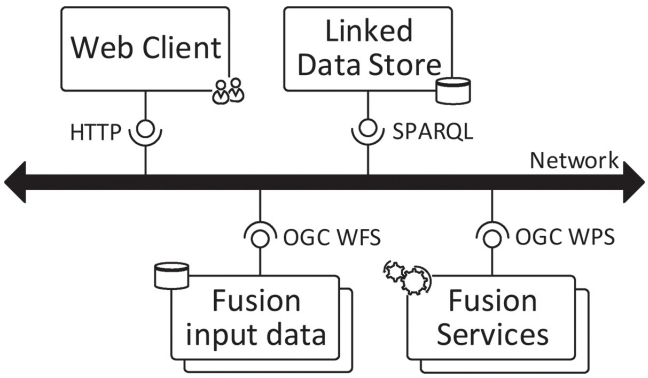


Figura 2.1: Infraestructura de servicios implementada para la fusión de datos espaciales [54].

En la Figura 2.1 se muestra la infraestructura que proponen los autores, la cual contiene 4 componentes: *a) Web Client*, la cual permite la creación y definición de flujos de trabajo de fusión, donde los flujos de trabajo creados comienzan con la selección de las fuentes. Una vez que se ha seleccionado el conjunto de datos de referencia y el conjunto de datos de destino, el sistema proporciona información sobre la posible comparabilidad de los conjuntos de datos mediante el análisis de los sistemas de referencia espacial utilizados y sus extensiones espaciales. Una vez establecidos los parámetros de fusión y haber seleccionado los datos, el componente *b) Fusion input data* recibe y almacena temporalmente los datos a fusionar. Si un usuario desea comparar y combinar los conjuntos de datos mediante procesos de fusión personalizada, se debe proporcionar una instancia de OGC WPS (*OGC Web Processing Service* [10]) adecuada dentro del componente. Posteriormente *c) Fusion Services* es el encargado de aplicar la técnica de fusión de datos definida. Para finalizar, los datos son alojados dentro del servicio de *d) Linked Data Store*, que es un repositorio de datos que

permite ser consumidos mediante un *endpoint*[4] *SPARQL*. En la Figura 2.2 se describen los pasos base establecidos en el flujo de FD, los autores indican que no deben considerarse como una secuencia estricta, sino mas bien como un conjunto procesos independientes que es posible omitirlos, reiterarlos o combinarlos de manera diferente a conveniencia de lo que desea el usuario.

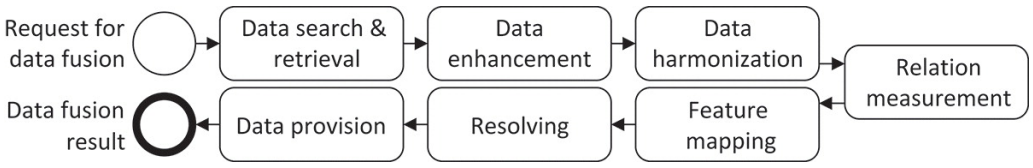


Figura 2.2: Proceso de fusión implementado en [54].

B) Fusión de datos sanitarios mediante orquestación de arquitectura orientada a servicios

En 2011 Venkatesh *et al.* [52] describieron una propuesta de fusión para datos médicos. Si bien este trabajo es está especialmente dirigido al tratamiento de datos espacio-temporales, cuenta con algunos aspectos que se relacionan dentro del tema de tesis propuesto. Los autores proponen una manera de realizar fusión de datos en el dominio de *Smart Healthcare*. La idea surge a partir de la necesidad de analizar datos de sensores utilizados por los médicos pues producen datos que pueden ayudar a mejorar la atención médica y reducir el desencadenamiento de otras enfermedades en los pacientes.

En este proyecto se propone la arquitectura que se muestra en la Figura 2.3, en la cual se consideran tres tipos de unidades de cuidados intensivos (ICU por sus siglas en inglés): pediatría, traumatología y cardíaca, cada una de estas contiene su propio conjunto de datos y tiene asignado

[4]Un *endpoint* es un punto de acceso a un conjunto de datos o servicios para que puedan consumirse siguiente un protocolo determinado en algún lenguaje (XML, JSON, SPARQL, etc.).

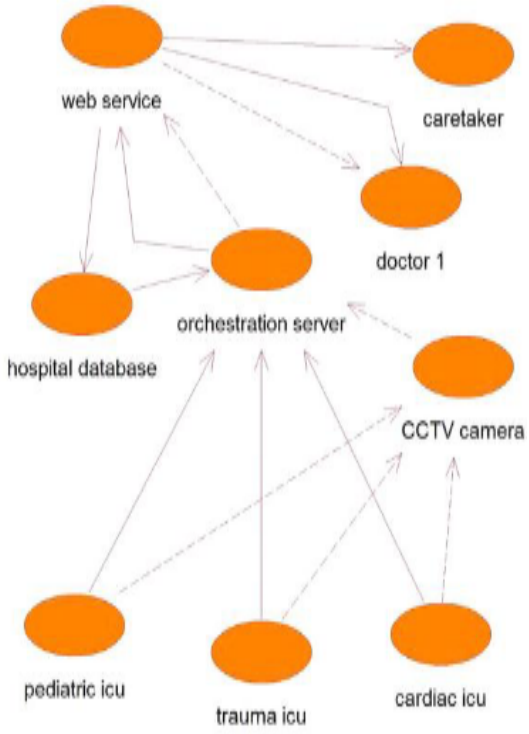


Figura 2.3: Arquitectura de fusión de datos implementada en [52].

un médico especializado. La fusión es necesaria cuando el médico envía información a un servicio web de concentración de datos. Todos los datos de las ICU se envían al servidor de orquestación, donde se almacenan en la base de datos junto con la transmisión de las cámaras CCTV de la ICU correspondiente. El orquestador es el encargado de organizar todos estos datos dependiendo del médico que haya realizado la carga de datos y así fusionarlos con el tipo de ICU.

C) Formalización de fusión de datos espaciales.

En 2017 Wiemann [53] propuso un modelo procesamiento relacionado con la fusión de datos orquestada mediante un servidor web. Esta propuesta se basa en el trabajo de Wiemann y Bernard [54], donde los autores proponen una implementación sobre un servicio de fusión de datos dentro

del área de datos espaciales. La arquitectura propuesta por Wiemann [53] se muestra en la Figura 2.4. A partir de la interacción del usuario con la plataforma web, éste indica aquellas variables que desea fusionar. Esta plataforma tiene servidores a su disposición, los cuales están dedicados al almacenamiento de los datos, proceso de pre-fusión y de fusión de los datos. A partir de esto, siguiendo las configuraciones especificadas por el usuario, se puede generar una fusión de datos espaciales personalizada con las variables que el usuario desee, pero teniendo en cuenta las fuentes que tenga disponibles la plataforma.

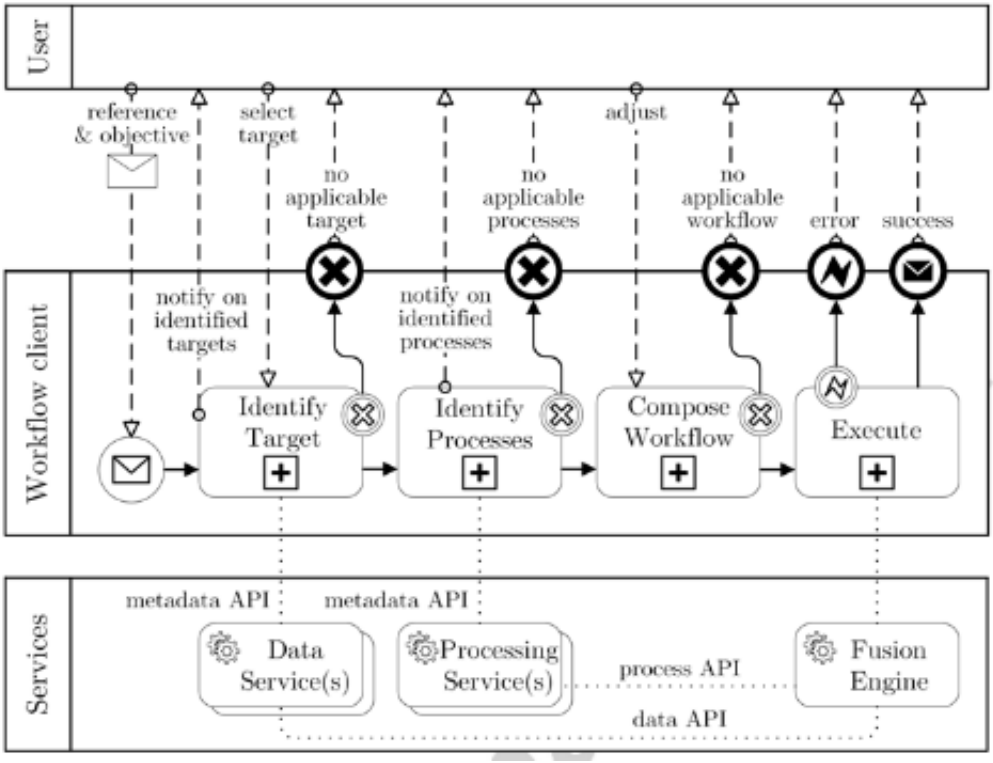


Figura 2.4: Flujo de trabajo y ejecución a través de la notación BPMN.

D) Plataforma de fusión de datos de autoevolución para modelos de agua a gran escala
En 2021 Li *et al.* [34] propusieron una plataforma para la adquisición y asimilación de datos. Estos datos son de aspectos científicos sobre la relación de los factores humanos y climáticos que interactúan y están relacionados con la escasez mundial de agua. Este proyecto implementa un método para abordar la fusión automatizada de datos transferibles, los cuales puedan usarse para simulación y observación del ciclo del agua a gran escala y de alta dimensionalidad. Este enfoque permite la adaptabilidad de los datos a varios formatos, tiene la capacidad de caracterizar la heterogeneidad temporal y espacial de la información y puede representar estructuras complejas para distintos fines científicos.

En la Figura 2.5 se muestra el flujo de trabajo propuesto, el cual va desde la recopilación de datos sin procesar hasta la decisión final de aplicar alguna técnica de *machine learning*. Los conjuntos de datos de diversas fuentes se recopilan, procesan y fusionan con la intersección de variables espacio-temporales, antes de su almacenamiento en un *datalake* basado en la nube. De acuerdo con los autores, la orquestación de datos es realizada en el módulo llamado *Data Managment*, donde internamente se aplican técnicas de minería de datos y consultas para ejecutar los procesos de fusión configurados por el usuario. El usuario es responsable de controlar en la plataforma la preparación de los conjuntos de datos de acuerdo con los objetivos y escenarios de que desee con la técnica de *machine learning* a emplear.

2.2.2 Servicios para fusión de datos

Existen algunos servicios disponibles en la nube que proporcionan servicios de fusión de datos, estos permiten crear flujos de trabajo que el usuario puede personalizar a su manera y así generar sus propios procesos de fusión de datos. Al usar estos servicios el usuario da por hecho que utilizará las

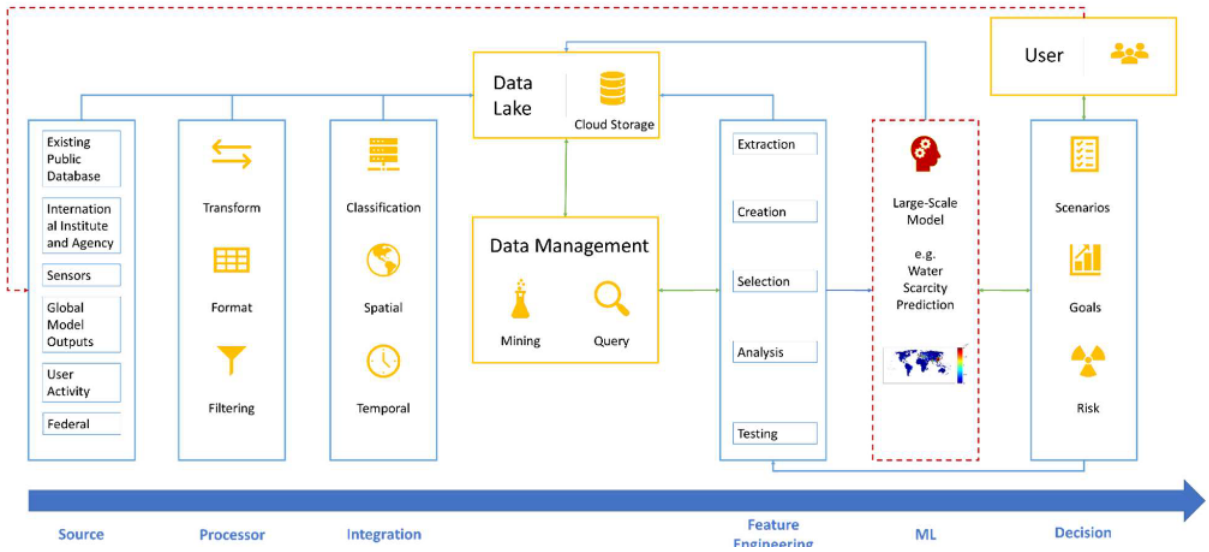


Figura 2.5: Arquitectura implementada para la fusión de datos automatizada en [34].

orquestaciones de datos fijas del proveedor para poder obtener un rendimiento aceptable, pues es el mismo proveedor quien define unos niveles mínimos de operación, lo que da confianza a los usuarios. Entre los servicios disponibles en la actualidad, los servicios más relevantes se describen brevemente a continuación.

Google Cloud Data Fusion

Cloud Data Fusion[5] es una plataforma propuesta por Google [36], la cual ayuda a crear servicios de flujos de procesamiento de datos con el fin de realizar integración de éstos de forma nativa en la nube de Google. Este servicio extiende la tecnología de código abierto de CDAP[6] [25]. El servicio se encarga de gestionar y compilar la integración de fuentes de datos de manera ágil. Para ello esta plataforma ofrece componentes de integración predefinidos que ayudan a los usuarios a agilizar sus procesos de flujo de trabajo. Este servicio hace uso de la tecnología sin servidores (*serverless*[7]), lo que ayuda a los usuarios a evitar los cuellos de botella en los procesos que defina. Esta definición de

[5]Cloud Data Fusion - https://cloud.google.com/data-fusion

[6]CDAP es una plataforma de aplicaciones para construir y gestionar aplicaciones de datos en entornos híbridos y multi-nube.

[7]En este modelo de servicio el usuario no requiere infraestructura propia, todo ejecuta en la infraestructura del proveedor de la nube, donde se realizan todas las tareas. El usuario sólo se encarga de definir lo que desea usar.

procesos la realiza sólo mediante una interfaz web o mediante línea de comandos. Lo que se persigue con este servicio es ayudar a los usuarios en el análisis de grandes volúmenes de datos, dejando la integración de las fuentes de datos a este servicio, incluso permitiendo que se realice la integración con mínimo esfuerzo para escribir los scripts de configuración. Para brindar esa funcionalidad emplea varios de los servicios de Google para la nube (Cloud Engines, Cloud Security, Cloud AI, Storage y Exploration and BI). Si bien el servicio proporciona un amplio abanico de componentes para usar en los procesos del usuario, estos tienen una dependencia completa a los servicios de Google, por lo que se presenta un escenario de *vendor lock-in*.

AWS Glue

Amazon proporciona el servicio *AWS Glue*[8] [50] para la creación de servicios de integración de datos sin necesidad de que el usuario opere servidores (*serverless*). La idea es facilitar el descubrimiento, la preparación y la combinación de datos para su análisis. Esto es como consecuencia de los recursos que ofrece Amazon para el procesamiento de los datos, aprendizaje automático y desarrollo de aplicaciones. Esta plataforma cuenta con todos elementos básicos para la creación de servicios de fusión de datos y que los usuarios puedan acceder fácilmente a sus datos. El modelo de *AWS Glue* está basado con el modelo tradicional ETL[14], con el que los ingenieros de datos y los desarrolladores pueden utilizar *AWS Glue Studio* para crear, ejecutar y supervisar visualmente flujos de trabajo. Asimismo, AWS Glue permite realizar transformaciones en los datos según el esquema y formato de los datos, por lo que en la integración que defina el usuario pueden mezclarse datos estructurados, semi-estructurados y relacionales. Uno de los objetivos que se persiguen con este servicio es que el usuario se enfoque más en los procesos de analítica de datos, en lugar de la integración. Para ellos el servicio de Amazon Glue configura, aprovisiona y gestiona de manera completa todos los recursos de cómputo disponibles en la infraestructura. Esta infraestructura se hace accesible al servicio mediante

[8] https://aws.amazon.com/es/glue

Apache Spark[9], de tal manera que la infraestructura puede escalar de manera rápida y eficiente según las necesidades de la fusión de datos. El funcionamiento de Amazon Glue está sustentado por los servicios de Amazon S3[10] y Redshift[11], aunque también permite la integración de Apache Kafka[12] y MongoDB[13].

Apache Airflow

Apache Airflow[14] [48] propiamente no es una plataforma para fusión de datos pero puede emplearse para ese fin. En realidad es una plataforma para crear, programar y monitorear flujos de trabajo. En estos flujos de trabajo se pueden poner componentes que realicen la fusión de datos, pero es el usuario quien debe organizarlos. En Airflow el usuario puede ejecutar sus tareas mediante una serie de trabajadores (procesos que se ejecutarán de acuerdo a las necesidades del usuario) mientras se cumplan las dependencias especificadas. Es decir, el usuario se debe de apegar a las reglas definidas por Airflow para que su flujo de trabajo ejecute con éxito. La configuración y organización de los componentes de los flujos de trabajo se definen mediante un lenguaje declarativo propio de Airflow. Este lenguaje permite organizar los flujos de trabajo como grafos acíclicos dirigidos (DAG) de tareas [27], con lo que la estructuración y redefinición de tareas es más intuitiva. En el DAG se definen las dependencias entre tareas y el orden de ejecución de las mismas. La idea es que mediante este lenguaje declarativo los flujos de trabajo se realizan mediante procesos independientes. Se cuenta con una interfaz de usuario facilita la visualización de las ejecuciones en producción, con lo cual el usuario podrá monitorear el progreso y solucionar problemas cuando sea necesario, todo en tiempo de ejecución. Airflow permite a los desarrolladores orquestar flujos de trabajo de distinta índole, no necesariamente de fusión de datos. Entre las tareas que se pueden orquestar están las de adquisición,

[9] https://spark.apache.org
[10] https://aws.amazon.com/es/s3
[11] https://aws.amazon.com/es/redshift
[12] https://kafka.apache.org
[13] https://www.mongodb.com
[14] https://airflow.apache.org

extracción, transformación, procesamiento y almacenamiento. Las tareas contenidas en un flujo de trabajo puede a su vez contener muchas otras tareas de manera interna, incluso disparar eventos para la ejecución de otras tareas.

2.2.3 Resumen

En esta sección se presenta una comparación cualitativa de las características funcionales de los métodos de orquestación de datos disponibles en el estado del arte que son lo más parecido al método propuesto. Esta comparativa sigue las características comúnmente evaluadas entre los servicios de orquestación actualmente disponibles [49].

En la Tabla 2.1 se muestran las características de las propuestas y herramientas de los proveedores de servicios de FD. Observando el cumplimiento de las características se puede ver que los diferentes trabajos tienen cierta relación, quizá no completa, con la propuesta de este trabajo de tesis. Si bien el número de características de las propuestas/herramientas es variada, las características mostradas en la tabla son las más comunes entre la mayoría de los trabajos:

1. *Enfoque de desarrollo*: Indica el objetivo con el cual fue desarrollado el trabajo.

2. *Posibilidad de añadir nuevas fuentes de datos*: Si la propuesta/herramienta tiene la capacidad de que los usuarios puedan adicionar nuevas fuentes de datos.

3. *Reducen los escenarios de dependencias*: Si las características de la propuesta/herramienta permiten la reducción de las dependencias *usuario-proveedor* y *fusión-orquestación*.

4. *Estandarización de datos personalizada*: Si la propuesta/herramienta permite adaptaciones personalizadas sobre los datos para normalizarlos.

5. *Permite datos no estructurados*: Si la propuesta/herramienta permite lectura y manejo de los datos no estructurados.

6. *Monitoreo*: Si la propuesta/herramienta permite la visualización de comportamiento de los recursos de cómputo en las diferentes etapas de los flujos de trabajo.

	***Google Cloud*[36]**	***AWS Glue*[50]**	***Airflow*[48]**	**A [54]**	**B [52]**	**C [53]**	**D [34]**
1	Integración de datos	ETL, Integración de datos	Orquestación, *Workflows*	Fusión de datos, Orquestación de *Workflows*	Fusión de datos, Orquestación de *Workflows*	Fusión de datos	Fusión de datos
2	Posible	Posible solo con códigos en Scala o Python	Posible	No Posible	No Posible	No Posible	No especificado
3	Reduce con CDAP [25]	Medianamente Posible	No especificado	No Posible	No Posible	No Posible	Medianamente Posible
4	Posible	Posible	Posible	No posible	No Posible	No Posible	No Posible
5	Posible	Posible	Posible	No posible	No Posible	No Posible	Posible
6	Posible	No, pero posible	Posible	Posible	No Posible	Posible	Medianamente Posible

Tabla 2.1: Comparación cualitativa de características funcionales de los métodos de orquestación de datos disponibles en el estado del arte.

Con base en la comparativa de la tabla se definieron las características que debería cumplir un método de orquestación para fusion de datos como el deseado en el Capítulo 1. De las propuestas y herramientas existentes se identificaron los mejores aspectos y se planteó la integración de éstos en el método propuesto; lo cual, en la medida de lo posible, se fue cumpliendo por partes. Aunado a ello, y como característica a destacar, es que se planteó que el método propuesto mitigara casi por completo, o completamente, en el mejor de los casos, las dependencias *usuario-proveedor* y *fusión-orquestación*. Con ello el usuario tendría amplias posibilidades para definir las tareas de procesamiento/análisis de datos que mejor considerara convenientes, incluso permitiendo la incorporación de otras herramientas (externas) al proceso de fusión.

3

Método de orquestación para servicios de fusión de datos definidos por variables espacio-temporales

En esta sección se describe, el método de la solución propuesta para enfrentar al problema mencionado previamente. La solución propuesta considera el diseño, desarrollo e implantación de un método de orquestación agnóstico de la infraestructura y determinando por esquemas basados en variables espacio-temporales que puedan permitir crear servicios dinámicos de fusión de datos (FD).

Se describen las fases del método de orquestación mediante una representación conceptual, así como el modelo de procesamiento agnóstico para crear servicios dinámicos de FD sobre un entorno de nube. A partir de este modelo también se describen algunos posibles esquemas de despliegue y acoplamiento de servicios de FD que se podrían obtener.

3.1 Descripción general

Como se mencionó previamente, un modelo de FD considera múltiples entradas de fuentes de datos (datos, características y decisiones); que son integrados mediante intersecciones o uniones por un ente *transformador* (proceso de BDA) en un solo resultado que es entregado a un servicio de consumidor (ya sea otro FD, otro BDA o *datawarehouse/datalake*).

Esto crea relaciones recursivas en un servicio de FD que no necesariamente pueden ser modeladas por un sistema ETL tradicional, pues este modelo tradicional considera un solo punto de obtención y un solo punto de entrega.

En cambio, en este trabajo de tesis se propone modelar un servicio de FD como un sistema de compuertas que sirva de guía para realizar la orquestación de los datos. Al emular un sistema de compuertas, el modelo propuesto puede absorber todas las múltiples entradas esperadas de un FD, así como las posibles interconexiones recursivas que se podrían presentar dependiendo de las necesidades de los usuarios finales. Antes de comenzar a explicar este método de orquestación es importante conocer aquellas variables que se ven involucradas, en la Tabla 3.1 se encuentran enlistadas dichas variables con su descripción.

Variable	**Descripción**
FD	Servicios de Fusión de Datos
$\xrightarrow{d_i}$	Transferencia de datos
Id	Identificador
$Path$	Ruta de origen
DS	Fuentes de datos
MDS	Múltiples fuentes de datos
DA	Proceso de análisis de datos
BDA	Conjunto de procesos de DA
DSk	Resumidero de datos
$MDSk$	Múltiples resumideros de datos
Sk	Destino de resultados
$SinkPool$	Conjunto de resumidero de datos

Tabla 3.1: Variables utilizadas dentro del método

Como se puede observar en la Figura 3.1 la construcción del sistema está compuesto por 6 fases, la cuales incluyen el proceso de fusión, división, conversión, consolidación de los datos y por último la visualización, la última fase llamada retorno nos permitiría regresar a alguna de las fases anteriores con el fin de consumir los resultados de alguna de las fases.

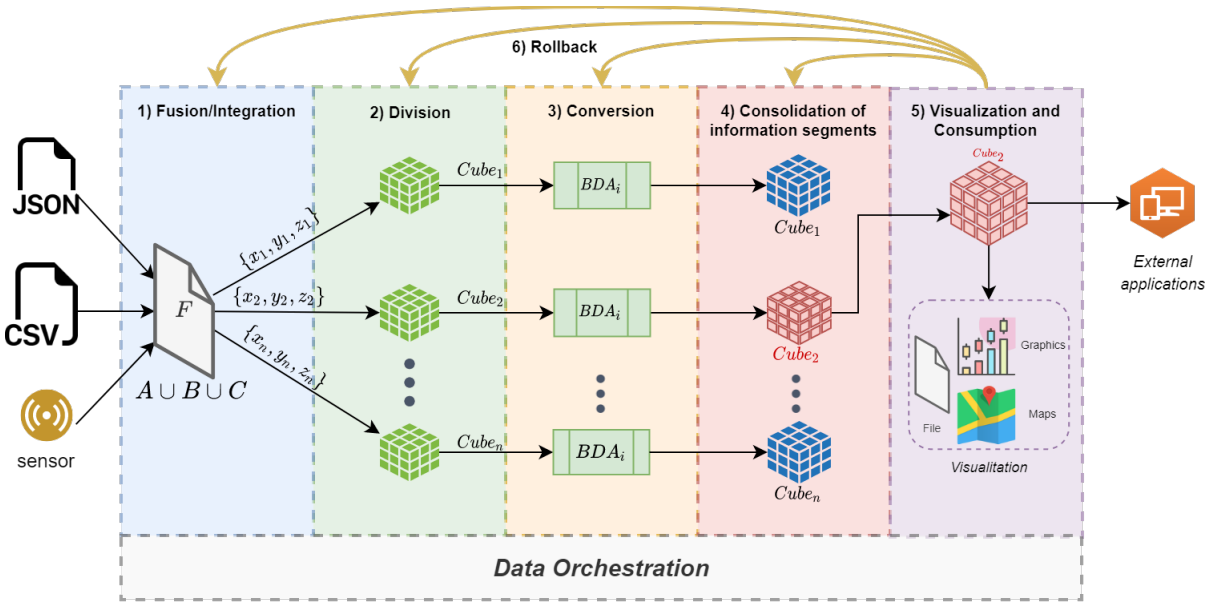


Figura 3.1: Método definido por espacial-temporal

Se ha implementado este método para ejecutar las fases de forma secuencial, asistiendo al usuario en cada fase mediante un esquema declarativo donde el personal define los parámetros necesarios en cada fase. Por lo tanto, los usuarios finales reciben asistencia a través de las etapas y al final de la última fase, se produce un sistema de ciencia de datos diseñado/elegido por los usuarios finales. Una vez definido el sistema, un conjunto de fuentes de datos seleccionadas pasan por todo el proceso de fusión y transformación para finalmente obtener información para la toma de decisiones.

Para la dependencia usuario-proveedor se creó un método de despliegue y acoplamiento de servicios FD que permitirá a los usuarios finales crear estos servicios mediante un esquema de cláusulas declarativas. Como se ha mencionado los servicios de FD comienzan con la entrada de fuentes de datos, una fuente de datos (DS_i)se declara de como se muestra la ecuación 3.1, para que una fuente de datos sea admitida debe de contener un identificador (Id_i) y la ruta ($Path_i$) en

donde se encuentra dicha fuente de datos.

$$DS_i = (Id_i, Path_i) \tag{3.1}$$

Ya teniendo declaradas las fuentes de datos, se genera lo que es un conjunto de múltiples fuentes de datos (MDS), que estas fuentes son las que se estarán trabajando durante el servicio de FD, esta múltiple fuente de datos se declara de la siguiente manera (ver ecuación 3.2), tiene como condición de contar con al menos una fuente de datos para hacer funcionar el servicio de FD.

$$MDS = \{DS_i, DS_{i+1}, \ldots, DS_n\}, (MDS \neq \emptyset) \tag{3.2}$$

Concluyendo la declaración de las fuentes de datos, ya es posible declarar los procesos de analítica, para declarar un proceso de analítica se hace de la siguiente forma (ver ecuación 3.5), de igual manera que una fuente de datos, se debe de declarar el Id y $Path$ de proceso de *Big Data*.

$$DA_j = (Id, Path) \tag{3.3}$$

Con esta declaración de los DA, es posible tener el conjunto de BDA (ecuación 3.4), teniendo como condicional que al menos exista un proceso.

$$BDA = \{DA_j, DA_{j+1}, \ldots, DA_m\}, , (BDA \neq \emptyset) \tag{3.4}$$

Por último es declarar el destino de los resultados obtenidos del conjunto BDA, este parte se

llama resumidero de datos y se expresa de la siguiente manera (ecuación ??):

$$DSk_k = (Id, Path) \tag{3.5}$$

Se pueden tener varios resumideros de datos, los cuales conforman el conjunto de resumidero de datos $MDSk$ (ver ecuación 3.6), de igual manera se espera que este conjunto al menos tenga un DSk.

$$MDSk = \{DSk_k, DSk_{k+1}, \ldots, MDSk_o\} \tag{3.6}$$

Los resumideros anteriores es en caso de que los datos se quieran almacenar para algún posterior uso en el futuro, pero estos datos pueden incluso ser dirigidos a otro proceso de BDA y seguir con el análisis o si bien es posible ser fuente de datos de otro servicio de FD, con esto podemos dar por hecho que tenemos una piscina de destinos ($SinkPool$) que se expresa en la ecuación 3.7, con este conjunto de destinos se debe elegir uno (Sk) para finalizar el proceso del servicio de FD.

$$Sk \in SinkPool, SinkPool = \{BDA \vee FD \vee MDSk\} \tag{3.7}$$

De esta forma, mediante las cláusulas declarativas anteriores, los usuarios finales pueden definir las variables del modelo de despliegue y acoplamiento (ver ecuación 3.8), en donde se expresa el flujo del servicio de FD que va desde las fuentes de datos, pasando por procesos de análisis de datos y terminando en un algún destino (Sk).

$$FD = \{MDS \xrightarrow{D_i} BDA \xrightarrow{R_i} Sk\}_{i=1}^{n} \tag{3.8}$$

Por ejemplo, podemos configurar una cláusula de la siguiente manera

$$FD = \{\{DS_1, DS_2\} \xrightarrow{spatioPatt} SpatioIntersection\} \tag{3.9}$$

significa que la FD se realizará mediante una intersección de la variable espacial (e.g. latitud-longitud), que en este ejemplo se establece como $spatioPatt = (latitudX, longitudY)$. De la misma forma se declara que el conjunto de las fuentes de datos MDS es $\{DS_1, DS_2\}$, donde DS_1 podría tomar el siguiente valor $DS_1 = (Merra, path = /Amazon/MERRA^1)$, y DS_2 podría tomar el valor $DS_2 = (EMAS, path = /google/EMAS^2)$.

El mismo procedimiento aplicaría para el resto de cláusulas, tales como DAs o BDA y los Sk finales. De esta forma, la ubicación de los componentes son definidos por el usuario final mediante este esquema declarativo, con lo cual se podría reducir la dependencia usuario-proveedor. Las fuentes MDS, los BDA y los SK podrían estar en infraestructuras distintas, compartiendo infraestructura o centralizadas en una misma infraestructura.

[1]MERRA (Modern-Era Retrospective Analysis for Research and Applications) es un proyecto satelital de la NASA que proporciona datos meteorológicos de distintas localidades desde 1980 con una semana de retraso a partir del tiempo actual. - `https://gmao.gsfc.nasa.gov/reanalysis/MERRA/`

[2]EMAS es un sistema que produce datos meteorológicos sobre el territorio mexicano a través de antenas. - `https://smn.conagua.gob.mx/es/observando-el-tiempo/estaciones-meteorologicas-automaticas-ema-s`

3.2 Fases del método

La construcción de un sistema de ciencia de datos involucra la participación de múltiples disciplinas (básicamente informática, estadística y el dominio de la ciencia, que en este caso es medicina, fuentes climatológicas)[11] En este contexto, basados solo en los métodos analíticos y de procesamiento, se diseñó un método de 6 fases para que los usuarios finales creen sistemas de ciencia de datos basados en la fusión de múltiples fuentes de datos mediante el uso de variables espacio-temporales (ver estas fases en la Figura 3.1). Las fases comienzan con la entrada de múltiples fuentes de datos, para llagar a la primera llamada fusión donde se realiza la integración de las fuentes, siguiendo a la fase de división donde se segmentan el conjunto de datos en cubos de datos y así llegar a la fase de conversión en donde se transforman los cubos de datos, después la fase de consolidación en donde se analiza los resultados de cada segmento, ya por último es la fase de visualización que a partir de los cubos de datos consolidados generar activos de información. La fase de retroceso nos ayuda a regresar a una fase de interés.

3.2.1 Fusión/Integración

La primera fase considera la identificación y selección de fuentes de datos (MDS) para ser procesados en un flujo de trabajo analítico. Este proceso es asistido por un lenguaje declarativo que solicita a los usuarios finales que declaren la ubicación de las fuentes de datos (la ruta del directorio o la ubicación de la nube donde se almacenan, un disco, una URL, una API de sensores, etc.) así como el tipo de fusión de datos a implementar (cualquiera de unión, adenda o intersección). Luego, se solicita a los usuarios finales que seleccionen las variables comunes en las fuentes de datos para fusionarlas en el flujo de trabajo. En esta tesis se están utilizando las variables *espacio-temporales* como variables comunes en las fuentes de datos, pero los usuarios finales pueden decidir qué variables son las más adecuadas para realizar la fusión de datos. Por ejemplo, si utilizamos bases de datos de estudios de salud, como el recuento de suicidios y el acceso a la atención médica, que se fusionaron

mediante el uso de variables espacio-temporales como el año, los ID de los estados, los ID de las ciudades, etc.

El resultado de esta fase es una nueva fuente de datos fusionada (F), que será indexada y utilizada en las siguientes fases del proceso analítico por los módulos de orquestación de datos y tareas.

Para llevar a cabo esta fase de fusión de los datos, se ejecuta el siguiente algoritmo en donde a partir de las declaraciones el usuario, se es posible fusionar las fuentes, este algoritmo va iterando fuente por fuente con el fin de juntarlas según las declaraciones, ya que las fuentes pueden tener diferentes tipos de fusión.

Algorithm 1 Algoritmo dedicado fusionar las fuentes de datos

Entrada: $MDS = \{DS_1, \ldots, DS_n\}, COLS = \{col_1, \ldots, col_n\}, COND = \{cond_1, \ldots, cond_n\}$
Salida: F

```
para pos ← 0, n hacer
  si pos != 0 entonces
    si COND[pos] = interseccion entonces
      F ← fusionInterseccion(DS_aux, col_aux, MDS[pos], COLS[pos])
    si no
      F ← fusionUnion(DS_aux, col_aux, MDS[pos], COLS[pos])
    fin si
  si no
    DS_aux ← MDS[pos]
    col_aux ← COLS[pos]
  fin si
fin para
devolver F =0
```

3.2.2 División de datos de la fuente fusionada

En esta fase, el sistema de ciencia de datos está lidiando con una fuente de datos fusionada (ver F en la figura 3.1), que se espera que sea más grande que una sola fuente de datos, lo que producirá problemas de eficiencia (en un escenario de grandes cúmulos de datos). Para hacer frente a este problema, los datos fusionados se segmentan en subconjuntos de datos n que se crean en forma de cubos de datos.

Esto representa una reducción de la dimensionalidad de los datos enfocada a reducir el espacio de procesamiento que observarán las próximas fases de este método, lo que se espera reduzca el tiempo empleado por las tareas de procesamiento. Se espera que esto mejore la experiencia de servicio de los usuarios finales. Los usuarios finales definen las variables (X,Y,Z) de estos cubos mediante un esquema declarativo (ver Figura 3.2). Por ejemplo, X podría ser un valor espacial, Y un valor temporal y Z una variable de interés determinada, incluso pueden varias estas asignaciones dependiendo la declaración del usuario.

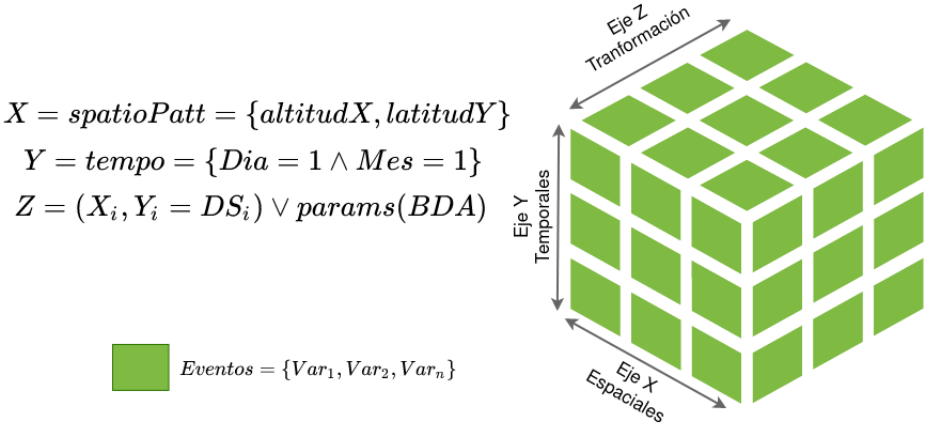


Figura 3.2: Cubo espacio-temporal.

Siguiendo el ejemplo anterior de la fuente de datos fusionada, la X era el año, la Y era un estado de México y la Z era una etiqueta de género. Como resultado, esta selección producirá tantos cubos de datos como X veces Y seleccionadas por los usuarios finales, lo que producirá resultados Z para cada valor X e Y. La figura 4.6 muestra que F se divide en n partes, donde cada parte es un subconjunto con un cierto valor para X, Y y Z. En este contexto, un usuario final puede definir un FD de la siguiente forma:

$$Cube_{v1} = (X, Y, Z) =$$

$$(X = spatioPatt = (latitud_X, longitud_Y \in DS_1,$$

$$Y = tempo = (dia = 1 \wedge mes = 1) \in DS_1,$$
$$Z \in DS_2 = (X, Y \in DS_1)$$

En muchos escenarios, las variables *espacio-temporales* en el conjunto de datos permiten a los analistas obtener diferentes puntos de vista de los datos de diferentes eventos de interés producidos en un espacio determinado y/o en un momento determinado. Estos estudios espacio-temporales son bastante útiles en el proceso de toma de decisiones. Por ejemplo, siguiendo con el ejemplo anterior, los cubos de datos representarán un subconjunto de los conjuntos de datos originales, divididos por año, estado y etiqueta de género.

La selección de variables en la fase anterior creará una división del conjunto de datos en forma de cubos de datos de búsqueda, que también crearán una matriz para cada valor de estas variables elegidas. Siendo el eje X todos aquellos valores de una variable espacial, y el eje Y los valores de una variable temporal.

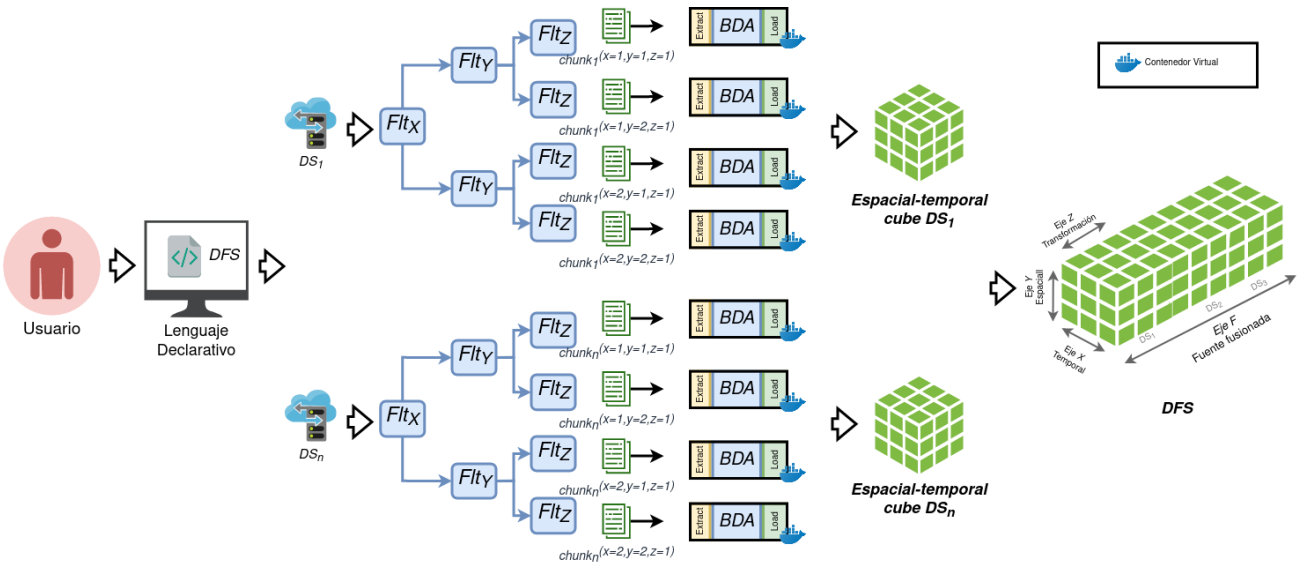


Figura 3.3: Representación conceptual de la solución propuesta.

En la fase de división existen dos tipos principales, temporal y espacial respectivamente, ambas basándose la función de balanceo de carga, esta consta con tres tipos algoritmos que puede utilizar el usuario para distribuir los cubos entre los trabajadores, el prototipo consta de tres balanceadores

disponibles: *Round Robin, Two Choices* y *PseudoRandom* (ver algoritmo).

Algorithm 2 Algoritmo de balanceadores de carga

Entrada: $Cubes, algorithmBalancer, workers$
Salida: $CubesBalanced$
1: $CubesBalanced = []$
2: **si** $algorithmBalancer ==' RR'$ **entonces**
3: $\quad CubesBalanced \leftarrow RoundRobin(Cubes, workers)$
4: **si no, si** $algorithmBalancer ==' PR'$ **entonces**
5: $\quad CubesBalanced \leftarrow PseudoRandom(Cubes, workers)$
6: **si no, si** $algorithmBalancer ==' TC'$ **entonces**
7: $\quad CubesBalanced \leftarrow TwoChoices(Cubes, workers)$
8: **si no**
9: $\quad CubesBalanced \leftarrow RoundRobin(Cubes, workers)$
10: **fin si**
11: **devolver** $CubesBalanced$
=0

El algoritmo 3 indica como es que se hace la división mediante espacial. Parte de la elección de la forma de balancear en la cual consiste si es por estado o municipio del país, en seguida de esto comienza a extraer los índices de los registros pertenecientes a la elección del espacial, para así llegar a la última función dedicada a realizar el balanceo de carga (algoritmo **??**) según el algoritmo a elección del usuario.

Por otro lado, el algoritmo 4 está dedicado a realizar la división mediante temporal. Esta división es similar a la espacial, solo que el listado de espacial cambia por el listado de rangos de fechas, es decir, el usuario pude elegir el tamaño de los subconjuntos de acuerdo a la fecha de inicio y fin (por ejemplo, rangos de fecha por cada dos años o rangos de fecha por cada 10 días, etc). Una vez teniendo este listado ya es posible identificar en las fuentes de datos sus segmentos (cubos de datos) y para finalizar realizar el balanceo de carga entre los trabajadores.

3.2.3 Conversión de datos en información útil segmentada

Esta fase recibe los cubos de datos y crea automáticamente un *transformador* para procesar cada cubo de datos. Un *transformador* encapsula servicios de análisis de Big Data (BDA) en microservicios

Algorithm 3 Algoritmo de división en espacial

Entrada: $MDS = \{DS_1, \ldots, DS_n\}, typeSpatial, typeBalance, algorithmBalancer, workers$
Salida: $CubesBalanced$

$Cubes = []$
si $typeSpatial =' state'$ **entonces**
  $spatial \leftarrow getList(typeSpatial =' state')$
si no, si $typeSpatial =' cities'$ **entonces**
  $spatial \leftarrow getList(typeSpatial =' cities')$
fin si
$lenSpatial \leftarrow len(spatial)$: longitud del arreglo de rangos
$lenMDS \leftarrow len(MDS)$: longitud del arreglo de fuentes de datos
▷ `Extraemos los registros de cada fuente por cada espacial (Segmentación)`
para $posSpatial \leftarrow 0, lenSpatial$ **hacer**
  $Cubes_{aux} = []$
  para $posDS \leftarrow 0, lenMDS$ **hacer**
    $mask = (DS[posDS].Spatial = spatial[posSpatial])$
    ▷ `Extramos los indices que cummplan la condición`
    $DS_{aux} \leftarrow getRecordsByDS(DS = DS[posDS], Index = mask)$
    $Cubes_{aux}[posDS] = DS_{aux}$
  fin para
  $Cubes[posRange] = Cubes_{aux}$
fin para
▷ `Balanceamos los` $Cubes$ `entre los trabajadores con un algoritmo de balanceo`
$CubesBalanced \leftarrow balanceCubes(algorithmBalancer, Cubes, workers)$
devolver $CubesBalanced$ =0

que pueden enviar consultas a un cubo de datos dado y convierte los resultados de esa consulta en información (ver figura 3.1 como un BDA_i recibe un $Cube_n$ y produce $Info_{Cube_n}$). Tenga en cuenta que un microservicio BDA es una pieza de software independiente que incluye todas las dependencias necesarias para ejecutarse en una computadora elegida por los usuarios finales.

Además, los usuarios finales pueden crear versiones de BDA creando una dimensión adicional definida por un criterio personalizado adicional. Por lo tanto, definimos una entidad llamada *Filtro* (Flt), que es una variación del BDA original, que producirá un nuevo $Info_{Cube_n}$. Por ejemplo, si un BDA implementa un algoritmo de agrupación (por ejemplo, $DBS_{\{}k - means\}$), el usuario final puede crear dos versiones del mismo para crear dos tareas diferentes (por ejemplo, Flt_1=$BDA_{\{}k - means\}$, $k = 1$ y Flt_1=$BDA_{\{}k - means\}$, $k = 2$). Una característica interesante

Algorithm 4 Algoritmo de división en temporal

Entrada: $MDS = \{DS_1, \ldots, DS_n\}, COLS = \{col_1, \ldots, col_n\}, COND = \{cond_1, \ldots, cond_n\}$ $startDate, endDate, sizeRange, typeRange, typeBalance, algorithmBalancer$

Salida: $CubesBalanced$

1: $ranges \leftarrow generateRanges(startDate, endDate, typeRaange, sizeRange)$: generamos las fechas correspondientes a cada rango
2: $lenRanges \leftarrow len(ranges)$: longitud del arreglo de rangos
3: $lenMDS \leftarrow len(MDS)$: longitud del arreglo de fuentes de datos
4: $Cubes = []$
▷ `Extraemos los registros de cada fuente por cada rango de fecha (Segmentación)`
5: **para** $posRange \leftarrow 0, lenRanges$ **hacer**
6: $Cubes_{aux} = []$
7: **para** $posDS \leftarrow 0, lenMDS$ **hacer**
8: **si** $posRanges = 0$ **entonces**
9: $mask = DS[posDS].Dates => ranges[0] \ \& \ DS[posDS].Dates <= ranges[posDS + 1]$
10: **si no**
11: $mask = DS[posDS].Dates > ranges[posDS] \ \& \ DS[posDS].Dates <= ranges[posDS + 1]$
12: **fin si**
13: $DS_{aux} \leftarrow getRecordsByDS(DS[posDS], mask)$
14: $Cubes_{aux}[posDS] = DS_{aux}$
15: **fin para**
16: $Cubes[posRange] = Cubes_{aux}$
17: **fin para**
▷ `Balanceamos los` $Cubes$ `entre los trabajadores con un algoritmo de balanceo`
18: $cubesBalanced \leftarrow balanceCubes(algorithmBalancer, Cubes, workers)$
19: **devolver** $CubesBalanced$ =0

de este modelo de procesamiento es que cada $Info_{Cube_n}$ también se administra como otro cubo, que se entrega a la siguiente fase.

Para esta fase, cada proceso de analítica mantiene una estructura general en donde influye la lectura, procesamiento y entrega de los cubos de datos, de manera general el algoritmo 5 indica como es que estos micros servicios trabajan, donde la función $procesoDA$, corresponde al algoritmo de analítica seleccionado por el usuario.

Algorithm 5 Algoritmo de procesos de analítica

Entrada: $Cubes, cols, params$
Salida: $CubesDA$
1: $CubesDA = []$
2: $lenCubesS \leftarrow len(Cubes)$: longitud del arreglo de fuentes de datos
3: **para** $posDS \leftarrow 0, lenCubes$ **hacer**
4: $colDS = cols[posDS]$DS $\leftarrow Cubes[posDS][colDS$
5: $CubesDA[posDS] \leftarrow procesoDA(DS, params[posDS])$
7: **fin para**
8: **devolver** $CubesDA$ =0

3.2.4 Consolidación de información segmentada

Una vez transformados los cubos de datos, corresponde al usuario tomar una primera decisión: "¿qué segmento de información me interesa?". En este contexto, el usuario puede optar por analizar cada uno de los resultados de cada segmento, para lo cual son enviados a herramientas de visualización y consumo (ver apartado 3.2.5). De esta forma, el usuario final puede comenzar a generar conocimiento que puede desencadenar en una decisión. Por ejemplo, si el usuario final detecta resultados sobresalientes para un año específico (una variable temporal), puede optar por explorar este segmento con otras herramientas BDA (ver sección 3.2.6).

3.2.5 Visualización y consumo

En esta fase se envían los $InfoCubos$, los cuales son automáticamente indexados y puestos a disposición para el consumo de los tomadores de decisiones a través de un repositorio. Los usuarios finales pueden elegir estos $InfoCubes$ para extraer información para crear activos de información (por ejemplo, mapas, gráficos o informes), que se producen automáticamente en esta fase de acuerdo a las selecciones del usuario.

Al igual que en la fase de Conversión de datos (sección 3.2.3), de manera general, para cada activo de información se mantiene una estructura que solo depende de la elección del usuario, si es un mapa, gráfica, etc. El algoritmo 6 se puede observar que apartir de los cubos de datos procesados,

es posible gnerar estos activos llendo de la mano con sus parametros de entrada.

Algorithm 6 Algoritmo de procesos de visualización

Entrada: $CubesDA, cols, params$
Salida: $CubesDA, plotDA$

1: $plotDA = []$
2: $lenCubesS \leftarrow len(Cubes)$: longitud del arreglo de fuentes de datos
3: **para** $posDS \leftarrow 0, lenCubes$ **hacer**
4: $colDS = cols[posDS]$DS $\leftarrow Cubes[posDS][colDS$
5: $plotDA[posDS] \leftarrow plotDA(DS, params[posDS])$
▷ `La funcion de guardar el gráfco depende del proceso de analítica ejecutado`
7: $savePlot(plotDA)$
8: **fin para**
9: **devolver** $CubesDA$ =0

3.2.6 Retroceder a una fase previa

Los usuarios finales pueden optar por ver los resultados y consumirlos para respaldar un proceso de toma de decisiones o regresar a las etapas anteriores mediante un proceso de reversión.

Esto significa que el sistema de ciencia de datos puede procesar datos de forma recursiva. En tal escenario, un usuario final puede decidir procesar un *Infocube* para fusionarlo con otros $InfoCubes$ (en la fase 1) o segmentarlo (en la fase 2) o convertirlo en nueva información (por ejemplo, enviar un $InfoCubes$ a la Fase 3 para ser procesada, por ejemplo, por una red neuronal).

3.3 Implementación

En esta sección se describirá la infraestructura utilizada, describiendo el equipo utilizado para el desarrollo del prototipo, también se expondrán los equipos de experimentación del prototipo y por último las métricas usada para medir el rendimiento de los servicios generados.

Cabe resaltar que el prototipo fue desarrollado en el lenguaje de programación Python, ya que es de gran ayuda para la implementación de los algoritmos de analítica. Por otro lado, se utilizó, la

plataforma de Docker, con el fin de desplegar estas aplicaciones en forma de contenedores virtuales de manera distribuida en cualquiera de los computadores disponibles.

3.3.1 Infraestructura

La infraestructura que se piensa utilizar para el desarrollo de este trabajo de tesis, así como la infraestructura de experimentación, se presentan a continuación.

Equipo para desarrollo

- Procesador: Intel Core i5-6200U
- Memoria: 8GB
- Disco Duro: 1 TB
- Cores: 2 físicos y 2 virtuales

Equipo para experimentación

Hostname	Sockets	Cores por Socket	Threads por Socket	RAM
Compute6	1	6	2	64GB
Compute8	2	8	1	64GB
Compute9	1	12	1	64GB
Compute10	1	12	1	64GB
Compute11	1	12	1	64GB
Compute12	1	12	1	64GB

Tabla 3.2: Equipo de cómputo que será utilizado para desplegar y evaluar los experimentos.

3.3.2 Métricas

Para poder medir el rendimiento del prototipo se dividió en tres, pero de manera general se estuvo midiendo los tiempos de servicio en donde se involucran las tareas de lectura, procesamiento y distribución. Los tiempos de servicio medidos fueron los siguientes:

- Tiempos de servicio de filtrado: Medición del tiempo de servicio en la división de las fuentes de datos en cubos.
- Tiempos de servicio de procesamiento: Medición de los tiempos en los procesos de analítica.
- Tiempos de servicio totales: Medición de los tiempos totales.

Durante los tiempos de servicio medidos, se expone una solución original (SO), esta conformada por los procesos de analítica sin realizar ningún proceso de división, es una de las formas de comparación para poder ver a detalle el rendimiento del prototipo conforme crecen los trabajadores.

4

Evaluación experimental y resultados

En este capítulo se describen tres estudios de caso y sus resultados obtenidos dentro del prototipo basado en el modelo de orquestación de datos propuesto en esta tesis, el primer estudio de caso se realiza en un contexto de datos climatológicos, el segundo caso corresponde a datos de defunciones y servicios médicos y para el último estudio se realizó con datos sintéticos.

4.1 Estudio de caso 1 - Datos meteorológicos

En esta primera evaluación del prototipo, se realizó una experimentación con dos fuentes de datos meteorológicos llamadas EMAS y MERRA, en donde ambos conjuntos de datos cuentan con variables espacio-temporales, este experimento se realizó con el fin de encontrar microclimas [1] dentro del territorio mexicano a través de los años.

Comenzando con las descripciones de las fuentes de datos, MERRA es un proyecto de la NASA

[1] Se llama microclima al clima que presenta unas características diferentes a las del resto de la zona en donde se encuentra.

donde refleja los avances recientes en el modelado atmosférico y la asimilación de datos, obteniendo las observaciones mediante satélites [22]. Este conjunto de datos cuenta con las variables presentadas dentro de la Tabla 4.1 después de realizar un preprocesamiento; cabe mencionar que solo se están seleccionando aquellos registros que se encuentran dentro del territorio mexicano.

Variable	Descripción
$antena$	Id/nombre de la antena
lat	Latitud de la antena
lon	Longitud de la antena
$date$	Fecha de registro
$HOURNORAIN$	Horas sin lluvia
$T2MMAX$	Temperatura máxima
$T2MMEAN$	Temperatura media
$T2MMIN$	Temperatura mínima
$TPRECMAX$	Precipitación máxima

Tabla 4.1: Variables de la fuente MERRA

Por otro lado, EMAS es un sistema de antenas distribuidas por el territorio mexicano, las cuales a través de sensores de medición, dispositivos eléctricos, electrónicos y mecánicos, montados sobre una estructura de soporte donde capturan datos meteorológicos. Las variables resultantes después de un preprocesamiento fueron las siguientes (ver Tabla 4.2):

Variable	Descripción
$estacion$	Id/nombre de la estación
$latitud$	Latitud de la estación
$longuitud$	Longitud de la estación
$fecha$	Fecha de registro
$Temperatura$	Temperatura máxima
$RapidezViento$	Rapidez del Viento (km/h)
$Humedad$	Humedad relativa
$Precipitacin$	Precipitación

Tabla 4.2: Variables de la fuente EMAS

4.1.1 Representación conceptual

Para este caso de estudio se diseñó un FD en donde al fusionar ambas fuentes de datos se aplicaría un *clustering* (Cl) con el fin de encontrar algún microclima dentro de las zonas de México. Esto se llevó a cabo realizando un filtrado (Flt_X) por año dentro para las dos fuentes (X), posteriormente ya que se cuenta con los fragmentos de años se realizó una fusión de estas fuentes en donde el punto de intersección serían las variables $fecha$ (X) por el lado del temporal y por el lado de espacial sería $latitud$ y $longuitud$ (eje Y). Una vez que sé obtuvo el archivo fusionado (cubo de datos) correspondientes a una fecha y un espacio, por ende este cubo se puede representar de la siguiente manera, $cube^{x=2000}$ donde el cubo está almacenando los datos correspondientes a un año en específico y por el lado de espacial (Y) todos los registros del territorio mexicano, es decir, no se realizó alguna división por estado o municipio de México con el fin de proyectar estos registros en un mapa del país.

Una vez terminado la fusión de las fuentes, el siguiente proceso de BDA es realizar un *clustering* con diferentes parámetros, para esto el eje Z del cubo se representó de la siguiente manera: $Z = [K = 3, K = 4, K = 5]$, tomando como base un algoritmo de $k - means$. Por último, al terminar este proceso de BDA se realiza un proceso de mapeo de los resultados donde se pueden observar la clasificación obtenida por el proceso de *clustering* y como se distribuye dentro del país.

La representación final de todo el flujo de trabajo de este experimento lo podemos observar en la Figura 4.1.

Las pruebas realizadas en este estudio de caso consisten en incrementar el número de trabajadores de procesamiento (fusión de datos y *clustering*), es decir, una vez que termine el proceso de Flt_X y se encuentre segmentados los cubos, se están distribuyendo en un número de trabajadores de uno a veinte nodos de procesamiento, donde en la última prueba de 20 trabajadores estaría dedicada a solo año (cada nodo trabajaría con un cubo de datos). tomando en cuenta que la solución original (SO) con la cual se está comparando consta en realizar los procesos de fusión de datos y *clustering*)

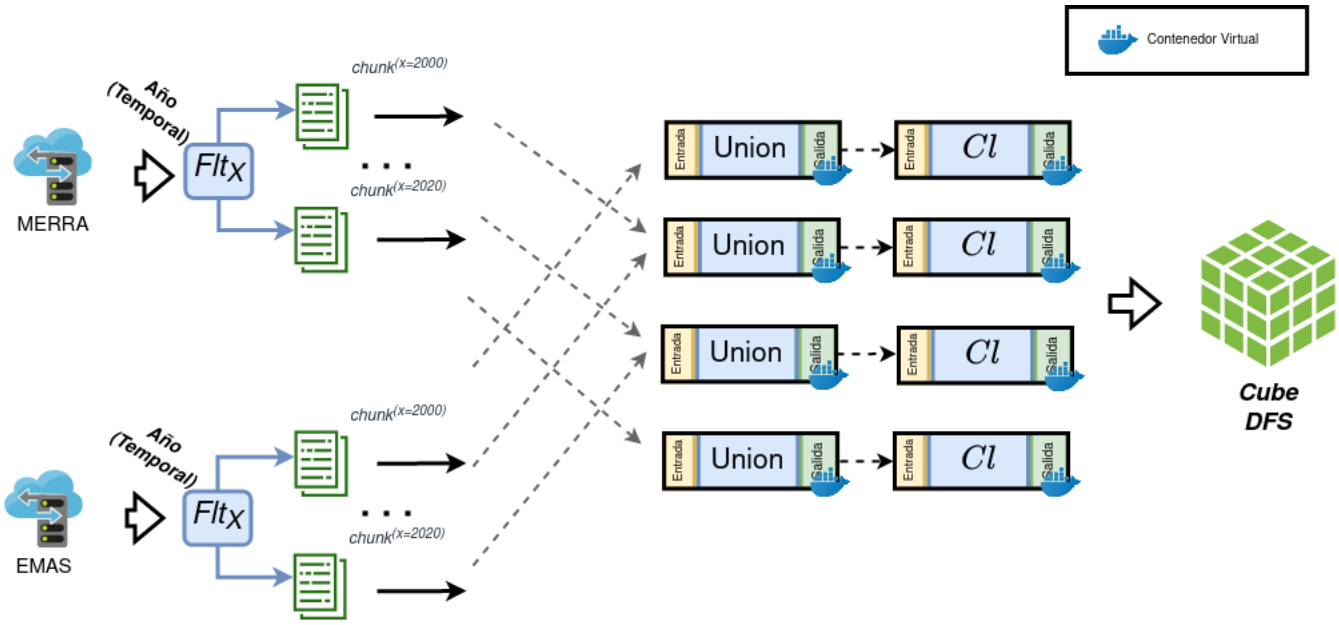


Figura 4.1: Representación conceptual del experimento 1

sin tener que realizar la segmentación de los datos.

En total se registraron 6 pruebas incrementando el número de trabajadores, estas pruebas se describen en la Tabla 4.3.

	Cantidad de trabajadores por prueba						
Nombre	$S.O.$	1	2	3	4	5	6
Temporal (Año)	0	1	1	1	1	1	1
Fusión(Unión)	1	1	2	5	10	15	20
Custering	1	1	2	5	10	15	20

Tabla 4.3: Pruebas realizadas

4.1.2 Tiempos de servicio del filtrado de datos

En esta sección se encuentran los tiempos obtenidos durante el filtrado del temporal sobre las fuentes de datos. El filtrado de temporal por años se aplicó un proceso de segmentación y balanceo de carga, el algoritmo utilizado para balancear la carga entre los nodos destino fue *TwoChoices* como condicional de balanceo es la cantidad de registros de cada cubo, por último se efectúa la distribución

de los cubos de datos a sus trabajadores destino. El conjunto de ambos datos cuenta con registros entre los años 2000 al 2020, donde los tiempos de servicio obtenidos en esta fase fueron los siguientes (ver Figura 4.2):

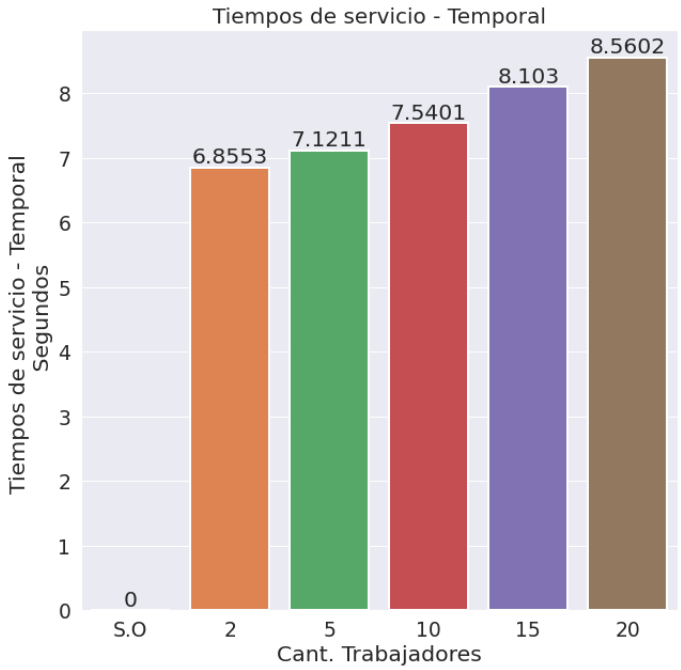


Figura 4.2: Tiempos de servicio del filtrado temporal

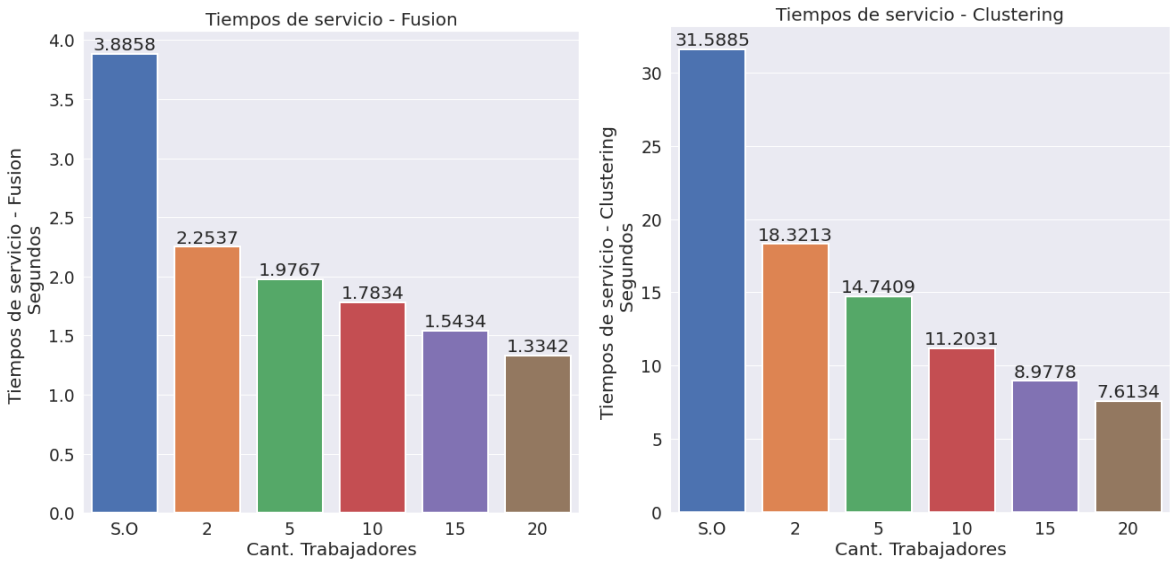
Se puede observar un comportamiento en donde los tiempos aumentan conforme los trabajadores aumentan, eso se esperaba debido a que los la labor del proceso Flt_X estaría filtrando, balanceando y distribuyendo los segmentos generados con cada vez más trabajadores, el crecimiento en los tiempos está en un aproximando de 25 % en el filtrado por años. Tomando en cuenta que en SO no se está aplicando este proceso.

4.1.3 Tiempos de servicio de fusión de datos y clustering

Dentro de esta sección se mostrarán los tiempos de servicio correspondientes a los procesos de analítica, en donde se mencionó la utilización de la fusión de datos siguiendo un proceso de *clustering* (Cl).

Como se han fragmentado los datos, una de las ventajas es que al momento de que el prototipo crea un cubo de datos, este es enviado al siguiente nodo con el fin de estar adelantando los procesos. Esto no afecta debido a que cada cubo de datos es independiente de los demás. Es por eso que se emplean múltiples trabajadores para realizar dichos procesos en paralelo.

Como se puede observar en la Figura 4.3 se puede observar una disminución en los tiempos de servicio en ambos procesos. Esta disminución se percibde de forma exponencial tanto en la fusión de datos (Figura 4.3(a)) como en el clustering (Figura 4.3(b)).



((a)) Fusión de datos (Unión) ((b)) Clustering

Figura 4.3: Tiempos de servicio en fusión y clustering

Dentro del proceso de fusión de datos se obtiene una mejoría del 65 % entre la solución original y el incremento de los trabajadores. Por otro lado, el proceso de *clustering* en donde se aplicó el mismo algoritmo variando el valor de k y posteriormente realizar un mapeo de los valores obtenidos por cada k, en esta fase se obtiene una mejoría del 76 % con respecto a la SO.

4.1.4 Tiempos de servicios totales

En esta última sección del estudio de caso, mostramos los tiempos de servicio totales de todo el servicio de FD. En la Figura 4.4 se puede observar que al momento de multiplicar el número de trabajadores, se observa una disminución general de los tiempos de forma exponencial.

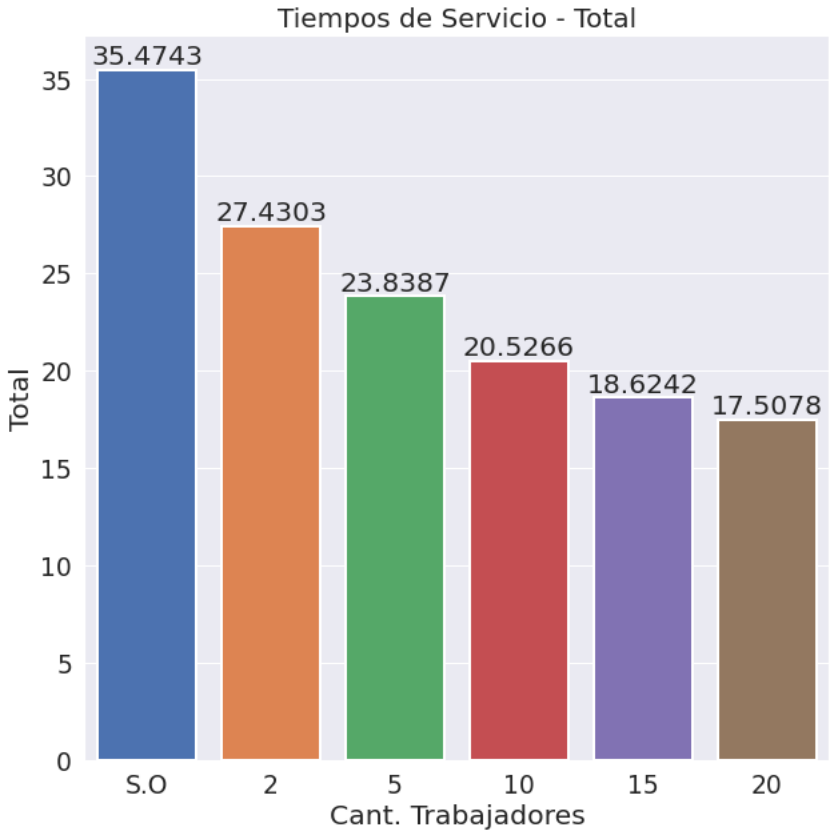
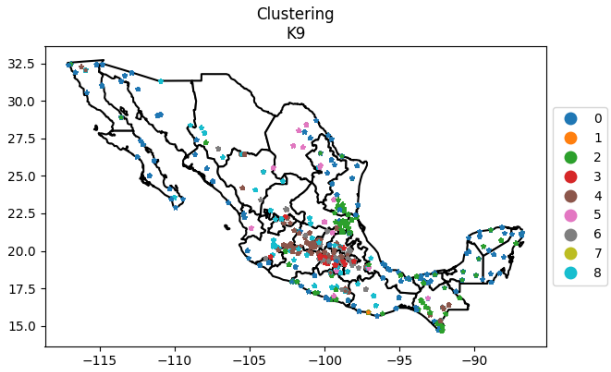


Figura 4.4: Suma de tiempos de servicio

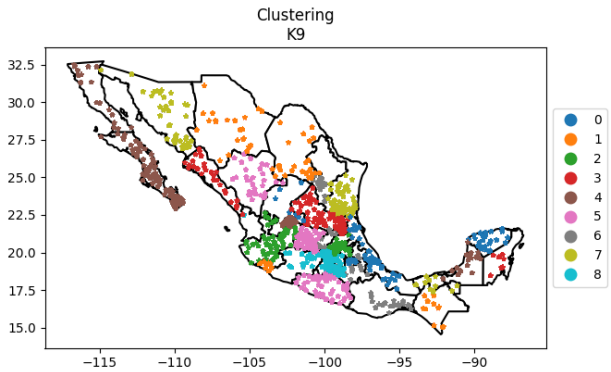
Por último, como resultado obtenemos conjunto de datos reducidos del total de ambas fuentes de datos ya procesados y así mismo imágenes del territorio mexicano con los resultados obtenidos del proceso del *clustering*.

Un ejemplo de los resultados obtenidos en el *clustering* lo podemos observar en la Figura 4.5. En donde claramente la diferencia de cantidades de registros es notoria a través de los años. Enfocándose dentro de la obtención de microclimas de país para el año 2000 (Figura 4.5(a)) no se perciben

microclimas con claridad, a excepción del grupo 3 se ubica sobre el centro del país, y unas semejanzas a este clima es sobre la costa de Jalisco donde espera que sea un clima semejante a algún grupo que se ubique por la orilla del país. Por otro lado, para el año 2020 (Figura 4.5(b))se puede observar que por el lado de Veracruz (grupo 0), en donde podemos catalogar un clima relacionado con las costas o playas, donde se espera que estos climas tengan temperaturas más altas que el centro del país, pero ocurre un hecho por la zona norte de Zacatecas en donde se presenta un clima relacionado con las costas de Veracruz, es decir, se puede dar por hecho que las temperaturas de esa zona han ido en aumento.



((a)) Año 2000



((b)) Año 2020

Figura 4.5: Resultados de mapeo del clustering

4.2 Estudio de caso 2 - Datos poblacionales

Para una segunda evaluación del prototipo, llevamos a cabo un estudio de caso con 2 conjuntos de datos reales. La primera fuente de datos cuenta con conteos de defunciones de personas con algún trastorno mental y que además consumían sustancias psicotrópicas, para diferentes entidades de México. El segundo conjunto de datos tiene datos estadísticos sobre personas con acceso a servicios médicos.

Dentro de la Tabla A.1 en el Anexo A.1.1, se encuentran descritas las variables que conforman las fuentes de datos, esta fuente tiene un tamaño de 15MB con un total de registros de 51,834, estos registros se encuentran en el rango de los años 2000 al 2020.

Las variables que conforman la fuente de datos macroeconómica la podemos observar en la Tabla A.1 en el Anexo A.1.2, esta fuente contiene datos en el mismo rango de fechas que la fuente de defunciones (año 2000 al 2020), consta de 12,305 registros y con un tamaño de 6 MB.

4.2.1 Representación conceptual

Para este caso de estudio se diseñó un FD para la comparación basado en una correlación de ambas fuentes de datos, filtrando por el sexo de la fuente de datos de defunciones (X), el espacial (Y), las variables de la fuente de datos de servicios médicos (Z). Posteriormente, cada fragmento pasó a un servicio que realiza la función de unión entre los fragmentos (cubos de datos) correspondientes de cada conjunto de datos, es decir, el $cube^{(x=1,y=1)}$ del conjunto de datos de muertes se procesó junto con el $cube^{(y=1,z=1)}$ del conjunto de datos médicos. Finalmente, a partir del resultado se realiza el cálculo del valor de correlación basándose en el coeficiente de Pearson (ver Figura 4.6). Para este caso, no necesitamos usar una variable temporal para realizar la fusión de datos, debido a que la fuente de datos sobre servicios médicos no cuenta con esta variable; por lo que el cubo resultante solo tiene 3 dimensiones, siendo las X las variables de un conjunto de datos contra las del otro conjunto de datos en Y, en conjunto por en un sector espacial (en este caso, Z).

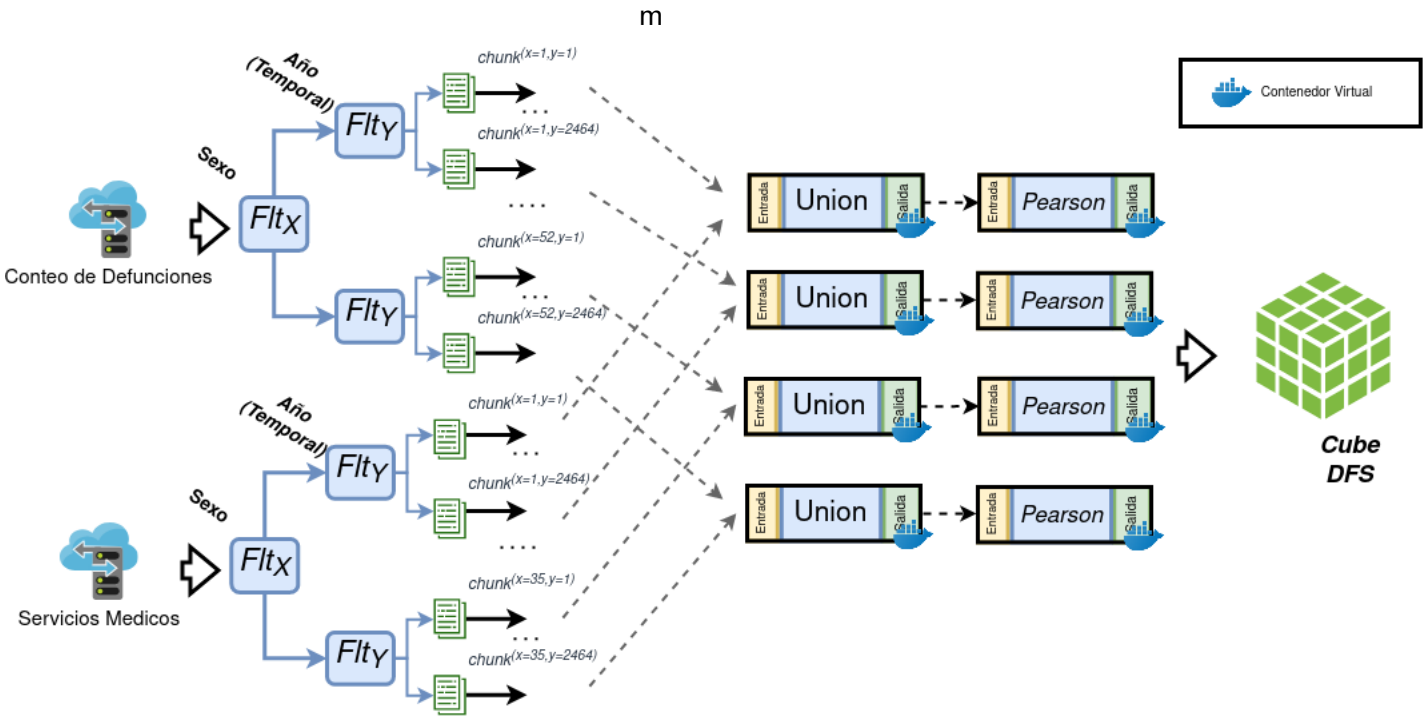


Figura 4.6: Representación conceptual del experimento 2

Dentro de las pruebas realizas, la fuente de datos de defunciones cuenta con 4 tipos de la variable sexo, los cuales son Hombres, Mujeres, NE/NA (No especificado) y Total, para esto el filtrado por sexo estaría llegando a un destino de cuatro trabajadores los cuales están dedicados a trabajar con una de los cuatro tipos de sexo. Ya que haya pasado sete filtrado, se estará pasando a un segundo filtrando, pero a través de la variable año (Temporal), la cual será segmentada mediante rangos de un año por cada sexo (primer filtrado), cabe resaltar que los datos corresponden del año 2000 al 2020. Así habría terminado el *map-reduce* de las fuentes de datos, y con esto poder tener distintos cubos de datos para fusionarlos y poder generar la correlación en base en los parámetros de entrada.

Las pruebas consisten en incrementar los trabajadores de procesamiento dedicados a realizar la fusión de los datos y la correlación de los mismos. Esto quiere decir, que por cada nodo encargado de temporal (4 nodos en total) existirán de uno a cinco trabajadores de fusión y correlación de forma secuencial. En total, la prueba final contendría cinco trabajadores de fusión y cinco trabajadores de correlación por cada sexo, es decir, 20 trabajadores por operación para las dos fuentes de datos.

Tomando en cuenta que nos estamos comparando a la solución original ($S.O.$), que en este caso es tomar por completo las dos fuentes de datos, aplicarles la fusión de datos y por último la correlación, sin realzar ninguna segmentación dentro de los datos.

En total se registraron 6 pruebas con diferentes trabajadores, estas se describen a continuación:

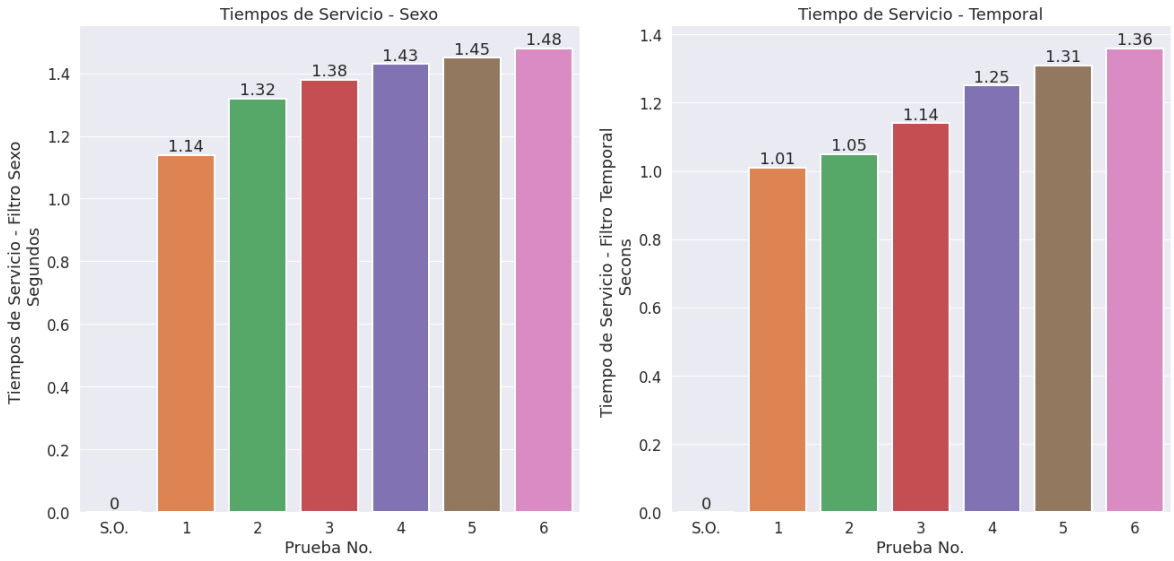
	Cantidad de trabajadores por prueba						
Nombre	$S.O.$	1	2	3	4	5	6
Sexo	0	1	1	1	1	1	1
Temporal	0	1	4	4	4	4	4
Fusión(Unión)	1	1	4	8	12	16	20
Correlación	1	1	4	8	12	16	20

4.2.2 Tiempos de servicio del filtrado de datos

Durante esta sección se encuentran los tiempos de servicios de los filtrados por los cuales fueron procesados las fuentes de datos. Los filtrados llevan a cabo el proceso de segmentación de los datos, balanceo de la carga entre los diferentes trabajadores y distribución de los cubos creados.

El filtrado de este caso de uso cuenta con dos fases, el filtrado mediante sexo y año como se mencionó anteriormente. El conjunto de datos cuenta con 4 tipos de sexo y con años del 2000 al 2020, los tiempos de servicio obtenidos en estas fases fueron los siguientes (Figura 4.7):

Aquí se presenta un comportamiento esperado, conforme crezcan el número de trabajadores para los filtrados, se estaría elevando el tiempo de servicio debido a que estaría filtrando y distribuyendo los diferentes segmentos de las dos fuentes de datos con cada vez más trabajadores, el crecimiento de los tiempos está en un aproximado del 29 % en el filtrado del sexo (Figura 4.7(a)) y un crecimiento del 34 % del filtrado de los años por rangos de uno (Figura 4.7(b)). Tomando en cuenta que la solución original no se le están implementando estos procesos.



((a)) Sexo ((b)) Año - Temporal

Figura 4.7: Tiempos de servicio en filtrado

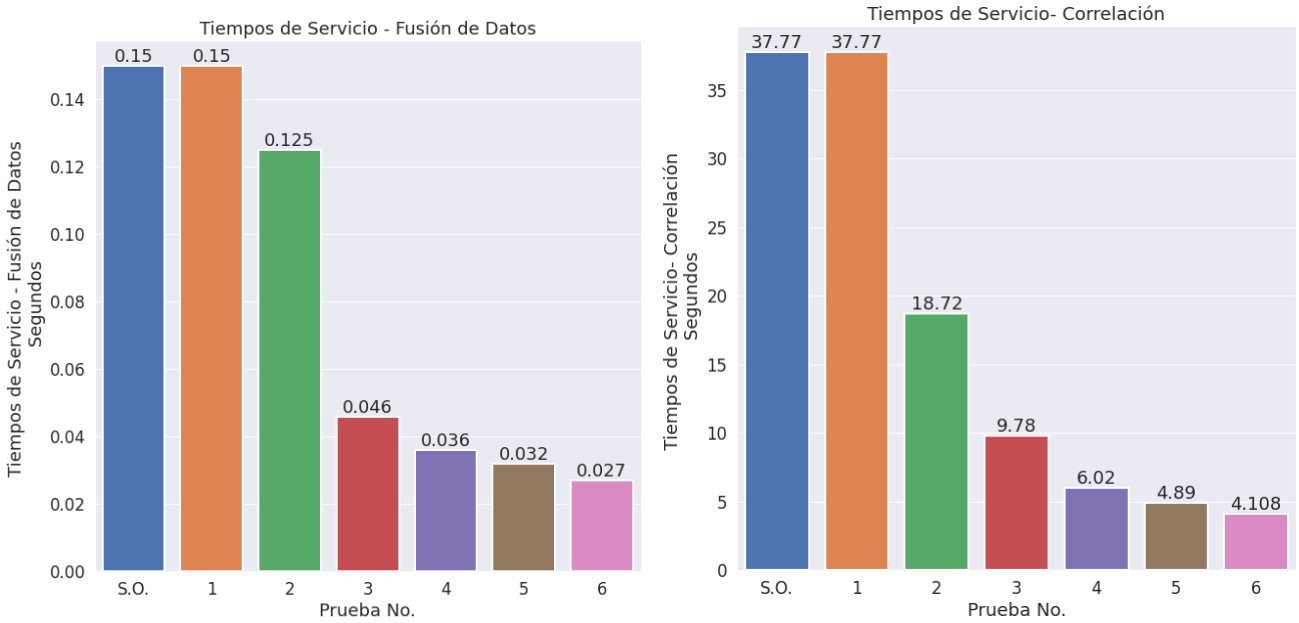
4.2.3 Tiempos de servicio de fusión de datos y correlación

Para esta sección se mostrarán los tiempos de servicio obtenidos en los procesos de analítica de fusión de datos y correlación aplicados a las fuentes de datos anteriormente mencionadas.

La ventaja de este prototipo es que empieza a tomar ventaja una vez que sé allá pasado los proceso de filtrado, debido a que la lectura, procesamiento y distribución de los cubos de datos son en fragmentos más pequeños que el conjunto de datos original y se distribuyen en múltiples trabajadores para realizar el proceso en paralelo.

En la Figura 4.9, en rasgos generales, se percibe una disminución de los tiempos de servicio en forma exponencial tanto para el proceso de fusión de datos (Figura 4.8(a)) como el proceso de correlación (Figura 4.8(b)). Enfocándonos en el proceso de fusión de datos, se obtiene una mejoría del 80 % entre un trabajador y utilizando 20 trabajadores. Ahora bien, la mejoría en los tiempos de servicio obtenidos dentro del proceso de correlación es del 89 % respecto a tiempos de un trabajador contra 20 trabajadores.

La solución original se ejecutó en un programa externo de forma secuencial, y prácticamente está dando los mismos resultados que en el prototipo con un solo trabajador tanto en fusión de datos como en correlaciones.



((a)) Fusión de datos (Unión)

((b)) Correlación

Figura 4.8: Tiempos de servicio en fusión y correlación

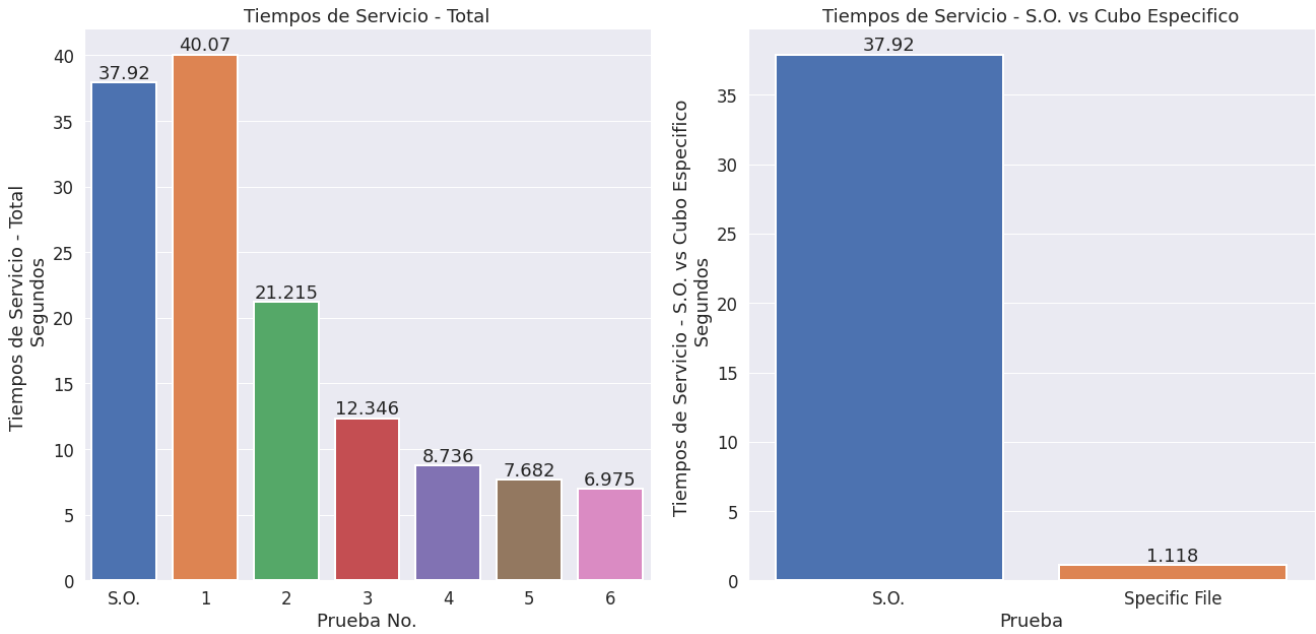
4.2.4 Tiempos de servicios totales

En este apartado mostraremos la suma de las diferentes transformaciones por las cuales pasaron las dos fuentes de datos. Recordando que las transformaciones fueron: dos procesos de filtrado, fusión de las fuentes y correlación.

En la Figura 4.12(a), podemos observar la suma de los tiempos de servicio de todas las transformaciones por las que pasaron las fuentes de datos. Como era de esperarse que la solución original ($S.O.$) tuviera un mejor rendimiento (5.6 % de diferencia) en comparación del prototipo con un solo trabajador, ya que este se le están añadiendo los procesos de filtrado. Contrarrestando esta situación, al momento de generar más trabajadores que realicen los procesos de fusión de datos y

correlación, se obtiene un rendimiento favorable del 81 % con respecto a $S.O.$.

Como se puede observar en la Figura 4.12(b) el tiempo que le toma al prototipo procesar un cubo en específico es del 97 % más efectivo si solo se estuviera ejecutando la solución original.



((a)) Tiempos de servicio total

((b)) Sexo y Año especifico vs $S.O.$

Figura 4.9: Tiempos de servicio totales y especificos

En este sentido, el cubo de datos resultante se construye a partir de los productos generados por cada uno de los filtros, permitiendo que las aplicaciones utilicen los conjuntos de datos reducidos (subconjuntos) en lugar de procesar el conjunto de datos completo. Por otro lado, analizando los resultados obtenidos, estos muestran un patrón en las variables observadas, pudiéndose identificar una correlación positiva de la variable correspondiente al número de suicidios por año (*total_suic*) con el porcentaje de la población en extrema pobreza (*prop_pob_extr_20*), y en menor medida con el nivel medio de escolaridad (*graproes*). Estos resultados se pueden ver en la Figura 4.10. Usando nuestro método, es posible analizar cada uno de los años por separado y encontrar, por ejemplo, patrones de crecimiento o disminución, y una vez que el usuario encuentra un patrón que le interesa, puede consultar el segmento de datos requerido y procesarlo recursivamente con más o un BDA

diferente.

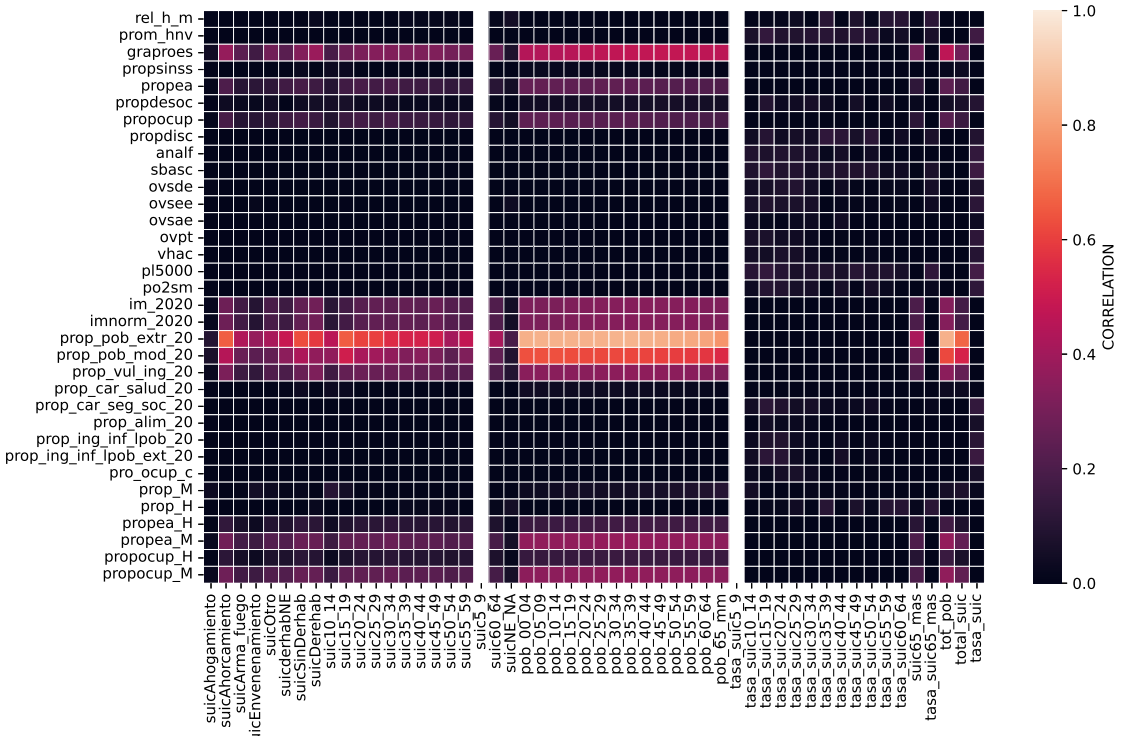


Figura 4.10: Resultado de correlación obtenida por el segmento X="Total" e Y="2022".

4.3 Estudio de caso 3 - Datos sintéticos

Para finalizar la evaluación del prototipo, se llevó a cabo un estudio de caso con fuentes de datos sintéticas, con el propósito de probar el prototipo con fuentes de datos más grandes que las anteriores mencionadas (superiores a 1GB) y así darnos cuenta del potencial que se le puede dar. Para eso se crearon un total de tres diferentes fuentes de datos, estas fuentes se crearon a partir de variables espacio temporales, siguiendo de 6 variables donde se generaron siguiendo una distribución diferente. Del lado de las variables espacio temporales, se generó una variable llamada $state$ basándose en un listado de los estados de México, por otro lado, la variable dedicada al temporal se creó a partir de

un rango de fechas desde el 1 de enero del 2000 hasta la actualidad.

A continuación presentaremos la descripción de las tres fuentes de datos sintéticos generadas. La estructura de la fuente 1 la podemos observar dentro de la Tabla 4.4, como se mencionó anteriormente se cuenta con una variable $state$ y $date$, siguiendo de 9 variables (A a la I) que fueron generadas a partir de valores aleatorios siguiendo tres distribuciones diferentes, esta fuente de datos cuenta con una cantidad de registros de 10 millones, con un tamaño aproximado a los 2.5GB.

Variable	**Descripción**
$state$	Nombre del estado del país
$date$	Fecha del registro
A	Valores distribución normal
B	Valores distribución gamma
C	Valores distribución chi-cuadrada
D	Valores distribución normal
E	Valores distribución gamma
F	Valores distribución chi-cuadrada
G	Valores distribución normal
H	Valores distribución gamma
I	Valores distribución chi-cuadrada

Tabla 4.4: Variables de fuente de datos sinteticos 1

La segunda fuente de datos contiene una estructura similar a la anterior, variando el orden de las distribuciones y los parámetros de configuración de las distribuciones (ver Tabla 4.5). El conjunto de datos tiene una cantidad igual de 10 millones de registros con un peso aproximado de 2.5GB.

Por último, la fuente de datos tres se aplicó algo similar, de igual manera cambiando el orden de las variables con sus distribuciones, tanto el tamaño del conjunto de datos como su cantidad de registros es similar a las fuentes anteriores, la descripción de las variables se encuentra descrita en la Tabla 4.6.

Los datos sintéticos se crearon con el fin de tener tres conjuntos de datos que pudiesen representar

Variable	Descripción
$state$	Nombre del estado del país
$date$	Fecha del registro
A	Valores distribución chi-cuadrada
B	Valores distribución normal
C	Valores distribución gamma
D	Valores distribución chi-cuadrada
E	Valores distribución normal
F	Valores distribución gamma
G	Valores distribución chi-cuadrada
H	Valores distribución normal
I	Valores distribución gamma

Tabla 4.5: Variables de fuente de datos sinteticos 2

Variable	Descripción
$state$	Nombre del estado del país
$date$	Fecha del registro
A	Valores distribución gamma
B	Valores distribución chi-cuadrada
C	Valores distribución normal
D	Valores distribución gamma
E	Valores distribución chi-cuadrada
F	Valores distribución normal
G	Valores distribución gamma
H	Valores distribución chi-cuadrada
I	Valores distribución normal

Tabla 4.6: Variables de fuente de datos sinteticos 3

algún tema de la vida real, en donde exista alguna correlación entre ellos y a partir de las de las tuplas de variables con mayor correlación, aplicar una regresión lineal y ver el comportamiento de los datos entre dos variables.

Las pruebas aplicadas dentro de este estudio de caso fueron similares a los casos anteriores, solo disminuyendo el incremento de 1 a 4 trabajadores por fase. Donde a partir del filtrado de espacial se empiezan a incrementar los trabajadores, para el filtrado de temporal, y los procesos de analítica.

4.3.1 Representación conceptual

Dentro de esta evaluación se diseñó un FD en donde se aplicaron dos procesos de filtrado, donde el primero fue a través de un espacial por estados (Flt_X), el segundo filtrado es por un temporal por años (Flt_Y), donde el rango de fechas de filtrado es por un año, a diferencia de los estudios de caso anteriores, este experimento es posible cambiar el fragmento de fechas a filtrar.

Ya terminado los filtrados por espacial-temporal, comienza el proceso de fusión de las fuentes, como condicional para llevar a cabo la fusión de estas son las variables $state$ y $fecha$. Posteriormente, el proceso Media-Clase ($MeanClass$) consta de dos partes, en donde se aplica un PCA a las variables resultantes del cubo de datos en la fusión (A ... I), con este análisis se obtendrán las variables más importantes generadas por el algoritmo de PCA. Ya teniendo en cuenta el número de variables importantes a seleccionar, se creará una variable de clase en donde aquellos valores por debajo de la media se les clasificarán con un 0 y los que sean igual o mayor que la media será un 1. Ahora bien, para conocer la relación entre estas fuentes, se ejecutó una correlación (nodo $Corr$) para cada cubo de datos, con esto podremos saber si existe una relación entre dos variables, aquí la combinación del cubo sobre el eje Z es utilizar dos métodos para obtener la correlación, los utilizados fueron $Z =$ [*pearson*, *spearman*], cabe resaltar que en este proceso de correlación se obtendrá las dos tuplas de variables con la correlación más alta entre los dos métodos aplicados y agregados al siguiente nodo de procesamiento.

El nodo siguiente corresponde a realizar una regresión lineal (nodo RL) en donde el eje Z de los cubos queda modificado con respecto al nodo de correlación anterior, es decir, el proceso de regresión lineal es dependiente de la correlación en este FD, ya que este provee las tuplas de variables a procesar. Este nodo estará dedicado a obtener una regresión y generar una imagen en donde se observe el comportamiento de las tuplas de variables entrantes.

El diseño conceptual de este servicio de FD lo podemos observar en la Figura 4.11:

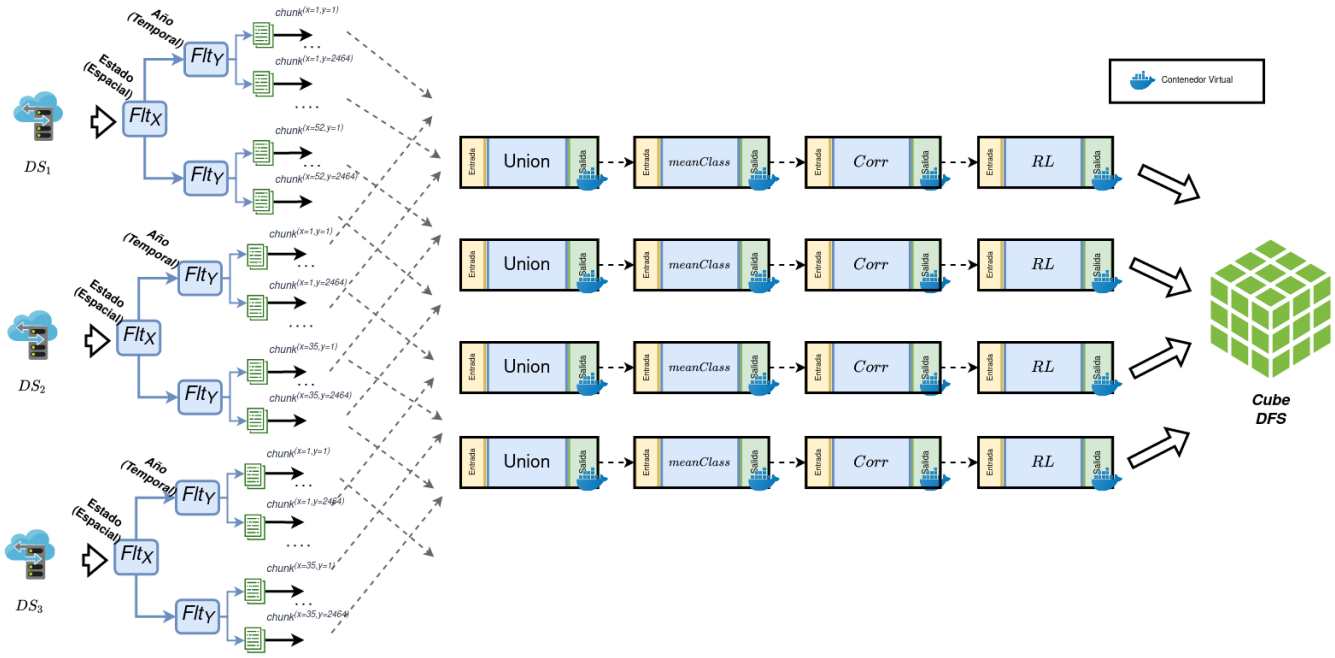


Figura 4.11: Representación conceptual del experimento 3

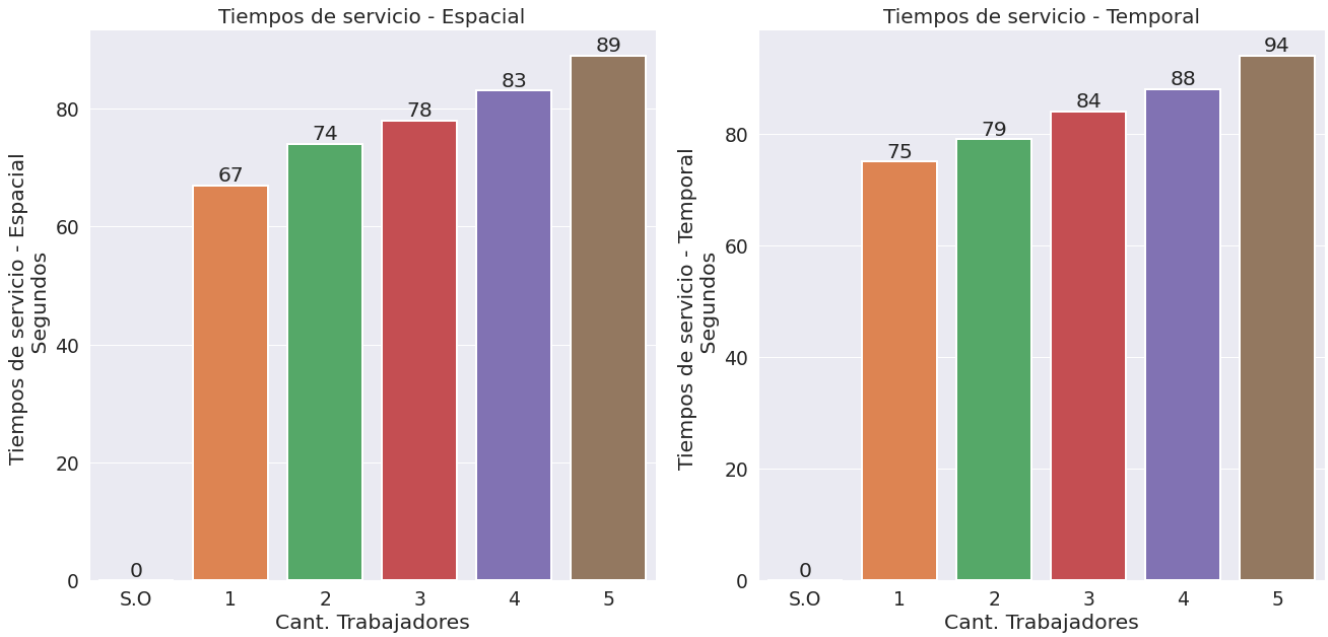
4.3.2 Tiempos de servicio del filtrado espacial-temporal

Dentro esta sección se encuentran los tiempos de servicio correspondientes a las fases de filtrado donde fueron procesadas la tres fuentes de datos. Como se ha ido mencionando, el filtrado de los datos corresponde a una segmentación, balanceo de carga y distribución de las fuentes de datos.

En la primera fase se realizó un filtrado por medio de espacial por estados de la república mexicana, los conjuntos de datos cuentan con registros en donde se ven involucrados todos los estados, para la segunda fase se efectuó un filtrado mediante año en donde los registros corresponden al año 2000 al 2022.

La solución original se ejecutó en un programa externo de forma secuencial, y prácticamente está dando los resultados similares al prototipo con un solo trabajador en los diferentes proceos aplicados.

De acuerdo con los resultados anteriores de los estudios de caso 1 y 2, se esperaba un comportamiento similar debido a que al realizar el filtrado tanto en espacial como en temporal, sus nodos destino fueron incrementando y por ende los tiempos de servicio irían aumentando en ambos



((a)) Tiempos de servicio en Espacial

((b)) Tiempos de servicio en Temporal

Figura 4.12: Tiempos de servicio en filtrado

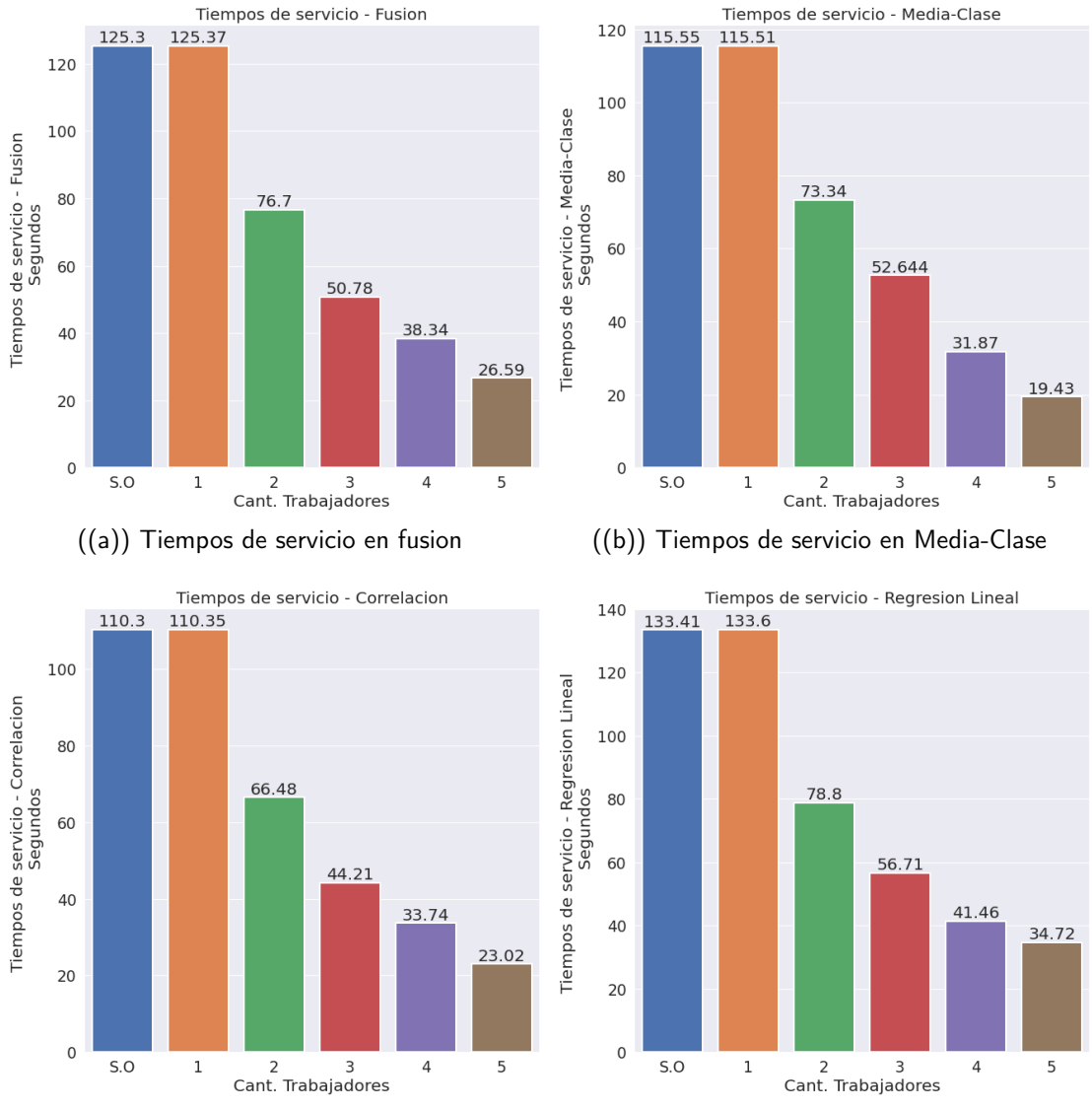
casos de filtrado (Figura 4.12). De acuerdo al filtrado espacial, estamos obteniendo un incremento del 32 % con respecto de 1 a 5 trabajadores. Ahora bien, en el filtrado por temporal por años el incremento obtenido es del 26 % aproximadamente, pero se puede notar que hay un incremento en la magnitud de los tiempos de servicio con respecto al filtrado de espacial. Con respecto a SO no se le aplicó un proceso de filtrado por eso su valor es 0 en cada caso.

4.3.3 Tiempos de servicio de los procesos

Continuando con las siguientes fases del servicio de FD, se encuentran los procesos de analítica de fusión de datos, Media-Clase, correlación y por último la regresión lineal. La ventaja en los tiempos que los tiempos empiezan a disminuir conforme vayan pasando los filtrados, debido a que los fragmentos de las fuentes de datos son más pequeños que el original y se distribuyen en trabajadores para ejecutarlos en paralelo.

En la Figura 4.13 tenemos los resultados de los tiempos de servicio de los procesos aplicados a

las fuentes de datos. En primera instancia se realizó una fusión de datos (Figura 4.13(a)) en donde se puede percibir una ganancia del 78 % en los tiempos. Siguiendo con el proceso de Media-Clase obtuvimos un rendimiento favorable en un 81 % al momento de incrementar los trabajadores. En el proceso de correlaciones se presentó un comportamiento similar en donde la ganancia obtenida fue de 80 % aproximadamente. Por último, la regresión lineal su ganancia obtenida fue 74 % con respecto a la SO.



((a)) Tiempos de servicio en fusion

((b)) Tiempos de servicio en Media-Clase

((c)) Tiempos de servicio en Correlación

((d)) Tiempos de servicio en Regressión Lineal

Figura 4.13: Tiempos de servicio en filtrado

4.3.4 Tiempos de servicio totales

Para finalizar este experimento se mostrarán la suma de los tiempos totales, en donde se ven involucrados todos los filtrados como los procesos de analítica. Así mismo se mostrarán algunos ejemplos en donde podremos percibir la utilidad de este prototipo. Así mismo se mostrará la ventaja principal del prototipo, por lo cual realiza la segmentación mediante un cubo de datos.

En la Figura 4.14, se presentan la suma de todos los tiempos de servicios de las transformaciones por las que pasaron las tres fuentes de datos. En términos generales, se presentó lo esperando en donde se obtiene una ganancia en el rendimiento del prototipo conforme se incrementan el número de trabajadores que procesan los cubos de datos en paralelo. La ganancia gel trabajador 5 con respecto a la SO es del 40 % aproximadamente, pero con respecto al utilizar un trabajador y aplicando el proceso de filtrado se obtiene una ganancia del 56 % aproximadamente, esto se debe a que era de esperarse que el prototipo con un solo trabajador fuera más lento con respecto a la SO ya que tiene procesos extra (Flt_X y Flt_Y), pero resaltando que el resultado es distinto porque la SO aplica los procesos a todo un conjunto de datos, en cambio, el prototipo está procesando los cubos de datos que pasaron por los filtros y se tendría la información resultante en segmentos.

La gran ventaja del prototipo, es que al momento de procesar en este caso un espacial y año en específico, es que el prototipo solo va a tomar ese cubo de datos, el cual paso por un filtrado antes de llegar a la fusión, Media-Clase, correlación y regresión lineal, en cambio, la solución original no estaría realizando ese proceso, por lo cual se tendría que generar otras acciones para poder llegar aun resultado similar.

Mostrando la utilidad del prototipo, podemos suponer que estas tres fuentes de datos sintéticos corresponden a registro de la vida real como educación, salud y economía, respectivamente para cada fuente de datos. Basándonos en este ejemplo, se realizan los filtrados y fusión de datos de las fuentes basándose en las variables espacio-temporales. Y obtenemos como resulta las variables descritas en la Tabla 4.7, en donde las variables con el mismo nombre corresponden a la fuente de

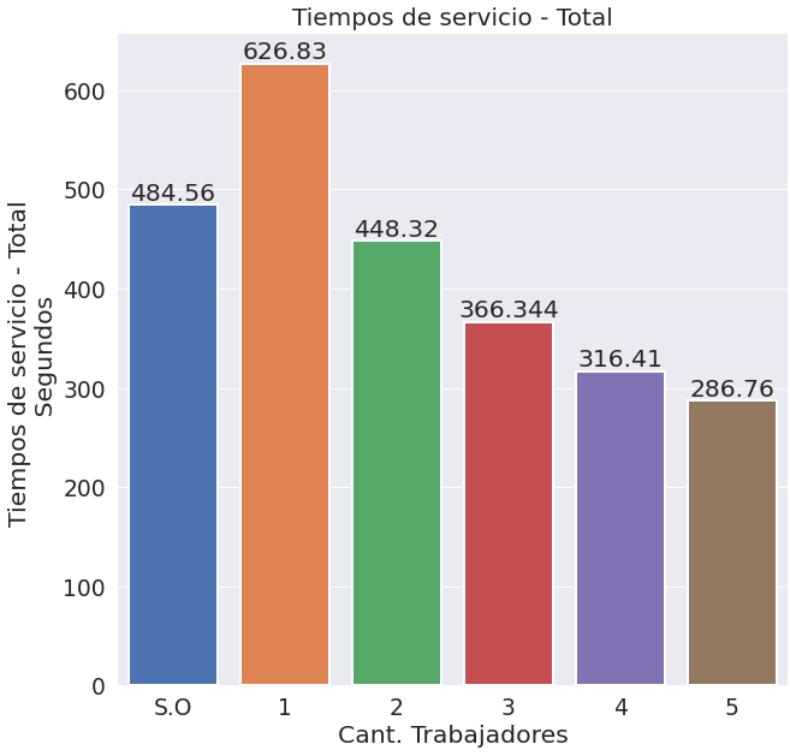


Figura 4.14: Tiempos de servicio totales

datos de educación (fuente uno), en cambio, las que cuentan con el sufijo **_X** corresponden a la fuente de datos salud (fuente dos) y las de **_Y** a la fuente de datos de economía (fuente 3).

Siguiendo con los procesos se aplica un Media-Clase, recordando que este proceso de analítica a través de PCA se obtienen el número de variables importantes en la fuente fusionada y genera una variable extra en donde aquellos valores que estén por debajo se clasifican con un 0 y los que sean igual o superior con un 1. Un ejemplo de los resultados obtenidos se presenta en la Tabla 4.8 que muestra las variables importantes para el estado de San Luis Potosí y del año 2001. En esta ocasión las variables importantes generadas fueron 3 y dio como resultado una variable correspondiente por cada fuente de datos.

Ahora bien, el siguiente proceso fue una correlación, la cual tiene como objetivo extraer las tuplas de variables con mayor correlación, y asignarlas al proceso siguiente de regresión lineal, una vez aplicada esta correlación para este mismo conjunto de San Luis Potosí, se configuró que detectara

Variable	Descripción
$state$	Nombre del estado del país
$date$	Fecha del registro
A	Valores distribución normal
B	Valores distribución gamma
C	Valores distribución chi-cuadrada
D	Valores distribución normal
E	Valores distribución gamma
F	Valores distribución chi-cuadrada
G	Valores distribución normal
H	Valores distribución gamma
I	Valores distribución chi-cuadrada
A_X	Valores distribución chi-cuadrada
B_X	Valores distribución normal
C_X	Valores distribución gamma
D_X	Valores distribución chi-cuadrada
E_X	Valores distribución normal
F_X	Valores distribución gamma
G_X	Valores distribución chi-cuadrada
H_X	Valores distribución normal
I_X	Valores distribución gamma
A_Y	Valores distribución gamma
B_Y	Valores distribución chi-cuadrada
C_Y	Valores distribución normal
D_Y	Valores distribución gamma
E_Y	Valores distribución chi-cuadrada
F_Y	Valores distribución normal
G_Y	Valores distribución gamma
H_Y	Valores distribución chi-cuadrada
I_Y	Valores distribución normal

Tabla 4.7: Variables de fuentes de datos fusionadas

Variable	Descripción
class_G	Valores distribución normal
class_D_X	Valores distribución chi-cuadrada
class_H_Y	Valores distribución chi-cuadrada

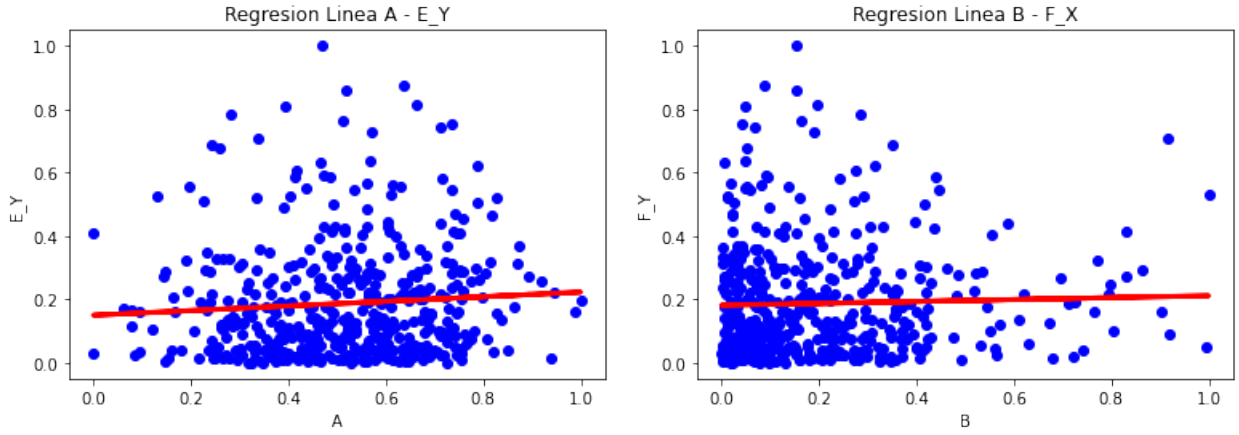
Tabla 4.8: Variables de fuentes de datos fusionadas

dos tuplas con mayor correlación, el resultado de estas dos tuplas se puede observar en la Tabla 4.9.

Variable 1	Variable 2	Descripción
A	E_Y	Variables de estudio y economía
B	F_X	Valores de estudio y salud

Tabla 4.9: Variables con mayor correlación

Como se mencionó anteriormente, estas dos tuplas de variables se están pasando al proceso de regresión lineal donde se generó una gráfica para cada una donde se observa el comportamiento (ver Figura 4.15.



((a)) Regresión entre A y E_Y ((b)) Regresión entre B y F_X

Figura 4.15: Variables con mayor correlación

Siguiendo con el ejemplo del caso de la educación, salud y económica, en la Figura 4.15(b) se puede suponer que grado de estudios tiene un efecto de generar casos de estrés presentados en alguna clínica. Por otro lado, en la Figura 4.15(b) podemos suponer un caso en donde el nivel de escolaridad de la población aumenta conforme la situación económica de la población es alta.

De esta manera se puede ver la utilidad del prototipo con volúmenes de datos grandes, aplicando algún tipo de filtrado para obtener información en específico, así como diferentes procesos de analítica, sabiendo que puedes crear tus propios servicios y consumir la información en la fase que deses.

5

Conclusiones y trabajo futuro

En este capítulo se presenta un resumen general del método para dar una idea del trabajo que se reporta en la tesis. Asimismo, se describen algunas limitaciones que pueden notarse del método propuesto. Posteriormente se listan las aportaciones que se obtuvieron con la realización del trabajo y, por último, se listan los aspectos que podrían abordarse a mediano y corto plazo para darle continuidad al trabajo desarrollado.

5.1 Resumen

En esta tesis se presentó un método de orquestación para servicios de fusión de datos, el cual está basado en variables espacio-temporales. El método consta de las siguientes fases:

- **Fusión:** se unen las fuentes de datos a partir de variables espacio-temporales.
- **División:** segmentación de la fuente de datos.
- **Conversión:** aplicar una proceso de analítica.

- **Consolidación:** selección de los cubos de datos de interés.
- **Visualización y consumo:** elección de los cubos para crear información.
- **Retorceder:** ver o consumir resultados de alguna fase anterior.

El argumento para obtener el método propuesto es que se desea mitigar o, en el mejor de los casos, evitar escenarios de fusión de datos que tengan dependencia de *usuario-proveedor* y *fusión-orquestador*. A partir del método propuesto se implementó un prototipo funcional, el cual fue probado en diversas pruebas preliminares, tres estudios de caso (dos con datos reales y uno con datos sintéticos). La evaluación experimental realizada mostró la viabilidad del método para dividir conjuntos de datos en segmentos más pequeños para su posterior procesamiento. Cada segmento se indexa y almacena en una estructura de cubo de datos para consultas posteriores sobre varios servicios de análisis y fusión de datos. El enfoque del método permite el despliegue de clones de Big Data Applications (BDA), permitiendo el procesamiento paralelo de múltiples segmentos del conjunto de datos en diferentes equipos. Así pues, el método ha permitido diseñar soluciones de servicio de análisis basadas en la declaración de variables, que pueden ser viables para procesar cualquier tipo de conjunto de datos con un conjunto de variables categóricas, espaciales y temporales.

A partir del trabajo realizado, se puede responder a las preguntas de investigación planteadas al inicio del trabajo:

1. Objetivo específico 1: *Definir un modelo de procesamiento agnóstico (de la infraestructura y/o plataforma) para crear servicios dinámicos de FD sobre un entorno de nube mediante cláusulas declarativas*. Puede verse que este objetivo se cumple con el método de orquestación de datos presentado, dado que mediante las cláusulas declarativas los usuarios pueden moldear los servicios de FD de acuerdo a sus necesidades y así mismo tener un control en cada fase presentada anteriormente. Así mismo, la utilización de la plataforma Docker nos permite, desplegar cualquiera aplicación en los computadores que cuenten con esta, sin preocuparse con alguna instalación externa a la de Docker.

2. Objetivo específico 2: *Crear un esquema de despliegue y acoplamiento de servicios de FD definido por variables espacio-temporales empleando el modelo anterior*. Puede verse que este objetivo se cumple con la declaración de los cubos de datos de las fuentes de entrada para comunicarse entra los diferentes fases de los servicios de FD, ya que con estos cubos se llevaría un control desde las fuentes fusionadas (F) hasta la segmentación deseada por el usuario.

5.2 Limitaciones

La definición del método se realizó con miras a cumplir con los objetivos planteados para la tesis. A partir de esta definición se definieron, estructuraron y organizaron los componentes del método. Estos componentes posteriormente se implementaron en un prototipo funcional que ha proporcionado resultados prometedores. No obstante, hay algunas cosas que se asumieron de manera inicial que ahora pueden verse como limitantes del método. A continuación se describen las más importantes.

1. El conjunto de directivas es limitado. Si bien el conjunto actual de directivas permite realizar las fusión de datos desde distintas fuentes, creemos que es posible incrementarlo a medida que se desee agregar más funcionalidad y otras prestaciones al método propuesto. Las directivas actuales son todas útiles, cuyo desempeño puede dar lugar a crear funcionalidades más especializadas.

2. El método de orquestación es de creación propia. Dada la naturaleza de las tareas a desarrollar, consideramos que lo mejor era iniciar los componentes de orquestación desde cero. Creemos que podría incluirse alguna funcionalidad de alguna de las variantes de Kubernetes[1]

3. Infraestructura de cómputo no distribuida. Propiamente esta no es una limitante tal cual, pero puede verse como tal. Para la experimentación y pruebas preliminares que se realizaron con el prototipo funcional del método se empleó una infraestructura de cómputo que si bien no es

[1] https://kubernetes.io

pequeña y tiene buenas capacidades en los recursos, es una infraestructura que está físicamente en una única localización, el centro de datos de Cinvestav Tamaulipas. Esta infraestructura fue más que suficiente para realizar toda la experimentación. Creemos que se podría emplear una infraestructura que esté repartida en diversas ubicaciones geográficas sin necesidad de hacer modificaciones al método. La configuración y puesta en marcha de esa infraestructura distribuida dependería del administrador de la nube. El usuario del prototipo funcional no notaría en dónde está ejecutando el prototipo, ni dónde están almacenados los datos.

5.3 Aportaciones

1. Un método de orquestación para la fusión de datos, el cual es independiente de la infraestructura en la que ejecuta (agnóstico).

2. Un prototipo funcional que implementa el método propuesto, el cual ha sido probado de manera exitosa con dos fuentes de datos reales y una sintética. Los resultados obtenidos son prometedores.

3. Un artículo de congreso aceptado en el *9th International Conference on Bioinformatics Research and Applications* (ICBRA 2022), que se realizará del 18 al 20 de septiembre de 2022 en Berlín, Alemania. Las actas se publicarán en el ACM Conference Proceedings (ISBN: 978-1-4503-8426-1). Se hará una presentación oral del trabajo enviado. En este trabajo se reporta el método propuesto y los resultados obtenidos en dos estudios de caso: uno sobre segmentación de regiones climáticas (microclimas) en México y otro sobre el pronóstico de la tasa de defunción por consumo de sustancias ilegales en México.

4. La aplicación del método, mediante el prototipo implementado en conjunción con una herramienta de análisis de datos (Xelhua[2]), en un estudio sobre causas de mortalidad en

[2]https://conacyt.mx/pronaces/pronaces-salud/ciencia-de-datos-y-salud/plataforma-para-la-gestion-aseguramiento-intercambio-analisis-y-preservacion-de-datos-medicos

México en el período de 2000 a 2020. Esto se realizó mediante una estancia académica en el Centro de Investigación en Salud Mental Global del Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz[3] bajo la supervisión de la Dra. Martha Cordero Oropeza[4].

5.4 Trabajo futuro

El plan a futuro es seguir mejorando el método agregando características no funcionales, además de seguir probando diferentes casos de uso para corroborar la utilidad de la orquestación usando variable espacio-temporales.

- *Incluir fuentes de datos no estructuradas*. Hasta la fecha se han realizado pruebas en la experimentación sobre 2 problemáticas reales y una ficticia. En estos escenarios se han considerado fuentes estructuradas de datos, provenientes de bases de datos relacionales reales y la generada a modo. En esos escenarios el desempeño del prototipo que implementa el método propuesto ha demostrado buenos resultados. Por cuestiones de tiempo e impedimentos técnicos no fue posible obtener datos de fuentes no estructuradas, p.ej. flujos de audio o video, flujos de texto, por mencionar algunas. Consideramos que en un futuro a mediano plazo podrían diseñarse e implementarse los módulos que permitan la adquisición y conformación de esas fuentes. Aunque eso no garantiza que se puedan conseguir las fuentes de datos, finalmente dependerá de su disponibilidad.

- *Integrar conectores nativos a bases de datos externas (MongoDB, Amazon S3 y Google Storage)*. En el aspecto tecnológico, el prototipo actualmente no usa un manejador de base de datos, sino que se emplean librerías de bajo nivel de Python. Se considera que en el mediano plazo podrían integrarse conectores nativos de los manejadores de bases de datos de los proveedores más conocidos: MongoDB, Amazon y Google. Aunque esto puede verse

[3] http://www.inprf.gob.mx
[4] http://www.inprf.gob.mx/psicosocialesnew/dpto_mod_intervencion.html

como regresar al escenario de dependencia *usuario-proveedor*, sería una característica más del método, la posibilidad de ser compatible con ese tipo de repositorios de datos, sin que dependan totalmente de ellos. Sería una compatilidad en un solo sentido, pues una vez reunidos los datos en el prototipo se eliminaría el vínculo con el proveedor.

- *Ampliar el conjunto de directivas del modelo declarativo*. Actualmente el conjunto de directivas definidas para crear los flujos de trabajo para fusión de datos contemplan las posibilidades mínimas requeridas para la tarea de fusión. No obstante, este conjunto podría extenderse conforme se deseen incorporar tareas más especializadas de fusión o la incorporación de herramientas de terceros (MongoDB, Amazon, Google). Por ejemplo, podrían incorporarse directivas para seleccionar solo variables específicas, unir datos de dos variables en una misma variable, segmentar texto, etc.

5.5 Datos de contacto

Todo el trabajo reportado en esta tesis (código de las aplicaciones desarrolladas, compendio de artículos revisados, datos generados y adquiridos por cuenta propia) está disponible a petición de los interesados mediante los siguientes correos.

José Carlos Morín García

- Correo: josemorin9842@gmail.com
- Página web: `https://www.tamps.cinvestav.mx/~jmorin/`

José Luis González Compeán

- Correo: joseluis.gonzalez@cinvestav.mx
- Página web: `http://www.adaptivez.org.mx/Adaptivez2015/index.php`

Iván López Arévalo

- Correo: ilopez@cinvestav.mx
- Página web: `https://www.tamps.cinvestav.mx/~ilopez/`

A

Anexos

A.1 Estudio de caso - Datos poblacionales

En este anexo se encuentran las tablas que describen las variables que conforman las fuentes de datos utilizadas en el estudio de caso.

A.1.1 Variables de defunciones

Variables	**Descripción**
cve_ent_mun	Clave del municipio
POB_TOT	Población total
analf	% de población analfabeta de 15 años o más
sbasc	% de población de 15 años o más sin educación básica
ovsde	% de ocupantes en viviendas sin drenaje ni excusado

ovsee	% de ocupantes en viviendas sin energía eléctrica
ovsae	% de ocupantes en viviendas sin agua entubada
ovpt	% de ocupantes en viviendas con piso de tierra
vhac	% de viviendas particulares con hacinamiento
pl5000	% de pob. que vive en localidades menores a 5 000 habitantes
po2sm	% de pob. ocupada con ingresos de hasta 2 salarios mínimos
IM_2020	Índice de marginación municipal
gm_2020_a	Grado de marginación municipal
im_2020	Índice de marginación municipal
rel_h_m	Relación hombres-mujeres
prom_hnv	Promedio de hijas e hijos nacidos vivos
graproes	Grado promedio de escolaridad
propsinss	Prop. de población sin afiliación a servicios de salud
propea	Prop. de población de 12 años y más económicamente activa
propdesoc	Prop. de población de 12 años y más desocupada
propocup	Prop. de población de 12 años y más ocupada
propdisc	Prop. de población con discapacidad
imnorm_2020	Índice de marginación municipal normalizado
prop_pob_extr_20	Prop. de pob. en pobreza extrema en el municipio
prop_pob_mod_20	Prop. de pob. en pobreza moderada en el municipio
prop_vul_ing_20	Prop. de pob. vulnerable por ingresos en el municipio
prop_car_salud_20	Prop. de pob. que carece de acceso a los servicios de salud
prop_car_seg_soc_20	Prop. de pob. que carece de acceso a la seguridad social
prop_alim_20	Prop. de pob. que carece de acceso a la alimentación

prop_ing_inf_lpob_20	Prop. de pob. con ingreso inferior a la línea de pobreza
prop_ing_inf_lpob_ext_20	Prop. de pob. con ingreso inferior a la línea de pobrezaextrema
pro_ocup_c	Media de ocupantes por cuarto en viviendas particulares
prop_M	Prop. de mujeres que residen habitualmente
prop_H	Prop. de hombres que residen habitualmente
propea_H	Prop. de hombres de 12 años y más económicamente activos
propea_M	Prop. de mujeres de 12 años y más económicamente activas
propocup_H	Prop. de hombres de 12 años y más ocupados
propocup_M	Prop. de mujeres de 12 años y más ocupadas
propdesocup_H	Prop. de hombres de 12 años y más desocupados
pdesocup_M	Prop. de mujeres de 12 años y más desocupadas
im_2020	Índice de marginación normalizado

Tabla A.1: Variables de defunciones

A.1.2 Variables de defunciones

Variables	Descripción
cve_ent_mun	Clave del municipio
pob_tot	Población total
analf	% de Pob. analfabeta de 15 años o más en el municipio
sprim	% de Pob. de 15 años o más sin primaria completa
sbasc	% de Pob. de 15 años o más sin educación básica
ovsde	% de ocupantes en viviendas sin drenaje ni excusado
ovsee	% de ocupantes en viviendas sin energía eléctrica
ovsae	% de ocupantes en viviendas sin agua entubada
ovpt	% de ocupantes en viviendas con piso de tierra
vhac	% de viviendas particulares con hacinamiento
pl5000	% de población que vive en localidades menores a 5 000 habitantes
po2sm	% de población ocupada con ingresos de hasta 2 salarios mínimos
gm	Grado de marginación municipal
im	Índice de marginación municipal
rel_h_m	Relación hombres-mujeres en el municipio
prom_hnv	Promedio de hijas e hijos nacidos vivos en el municipio
graproes_M	Grado promedio de escolaridad de las mujeres
graproes_H	Grado promedio de escolaridad de los hombres
graproes	Grado promedio de escolaridad en el municipio
propsinss	Prop. de pob. sin afiliación a servicios de salud
propea	Prop. de pob. de 12 años y más económicamente activa
propea_H	Prop. de hombres de 12 años y más económicamente activos

propea_M	Prop. de mujeres de 12 años y más económicamente activas
propdesoc	Prop. de población de 12 años y más que no tenían trabajo
propdesocup_H	Prop. de hombres de 12 años y más desocupados en el municipio
pdesocup_M	Prop. de mujeres de 12 años y más desocupadas en el municipio
propocup	Prop. de población de 12 años y más que trabajaron
propocup_H	Prop. de hombres de 12 años y más ocupados en el municipio
propocup_M	Prop. de mujeres de 12 años y más ocupadas en el municipio
proconlim	Prop. de pob. con discapacidad en el municipio
propslim	Prop. de pob. que no tiene discapacidad
prop_pob	Prop. de pob. en situación de pobreza
prop_pob_extr	Prop. de pob. en pobreza extrema en el municipio
prop_pob_mod	Prop. de pob. en pobreza moderada en el municipio
prop_vul_ing	Prop. de pob. vulnerable por ingresos en el municipio
prop_car_salud	Prop. de pob. que carece de acceso a los servicios de salud
prop_car_seg_soc	Prop. de pob. que carece de acceso a la seguridad social
prop_alim	Prop. de pob. que carece de acceso a la alimentación
prop_ing_inf_lpob	Prop. de pob. con ingreso inferior a la línea de pobreza
prop_ing_inf_lpob_ext	Prop. de pob. con ingreso inferior a la línea de pobreza extrema
prom_ocup	Media de ocupantes en viviendas particulares
pro_ocup_c	Media de ocupantes por cuarto en viviendas particulares habitadas
prop_M	Prop. de mujeres que residen habitualmente en el municipio
prop_H	Prop. de hombres que residen habitualmente en el municipio
prop_hojef_M	Prop. de hogares con jefatura femenina
prop_hojef_H	Prop. de hogares con jefatura masculina

pe_inac	Población no económicamente activa
prop_una_car_soc_min	Prop. de Pob. con al menos una carencia social
prop_rez_educa	Prop. de Pob. con rezago educativo
prop_tres_car_soc_min	Prop. de Pob. con tres o más carencias sociales
prop_no_pob_no_vul	Prop. de Pob. no pobre ni vulnerable
prop_car_alim	Prop. de Pob. que carece de acceso a la alimentación
prop_car_cal_vivi	Prop. de Pob. que carece de calidad de vivienda
prop_vul_car_soc	Prop. de Pob. vulnerable por carencias sociales
prop_pein_H	Prop. de hombres no económicamente activos
propdesoc_M	Prop. de mujeres de 12 años y más desocupadas en el municipio
prop_pein_M	Prop. de mujeres no económicamente activas
prop_pe_inac	Prop. de Pob. no económicamente activa
prop_car_serv_bas_viv	Prop. de Pob. que carece de servicios básicos en su vivienda

Tabla A.2: Variables de macroeconómicas

Bibliografía

[1] Abbas, H. F. (2021). Management of network service orchestration and 5g networks. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(10):1109–1114.

[2] Ahmad, Z., Nazir, B., and Umer, A. (2021). A fault-tolerant workflow management system with quality-of-service-aware scheduling for scientific workflows in cloud computing. *International Journal of Communication Systems*, 34(1):e4649.

[3] Bansal, S. K. (2014). Towards a semantic extract-transform-load (etl) framework for big data integration. In *2014 IEEE International Congress on Big Data*, pages 522–529.

[4] Behl, A., Dutta, P., Lessmann, S., Dwivedi, Y. K., and Kar, S. (2019). A conceptual framework for the adoption of big data analytics by e-commerce startups: a case-based approach. *Information systems and e-business management*, 17(2):285–318.

[5] Bernard, L., Kanellopoulos, I., Annoni, A., and Smits, P. (2005). The european geoportal—one step towards the establishment of a european spatial data infrastructure. *Computers, Environment and Urban Systems*, 29(1):15–31. Geoportals.

[6] Blasch, E., Steinberg, A., Das, S., Llinas, J., Chong, C., Kessler, O., Waltz, E., and White, F. (2013a). Revisiting the jdl model for information exploitation. In *Proceedings of the 16th International Conference on Information Fusion*, pages 129–136.

[7] Blasch, E., Steinberg, A., Das, S., Llinas, J., Chong, C., Kessler, O., Waltz, E., and White, F. (2013b). Revisiting the jdl model for information exploitation. In *Proceedings of the 16th International Conference on Information Fusion*, pages 129–136.

[8] Camargo-Vega, J. J., Camargo-Ortega, J. F., and Joyanes-Aguilar, L. (2015). Conociendo Big Data. *Revista Facultad de Ingeniería*, 24:63 – 77.

[9] Castanedo, F. (2013). A review of data fusion techniques. *The Scientific World Journal*, 2013:1–2.

[10] Castronova, A. M., Goodall, J. L., and Elag, M. M. (2013). Models as web services using the open geospatial consortium (ogc) web processing service (wps) standard. *Environmental Modelling & Software*, 41:72–83.

[11] Chang, W. L., Grady, N., et al. (2019). Nist big data interoperability framework: Volume 1, definitions.

[12] Chen, Y., Argentinis, J. E., and Weber, G. (2016). Ibm watson: how cognitive computing can be applied to big data challenges in life sciences research. *Clinical therapeutics*, 38(4):688–701.

[13] Dasarathy, B. V. (1997). Sensor fusion potential exploitation-innovative architectures and illustrative applications. *Proceedings of the IEEE*, 85(1):24–38.

[14] Diouf, P. S., Boly, A., and Ndiaye, S. (2018). Variety of data in the etl processes in the cloud: State of the art. In *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, pages 1–5. IEEE.

[15] Dubey, A. K., Kumar, A., and Agrawal, R. (2021). An efficient aco-pso-based framework for data classification and preprocessing in big data. *Evolutionary Intelligence*, 14(2):909–922.

[16] Durrant-Whyte, H. F. (1990). Sensor models and multisensor integration. In *Autonomous robot vehicles*, pages 73–89. Springer.

[17] El-Seoud, S., El-Sofany, H., Abdelfattah, M., and Mohamed, R. (2017). Big data and cloud computing: Trends and challenges. *International Journal of Interactive Mobile Technologies (iJIM)*, 11:34.

[18] Elbanna, S. (2006). Strategic decision-making: Process perspectives. *International Journal of Management Reviews*, 8(1):1–20.

[19] Epitropou, V., Karatzas, K., Karppinen, A., Kukkonen, J., and Bassoukos, A. (2012). Orchestration services for chemical weather forecasting models in the frame of the pescado project. In *8th International Conference on Air Quality—Science and Application; Athens*, pages 19–23.

[20] Fathi, M., Kashani, M. H., Jameii, S. M., and Mahdipour, E. (2021). Big data analytics in weather forecasting: a systematic review. *Archives of Computational Methods in Engineering*, pages 1–29.

[21] Gajić, V. (2013). Historical review of study on weather influence on people and development of medical meteorology. *ABC-časopis urgentne medicine*, 13(2-3):65–69.

[22] Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B. (2017). The modern-era retrospective analysis for research and applications, version 2 (merra-2).

[23] Georgakopoulos, D., Hornick, M., and Sheth, A. (1995). An overview of workflow management: From process modeling to workflow automation infrastructure. *Distributed and parallel Databases*, 3(2):119–153.

[24] Gorton, I. and Klein, J. (2015). Distribution, data, deployment: Software architecture convergence in big data systems. *IEEE Software*, 32(3):78–85.

[25] Gowat, R. C. (2020). Blockchain traceability for the counterfeit detection and avoidance program (cdap) final report. Technical report, Accenture Federal Services Arlington United States.

[26] Grasse, D. and Nelson, G. (2006). Base sas® vs. sas® data integration studio: Understanding etl and the sas tools used to support it. *SAS Users Group International*.

[27] Healy, P. and Nikolov, N. S. (2001). How to layer a directed acyclic graph. In *International Symposium on Graph Drawing*, pages 16–30. Springer.

[28] Hilker, T., Wulder, M. A., Coops, N. C., Linke, J., McDermid, G., Masek, J. G., Gao, F., and White, J. C. (2009). A new data fusion model for high spatial-and temporal-resolution mapping of forest disturbance based on landsat and modis. *Remote Sensing of Environment*, 113(8):1613–1627.

[29] Hofacker, C. F., Malthouse, E. C., and Sultan, F. (2016). Big data and consumer behavior: Imminent opportunities. *Journal of consumer marketing*.

[Hollingsworth] Hollingsworth, D. Workflow management coalition the workflow management coalition specification workflow management coalition the workflow reference model.

[31] Janssen, M., van der Voort, H., and Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of business research*, 70:338–345.

[32] King, R. C., Villeneuve, E., White, R. J., Sherratt, R. S., Holderbaum, W., and Harwin, W. S. (2017). Application of data fusion techniques and technologies for wearable health monitoring. *Medical engineering & physics*, 42:1–12.

[33] Li, G., Tan, J., and Chaudhry, S. S. (2019). Industry 4.0 and big data innovations.

[34] Li, X., Vernon, C., Chen, M., Wang, H., and Hou, Z. (2021). A self-evolution data fusion platform for large-scale water models.

[McHugh] McHugh, B. Simplifying big data with data orchestration.

[36] Naik, N. (2016). Connecting google cloud system with organizational systems for effortless data analysis by anyone, anytime, anywhere. In *2016 IEEE International Symposium on Systems Engineering (ISSE)*, pages 1–6. IEEE.

[37] OHDSI (2021). Omop common data model.

[38] Okada, M. (2021). Big data and real-world data-based medicine in the management of hypertension. *Hypertension Research*, 44(2):147–153.

[39] Opara-Martins, J., Sahandi, R., and Tian, F. (2016a). Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective. *Journal of Cloud Computing*, 5(1):1–18.

[40] Opara-Martins, J., Sahandi, R., and Tian, F. (2016b). Critical analysis of vendor lock-in and its impact on cloud computing migration: A business perspective. *Journal of Cloud Computing*, 5(1).

[41] Papadakis-Vlachopapadopoulos, K., Dimolitsas, I., Dechouniotis, D., Tsiropoulou, E. E., Roussaki, I., and Papavassiliou, S. (2021). On blockchain-based cross-service communication and resource orchestration on edge clouds. *Informatics*, 8(1).

[42] Paščinski, U., Trnkoczy, J., Stankovski, V., Cigale, M., and Gec, S. (2018). Qos-aware orchestration of network intensive software utilities within software defined data centres.

[43] Pfeuffer, A. and Dietmayer, K. (2018). Optimal sensor data fusion architecture for object detection in adverse weather conditions. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 1–8.

[44] Quiroz, J. C., Chard, T., Sa, Z., Ritchie, A. G., Jorm, L., and Gallego, B. (2021). Extract, transform, load framework for the conversion of health databases to omop. *medRxiv*.

[45] Rajaraman, V. (2016). Big data analytics. *Resonance*, 21(8):695–716.

[46] Riahi, Y. (2018). Big data and big data analytics: Concepts, types and technologies. *International Journal of Research and Engineering*, 5:524–528.

[47] Schultz, M., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L., Mozaffari, A., and Stadtler, S. (2021). Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A*, 379(2194):20200097.

[48] Singh, P. (2019). Airflow. In *Learn PySpark*, pages 67–84. Springer.

[49] Stich (2021). Apache airflow vs. google cloud dataflow vs. stitch - compare features, pricing, services, and more.

[50] Sudhakar, K. (2018). Amazon web services (aws) glue. *International Journal of Management, IT and Engineering*, 8(9):108–122.

[51] Talukder, A., Elshambakey, M., Wadkar, S., Lee, H., Cinquini, L., Schlueter, S., Cho, I., Dou, W., and Crichton, D. J. (2017). Vifi: Virtual information fabric infrastructure for data-driven discoveries from distributed earth science data. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–8. IEEE.

[52] Venkatesh, V., Raj, P., Gopalan, K., and Rajeev, T. (2011). Healthcare data fusion and presentation using service-oriented architecture (soa) orchestration mechanism. *IJCA Special Issue on Artificial Intelligence Techniques-Novel Approaches & Practical Applications*, 2:17–23.

[53] Wiemann, S. (2017). Formalization and web-based implementation of spatial data fusion. *Computers & Geosciences*, 99:107–115.

[54] Wiemann, S. and Bernard, L. (2016). Spatial data fusion in spatial data infrastructures using linked data. *International Journal of Geographical Information Science*, 30(4):613–636.

[55] Wu, H. (2021). Optimization and design based on data visualization of enterprise-class crawler systems for shipping. In *Journal of Physics: Conference Series*, volume 1746, page 012078. IOP Publishing.

[56] Wu, Y. (2020). Cloud-edge orchestration for the internet-of-things: Architecture and ai-powered data processing. *IEEE Internet of Things Journal*.

[57] Xu, M., Cui, L., Wang, H., and Bi, Y. (2009). A multiple qos constrained scheduling strategy of multiple workflows for cloud computing. In *2009 IEEE International Symposium on Parallel and Distributed Processing with Applications*, pages 629–634.

[58] Yusifov, F. F., Axundova, N. E., et al. (2021). Analysis of demographic indicators based on e-demography data system. *İTP Jurnalı*.